Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Artículos

IberLEF 2022: Resúmenes de las tareas de evaluación

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Comité Editorial

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Jaén
Año de edición: 2022

Publicado por: Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, EdificioA3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Ramón López-Cózar Delgado	Universidad de Granada (España)
Bernardo Magnini	Fondazione Bruno Kessler (Italia)
Nuno J. Mamede	Instituto de Engenharia de Sistemas e Computadores (Portugal)
M. Antonia Martí	Universidad de Barcelona (España)
M. Teresa Martín Valdivia	Universidad de Jaén (España)
Patricio Martínez-Barco	Universidad de Alicante (España)
Eugenio Martínez Cámara	Universidad de Granada (España)
Paloma Martínez Fernández	Universidad Carlos III (España)
Raquel Martínez Unanue	Universidad Nacional de Educación a Distancia (España)
Ruslan Mitkov	University of Wolverhampton (Reino Unido)
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Mariana Lara Neves	Bundesinstitut für Risikobewertung (Alemania)
Lluís Padró	Universidad Politécnica de Cataluña (España)
Manuel Palomar	Universidad de Alicante (España)
Ferrán Pla	Universidad Politécnica de Valencia (España)
German Rigau	Universidad del País Vasco (España)
Horacio Rodríguez	Universidad Politécnica de Cataluña (España)
Paolo Rosso	Universidad Politécnica de Valencia (España)
Leonel Ruiz Miyares	Centro de Lingüística Aplicada de Santiago de Cuba (Cuba)
Horacio Saggion	Universidad Pompeu Fabra (España)
Emilio Sanchís	Universidad Politécnica de Valencia (España)
Encarna Segarra	Universidad Politécnica de Valencia (España)
Thamar Solorio	University of Houston (Estados Unidos de América)
Maite Taboada	Simon Fraser University (Canadá)
Mariona Taulé	Universidad de Barcelona (España)
Juan-Manuel Torres-Moreno	Laboratoire Informatique d'Avignon / Université d'Avignon (Francia)
José Antonio Troyano Jiménez	Universidad de Sevilla (España)
L. Alfonso Ureña López	Universidad de Jaén (España)
Rafael Valencia García	Universidad de Murcia (España)
René Venegas Velásques	Pontificia Universidad Católica de Valparaíso (Chile)
Felisa Verdejo Maíllo	Universidad Nacional de Educación a Distancia (España)
Karin Vespoor	University of Melbourne (Australia)
Manuel Vilares	Universidad de Vigo (España)
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)

Revisores adicionales

Laura Alonso Alemany	Universidad Nacional de Córdoba (Argentina)
Ana-Maria Bucur	University of Bucharest (Rumanía)
Óscar Araque Iborra	Universidad Politécnica de Madrid (España)
Marco Casavantes	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Riccardo Cervero	Universidad Politécnica de Valencia (España)
Elisabet Comelles	Universidad de Barcelona (España)
Laritza Coello	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Víctor Manuel Darriba Bilbao	Universidad de Vigo (España)

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Preámbulo

La revista *Procesamiento del Lenguaje Natural* pretende ser un foro de publicación de artículos científico-técnicos inéditos de calidad relevante en el ámbito del Procesamiento de Lenguaje Natural (PLN) tanto para la comunidad científica nacional e internacional, como para las empresas del sector. Además, se quiere potenciar el desarrollo de las diferentes áreas relacionadas con el PLN, mejorar la divulgación de las investigaciones que se llevan a cabo, identificar las futuras directrices de la investigación básica y mostrar las posibilidades reales de aplicación en este campo. Anualmente la SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural) publica dos números de la revista, que incluyen artículos originales, presentaciones de proyectos en marcha, reseñas bibliográficas y resúmenes de tesis doctorales. Esta revista se distribuye gratuitamente a todos los socios, y con el fin de conseguir una mayor expansión y facilitar el acceso a la publicación, su contenido es libremente accesible por Internet.

Las áreas temáticas tratadas son las siguientes:

- Modelos lingüísticos, matemáticos y psicolingüísticos del lenguaje
- Lingüística de corpus
- Desarrollo de recursos y herramientas lingüísticas
- Gramáticas y formalismos para el análisis morfológico y sintáctico
- Semántica, pragmática y discurso
- Lexicografía y terminología computacional
- Resolución de la ambigüedad léxica
- Aprendizaje automático en PLN
- Generación textual monolingüe y multilingüe
- Traducción automática
- Reconocimiento y síntesis del habla
- Extracción y recuperación de información monolingüe, multilingüe y multimodal
- Sistemas de búsqueda de respuestas
- Análisis automático del contenido textual
- Resumen automático
- PLN para la generación de recursos educativos
- PLN para lenguas con recursos limitados
- Aplicaciones industriales del PLN
- Sistemas de diálogo
- Análisis de sentimientos y opiniones
- Minería de texto
- Evaluación de sistemas de PLN
- Implicación textual y paráfrasis

El ejemplar número 69 de la revista *Procesamiento del Lenguaje Natural* contiene trabajos correspondientes a dos apartados diferentes: comunicaciones científicas y resúmenes de las tareas de evaluación competitiva de la edición del año 2022 del foro de evaluación *Iberian Language Evaluation Forum* (IberLEF). Todos ellos han sido aceptados mediante el proceso de revisión

tradicional en la revista. Queremos agradecer a los miembros del Comité Asesor y a los revisores adicionales la labor que han realizado.

Se recibieron 40 trabajos para este número, de los cuales 30 eran artículos científicos y 10 resúmes de las tareas de evaluación competitiva del foro de evaluación IberLEF 2022. De entre los 30 artículos recibidos, 16 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 53%.

El Comité Asesor de la revista se ha hecho cargo de la revisión de los trabajos. Este proceso de revisión es de doble anonimato: se mantiene oculta la identidad de los autores que son evaluados y de los revisores que realizan las evaluaciones. En un primer paso, cada artículo ha sido examinado de manera ciega o anónima por tres revisores. En un segundo paso, para aquellos artículos que tenían una divergencia mínima de tres puntos (sobre siete) en sus puntuaciones, sus tres revisores han reconsiderado su evaluación en conjunto. Finalmente, la evaluación de aquellos artículos que estaban en posición muy cercana a la frontera de aceptación ha sido supervisada por más miembros del comité editorial. El criterio de corte adoptado ha sido la media de las tres calificaciones, siempre y cuando hayan sido iguales o superiores a 5 sobre 7.

La elaboración de este número ha contado con la aportación del Vicerrectorado de Política Científica, Investigación y Transferencia de la Universidad de A Coruña, con cofinanciación del Convenio de Acciones Estratégicas I+D+i para 2022 entre la Consellería de Cultura, Educación y Universidad de la Xunta de Galicia y la Universidad de A Coruña.

Septiembre de 2022
Los editores.

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Preamble

The *Natural Language Processing* journal aims to be a forum for the publication of high-quality unpublished scientific and technical papers on Natural Language Processing (NLP) for both the national and international scientific community and companies. Furthermore, we want to strengthen the development of different areas related to NLP, widening the dissemination of research carried out, identifying the future directions of basic research and demonstrating the possibilities of its application in this field. Every year, the Spanish Society for Natural Language Processing (SEPLN) publishes two issues of the journal that include original articles, ongoing projects, book reviews and summaries of doctoral theses. All issues published are freely distributed to all members, and contents are freely available online.

The subject areas addressed are the following:

- Linguistic, Mathematical and Psychological models to language
- Grammars and Formalisms for Morphological and Syntactic Analysis
- Semantics, Pragmatics and Discourse
- Computational Lexicography and Terminology
- Linguistic resources and tools
- Corpus Linguistics
- Speech Recognition and Synthesis
- Dialogue Systems
- Machine Translation
- Word Sense Disambiguation
- Machine Learning in NLP
- Monolingual and multilingual Text Generation
- Information Extraction and Information Retrieval
- Question Answering
- Automatic Text Analysis
- Automatic Summarization
- NLP Resources for Learning
- NLP for languages with limited resources
- Business Applications of NLP
- Sentiment Analysis
- Opinion Mining
- Text Mining
- Evaluation of NLP systems
- Textual Entailment and Paraphrases

The 69th issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers and summaries of the shared-tasks of the edition of 2022 of the evaluation forum *Iberian Languages Evaluation Forum* (IberLEF). All of these were accepted by a peer review process. We would like to thank the Advisory Committee members and additional reviewers for their work.

Forty papers were submitted for this issue, from which thirty were scientific papers and ten were summaries of the evaluation tasks of the evaluation forum IberLEF 2022. From these thirty papers, we selected sixteen (53%) for publication.

The Advisory Committee of the journal has reviewed the papers in a double-blind process. Under double-blind review the identity of the reviewers and the authors are hidden from each other. In the first step, each paper was reviewed blindly by three reviewers. In the second step, the three reviewers have given a second overall evaluation of those papers with a difference of three or more points out of seven in their individual reviewer scores. Finally, the evaluation of those papers that were in a position very close to the acceptance limit were supervised by the editorial board. The cut-off criterion adopted was the mean of the three scores given, as long as it is equal or greater than 5 out of 7.

The preparation of this issue has been supported partially by the Vice-Rectorate for Science Policy, Research and Transfer of the University of A Coruña, with co-funding from the R&D Agreement on Strategic Actions for 2022 between the Department of Culture, Education and University of the Xunta de Galicia and the University of A Coruña.

September 2022
Editorial board.

Sociedad Española para el Procesamiento del Lenguaje Natural

ISSN: 1135-5948

Artículos

IberLEF 2022: Resúmenes de las tareas de evaluación

Artículos

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 15-26 recibido 31-03-2022 revisado 19-05-2022 aceptado 23-05-2022

A computational psycholinguistic evaluation of the syntactic abilities of Galician BERT models at the interface of dependency resolution and training time

Una evaluación psicolingüístico-computacional de las capacidades sintácticas de los modelos BERT para el gallego en la intersección entre la resolución de dependencias y el tiempo de entrenamiento

Iria de-Dios-Flores, Marcos Garcia
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
{iria.dedios, marcos.garcia.gonzalez}@usc.gal

Abstract: This paper explores the ability of Transformer models to capture subject-verb and noun-adjective agreement dependencies in Galician. We conduct a series of word prediction experiments in which we manipulate dependency length together with the presence of an attractor noun that acts as a lure. First, we evaluate the overall performance of the existing monolingual and multilingual models for Galician. Secondly, to observe the effects of the training process, we compare the different degrees of achievement of two monolingual BERT models at different training points. We also release their checkpoints and propose an alternative evaluation metric. Our results confirm previous findings by similar works that use the agreement prediction task and provide interesting insights into the number of training steps required by a Transformer model to solve long-distance dependencies.
Keywords: BERT models, Galician, targeted syntactic evaluation, agreement dependencies.

Resumen: Este trabajo analiza la capacidad de los modelos Transformer para capturar las dependencias de concordancia sujeto-verbo y sustantivo-adjetivo en gallego. Llevamos a cabo una serie de experimentos de predicción de palabras manipulando la longitud de la dependencia junto con la presencia de un sustantivo intermedio que actúa como distractor. En primer lugar, evaluamos el rendimiento global de los modelos monolingües y multilingües existentes para el gallego. En segundo lugar, para observar los efectos del proceso de entrenamiento, comparamos los diferentes grados de consecución de dos modelos monolingües BERT en diferentes puntos del entrenamiento. Además, publicamos sus puntos de control y proponemos una métrica de evaluación alternativa. Nuestros resultados confirman los hallazgos anteriores de trabajos similares que utilizan la tarea de predicción de concordancia y proporcionan una visión interesante sobre el número de pasos de entrenamiento que necesita un modelo Transformer para resolver las dependencias de larga distancia.
Palabras clave: Modelos BERT, Gallego, evaluación sintáctica dirigida, dependencias de concordancia.

1 Introduction

Current language models (LMs) based on deep neural network architectures obtain impressive performance on most NLP tasks, including semantic and syntactic applications (Devlin et al., 2019). In fact, it has been argued that LSTM and Transformer models may encode syntactic information captured in an unsupervised manner from unlabeled text (Lin, Tan, and Frank, 2019; Hewitt and Manning, 2019).

To explore the syntactic capabilities of LMs, various studies have probed their grammatical competence by analyzing the resolution of long-distance dependencies using a

ISSN 1135-5948. DOI 10.26342/2022-69-1

word prediction task, often known as targeted syntactic evaluation or TSE (Linzen, Dupoux, and Goldberg, 2016; Gulordava et al., 2018). Inspired by classical psycholinguistic experiments on human sentence processing, this task consists on comparing the model probabilities for a correct and incorrect alternative in the context of the targeted syntactic phenomena. For instance, given the well-known sentence in psycholinguistic research "The **key** to the cabinets **is**|***are** on the table", a model that correctly identifies the dependency between the subject (key) and the verb should assign a higher probability to the singular form (is) than to the plural form (are), despite the presence of an intervening plural attractor noun (cabinets).

First studies have shown that recurrent neural networks (RNNs) are able to solve most cases of the agreement prediction task in several languages (though mostly in English) and scenarios (Bernardy and Lappin, 2017; Gulordava et al., 2018; Kuncoro et al., 2018b). The interest in the assessment of syntactic abilities of language models grew with the popularization of Transformer architectures (Vaswani et al., 2017), which learn relations between words using a self-attention mechanism, and seem to perform better than models based on RNNs (Devlin et al., 2019). In this respect, recent works are putting the focus on the training procedure: Pérez-Mayos, Ballesteros, and Wanner (2021) measure the impact of the amount of training data on syntactic probing, while Wei et al. (2021) analyze the influence of word frequency on subject-verb number agreement. However, to the best of our knowledge, there are still no studies exploring the models' performance along the training process, i.e., how many training steps do they need to solve long-distance dependencies. This is one of the goals of the present work.

On the evaluation side, it has been argued that instead of comparing the probability of a single correct|incorrect pair (as is|are in the example above), these experiments should use large lists of pairs representing the same phenomena (e.g., exist|exists) to observe the model's *systematicity* (i.e., in how many pairs a model succeeds) and its *likely behaviour* (the probability of generating a correct inflection) (Newman et al., 2021). Nevertheless, this type of evaluation requires large sets of target pairs and, what is more substantial, it assumes a total independence between syntax and semantics —something which is controversial from a linguistic and psycholinguistic point of view.

Taking the above into account, in this paper we investigate the ability of Transformer models to capture fundamental linguistic operations such as dependency resolution in a less studied language, Galician. Following previous research inspired by both computational modeling and psycholinguistic research, we conduct a series of word prediction experiments using a dataset that targets two types of agreement dependencies (subject-verb dependencies and noun-adjective dependencies) in two experimental conditions (short and long-distance dependencies) while also manipulating the presence of an attractor noun that acts as a lure (e.g. "O **neno** que xogaba onte alí coa nena é **alto**|***alta**").[1] First, we evaluate the overall performance of the existing monolingual and multilingual models for Galician. Secondly, in order to observe the effects of the training process, we compare the different degrees of achievement of two monolingual BERT models (which vary on the number of hidden layers, vocabulary size, and initialization) at various training steps.

In addition, we propose an alternative evaluation metric of accuracy, which puts the focus on the probability distance between the correct and the incorrect alternatives for a given semantic plausible word. This metric allows us to explore a model's confidence in producing syntactically and semantically well-formed expressions and to compare models with different vocabulary sizes.

Our contributions are the following: (i) 34 checkpoints of two BERT models for Galician which allow to explore the effects of the training steps on different tasks; (ii) a novel metric for targeted syntactic evaluation which focuses on the probability distribution between correct and incorrect alternatives; (iii) a careful comparison of the models' performance on two agreement dependencies, analyzing the impact of the learning steps, the amount of training data, the model initialization, and the depth of the neural network.

[1] Here *alto* ('tall' in masculine) agrees in its gender features with the correct antecedent *neno* ('boy'), while *alta* ('tall' in feminine) agrees in gender with the structurally irrelevant noun *nena* ('girl'), hence producing an ungrammatical dependency.

Our results confirm previous findings by similar works using the agreement prediction task and provide interesting insights into the number of training steps required by a Transformer model to solve long-distance dependencies, as they already achieve high performance at early checkpoints when trained on enough data.

The rest of this paper is organized as follows: Section 2 introduces previous work about targeted syntactic evaluation on neural language models. Then, in Section 3, we present the main characteristics of the models used for the experiments and the different checkpoints provided by our study. In Section 4 we describe our methodology, including the rationale behind the evaluation metric proposed here. Finally, the results are presented and discussed in Section 5, while Section 6 draws the conclusions of the work.

2 Background

Linzen, Dupoux, and Goldberg (2016) introduced the *number prediction task* to evaluate the performance of language models on long-distance agreement, and their results suggested that even if LSTM models seem not to generalize syntactic structures, they identify subject-verb agreement dependencies. Inspired by this paper, various studies explored the behaviour of LSTMs models on a variety of languages and syntactic phenomena, arguing that these networks may achieve near-human performance in some agreement experiments (Bernardy and Lappin, 2017; Gulordava et al., 2018; Kuncoro et al., 2018b) even though they may be alternative explanations (such as surface-based heuristics) that explain the models' success (Kuncoro et al., 2018a; Linzen and Leonard, 2018). Moving forward, Marvin and Linzen (2018) published a new dataset in English which includes not only subject-verb agreement items, but also other dependencies (e.g. anaphora, negative polarity items) and more complex constructions. Their results showed that although the behavior of various RNNs on this dataset is far from human performance, they obtain competitive results in various settings. Taking a different approach, Lakretz et al. (2019) were able to identify individual cells on a LSTM model which encode information about grammatical number and plurality in English, suggesting that the network effectively captures some morphosyntactic information from raw text.

The growing interest in this research area motivated the release of *SyntaxGym*[2], an online platform for targeted evaluation of language models (Gauthier et al., 2020), as well as datasets in different languages, such as Mueller et al. (2020) (for English, French, German, Hebrew, and Russian), Pérez-Mayos et al. (2021) (for Spanish), or Garcia and Crespo-Otero (2022) (for Galician and Portuguese), which is the one used in this work.

Unlike LSTMs, Transformer architectures (Vaswani et al., 2017) use a non recurrent neural network which learns relations between words using a self-attention mechanism, and they can be interpreted as induced-structure models (Henderson, 2020). On a comparison of LSTM and Transformer architectures, Tran, Bisazza, and Monz (2018) found that the former slightly outperform Transformers on English subject-verb agreement. In this respect, the release of large Transformer-based models, such as BERT (Devlin et al., 2019) and its variants, gave rise to a larger interest in exploring their linguistic abilities. Given that training these models is computationally expensive, most studies explore publicly available resources (Goldberg, 2019; Mueller et al., 2020). Among others, Pérez-Mayos, Ballesteros, and Wanner (2021) have shown that more training data yields better performance in most syntactic tasks in English, and Pérez-Mayos et al. (2021) compared multilingual and monolingual models in English and Spanish: the results seem to indicate that the syntactic generalization of each model type is language-specific, as some multilingual architectures work better than the monolingual ones in some scenarios and vice-versa. More recently, Garcia and Crespo-Otero (2022) evaluated a variety of BERT models for Galician and Portuguese and found that monolingual ones seem to properly identify some agreement dependencies across intervening material such as relative clauses but struggle when dealing with others, like inflected infinitives.

More related to our project, Wei et al. (2021) trained BERT models for English controlling the training data, and found that word frequency during the learning phase influences the prediction performance of a

[2] `https://syntaxgym.org/`

model on subject-verb agreement dependencies (something which was previously suggested by Marvin and Linzen (2018)). However, to the best of our knowledge, there are no studies analyzing the impact of training time on the syntactic abilities of Transformer models, possibly because intermediate training checkpoints are not often available (although Sellam et al. (2022) just released several checkpoints at different training steps of BERT models for English).

This paper presents a detailed comparison between monolingual and multilingual BERT models for Galician. On the one hand, we explore the models' behaviour regarding linguistic properties of the test items, such as the length of the target dependency or the presence of attractors. On the other hand, we compare the models' performance taking into account several parameters, such as the number of hidden layers, the amount of training data, their initialization, or the number of their training steps.

3 Galician BERT models

In our experiments we compare the following monolingual and multilingual models:

- **mBERT**, which is the official multilingual BERT (base, with 12 layers).
- **Bertinho-small** (with 6 hidden layers) and **Bertinho-base** (with 12 hidden layers) published by Vilares, Garcia, and Gómez-Rodríguez (2021). These two models have a vocabulary of 30k tokens, and have been trained on the Galician Wikipedia (with about 45M words).
- **BERT-small** (6 hidden layers) and **BERT-base** (12 layers) released by Garcia (2021), both trained on a corpus of about 550M tokens. BERT-small has a vocabulary size of 52k tokens and has been trained during 1M steps. BERT-base has been initialized from mBERT (which includes Galician as on of its 102 languages). It has a vocabulary size of 119,547 and has been trained during 600 steps.

Additionally, we release the checkpoints of the latter two monolingual BERT models (BERT-small and BERT-base) mentioned above (Garcia, 2021). These models have been trained on a corpus which combines the Galician Wikipedia (April 2020 dump), SLI GalWeb (Agerri et al., 2018, composed by crawled texts from various domains), CC-100 (Wenzek et al., 2020), and other data crawled from online newspapers. It was semi-automatically cleaned by removing duplicate sentences, and utterances with many symbols and punctuations. The models were trained with a masked language modeling objective on a single Titan XP GPU (12GB), with batch sizes of 208 (small) and 198 (base), and using the *transformers* library (Wolf et al., 2020). For each model, we saved a checkpoint every 25k steps (about 12h and 26h for the small and base models respectively) up to 425k steps.[3] To avoid confusion with the originally published models BERT-small and BERT-base, these newly released checkpoints will be referred to as Check-small and Check-base.

4 Methodology

4.1 Research questions and experimental design

This work aims to explore the following research questions:

- **Q1:** Are Galician BERT LMs able to resolve agreement dependencies?
- **Q2:** Does model accuracy vary as a function of dependency type?
- **Q3:** Are Galician BERT LMs subject to interference effects from structurally-irrelevant attractor nouns?
- **Q4:** Does accuracy improve with training time?

To provide an (at least tentative) answer to these questions, we created a word prediction task that we run in the different models under evaluation (i.e. mBERT, Bertinho-small and base, BERT-small and base and the different checkpoints of Check-small and base. Our task had a factorial design which manipulated the type of dependency (noun-adjective vs. subject-verb agreement), the amount of intervening material (short vs. long) and the presence or absence of an intervening but structurally irrelevant noun that mismatches in agreement features with the head word (no attractor vs attractor). A

[3] All checkpoints are available at `https://github.com/marcospln/galician_bert_checkpoints`

dep	length	attr	example
noun-adj	short	no	O **neno** que xogaba onte alí é **alto**|*alta.
		yes	O **neno** que xogaba onte alí coa nena é **alto**|*alta.
	long	no	O **neno** que xogaba no parque inaugurado recentemente é **alto**|*alta.
		yes	O **neno** que xogaba no parque inaugurado recentemente coa nena é **alto**|*alta.
subj-verb	short	no	O **neno** que xogaba onte alí **aparece**|*aparecen na televisión.
		yes	O **neno** que xogaba onte alí cos outros nenos **aparece**|*aparecen na televisión.
	long	no	O **neno** que xogaba no parque inaugurado recentemente **aparece**|*aparecen na televisión.
		yes	O **neno** que xogaba no parque inaugurado recentemente cos outros nenos **aparece**|*aparecen na televisión.

Table 1: Sample set of the experimental conditions for noun-adjective and subject-verb agreement dependencies. *dep* is the target dependency, and *att* refers to the presence/absence of an attractor word. Words in bold are in a dependency relation, and underlined words are attractors, which agree with the wrong alternative marked with *. The base sentence (i.e. short without attractor) for noun-adjective dependencies means "The boy who was playing there yesterday is tall", and for subject-verb agreement dependencies means "The boy who was playing there yesterday appears on TV".

sample set of the experimental conditions is shown in Table 1.

We will pay particular attention to the models' accuracy for the experimental conditions at different steps in the training process by testing checkpoints at every 25k steps up to 425k for both Check-small and base. This will also allow us to investigate not only the effects of the training steps (on the various checkpoints) but also to make, among others, the following comparisons: (a) the impact of the training data and vocabulary size, comparing BERT-small and Bertinho-small; (b) the influence of the hidden layers, using Bertinho-base and Bertinho-small; (c) the effects of fine-tuning on monolingual data, comparing mBERT with our BERT-base (initialized from mBERT and fine-tuned in Galician). This inquiry is possible thanks to the public availability of the dataset described in the next section.

4.2 The dataset

In order to run the experiments described above, we have used a subset of the dataset released by Garcia and Crespo-Otero (2022)[4] —the first and only available dataset for targeted syntactic evaluation for Galician and Portuguese (number and gender) agreement dependencies. We limit ourselves to a subset of the dataset by choosing items with the structure of those in Table 1.

For noun-adjective agreement dependencies, the dataset contains 2,112 sentences. In order to avoid possible confounds, gender was conterbalanced so that half of the items had a feminine target and the other half a masculine one. For subject-verb agreement dependencies, the dataset contains 4,368 sentences. Similarly, number was counterbalanced so that half of the items had a singular target and the other half a plural one. It is worth noting that, overall, there are less experimental items without an attractor than those with an attractor. This is because the original dataset contained variants of the attractors to avoid potential lexical biases. Further details on the dataset building procedure are available in Garcia and Crespo-Otero (2022), but it must be noted that to create the dataset, the authors selected as target (masked) words only those forms appearing in the vocabulary of the (monolingual and multilingual) models in order to allow the evaluation of all models using the same number of experimental items.

4.3 Evaluation metrics

Before moving into the results of the experiments, some notes on evaluation procedures

[4] https://github.com/crespoalfredo/PROPOR2022-gl-pt

Sentence				
As **nenas** que xogaban onte alí co outro neno *ten|**teñen** fame. *'The girls who were playing there yesterday with the other boy *is|are hungry.'*				
Model	**Prob Corr**	**Prob Wrong**	**0/1 accuracy**	**PD accuracy**
mBERT	0.0007	0.0005	1	0.236
Bertinho-base	0.1856	0.0030	1	0.968

Table 2: Example of a sentence where both models (mBERT and Bertinho-base) give higher probability to the correct inflection (*Prob Corr* vs *Prob Wrong*).

deserve mentioning. Contrary to the standard evaluation procedure which considers that a model succeeds if it assigns a higher probability to the correct than to the incorrect form (assigning thus either a 0 or a 1), Newman et al. (2021) propose the use of a large set of correct/incorrect pairs on each sentence to probe a model. In this vein, they measure the model's *systematicity* (in how many pairs per sentence the model prefers the correct alternative) and *likely behaviour* (the probability of generating a correct inflection). In this method, large lists of verbs are gathered using corpus frequencies, which are then used to replace the original pairs of each context. Hence, using this strategy involves the evaluation on many potentially semantically infelicitous (or implausible) sentences.

By contrast, we put the focus on the probability distance between the correct and the incorrect alternatives provided by the dataset, and propose a new metric dubbed *Probability Distance* (PD accuracy). It is calculated by normalizing the probabilities of the correct and incorrect targets (e.g. *teñen* vs *ten*) via percentages, and then substracting the percentage of the incorrect form from the percentage of the correct one. The rationale behind this metric is motivated by the observation that traditional binary metrics obscure possible effects because subtle differences such that between 0.51 vs 0.49 would be immediately translated to 1 or 0 (thus obtaining the same accuracy as for large differences such as 0.85 vs 0.15). Instead, PD accuracies are circumscribed to a probability space which only includes a semantically plausible target pair, and accuracy is calculated within that narrowed probability space, keeping the original distance between correct and incorrect words. To demonstrate this, Table 2 shows the results for a subject-verb long agreement dependency with an attractor for which mBERT and Bertinho-base give a higher probability for the correct inflection. Nonetheless, while 0/1 accuracy focuses on the most likely alternative, and would thus assign a 1 in both cases, PD accuracy brings the distance between the correct and incorrect alternatives to the fore. To further demonstrate this, Figure 1 shows the mean accuracy for BERT-base's long sentences with an intervening attractor (i.e. examples such as the one in Table 2). We are only showing BERT-base results for long sentences with an attractor for the sake of simplicity, as these are the cases where models tend to fail. What can be observed here is that binary metrics clearly overestimate the systematicity of language models, as accuracy drops once PD accuracy is calculated. Critically, PD accuracy also provides a better threshold for comparison between models with different vocabulary sizes.

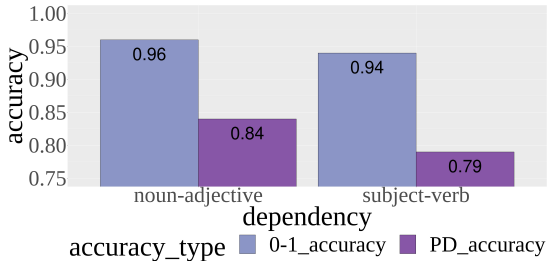


Figure 1: Mean accuracy for BERT-base's long sentences with an intervening attractor using 0-1 and PD accuracy metrics.

5 Results and Discussion

We will first report and discuss the accuracy provided by the five available models and then, we will focus on the results for Check-small and Check-base at different training checkpoints.

5.1 Published models

Overall accuracy: Figure 2 shows the overall accuracy for each model for noun-

adjective and subject-verb agreement dependencies, regardless of dependency length of the presence/absence of an attractor. Several ideas related to our research questions can be tackled at this point: first, there is a decline in accuracy for the monolingual models (BERT-base>BERT-small>Bertinho-base>Bertinho-small), suggesting that the amount of training data heavily influences the models' performance. Interestingly, mBERT's results resemble those from Bertinho-base, most possibly because they have been trained with the same Galician data (Wikipedia) and have a similar architecture (same number of hidden layers and dimensionality). BERT-base and BERT-small show a relatively acceptable performance, while the other three models (Bertinho-base, Bertinho-small and mBERT) are closer to chance performance (see Q1). This is particularly true for subject-verb agreement dependencies, while all models except Bertinho-small provide better results for noun-adjective dependencies (see Q2). In line with the results of Goldberg (2019) for English, Bertinho-small obtained slightly better accuracy than its base variant on subject-verb agreement.

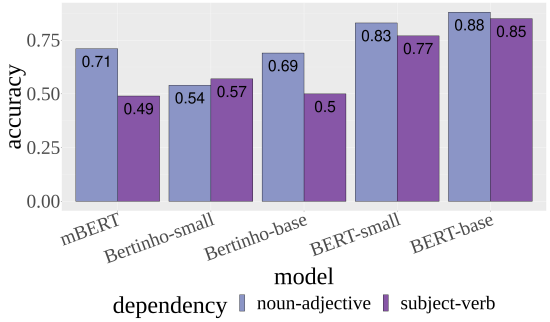


Figure 2: Mean accuracy by dependency type for the five models under investigation.

Accuracy per condition: Figure 3 provides a closer picture of the models' performance for noun-adjective and subject-verb dependencies when looking at the four different experimental conditions. These results tap directly into the possible presence of agreement attraction effects (see Q3). Based on previous studies, out of the four experimental conditions, short sentences with no attractor were predicted to be the easiest ones, while long sentences with an attractor were predicted to be the hardest ones. This was borne out in the results, confirming that Galician BERT models are lured by structurally irrelevant mismatching nouns that intervene in agreement dependencies between the head and the target. Critically, the emergence of attraction effects is mediated by the distance between the head and the target such that longer dependencies are more prone to give rise to attraction effects. Nonetheless, it must be noted that not all models show equally strong attraction effects (aligning with the accuracy decline described above), and that attraction effects are steeper in subject-verb agreement dependencies than in noun-adjective agreement dependencies — something which was foreseeable on the basis of Figure 2.

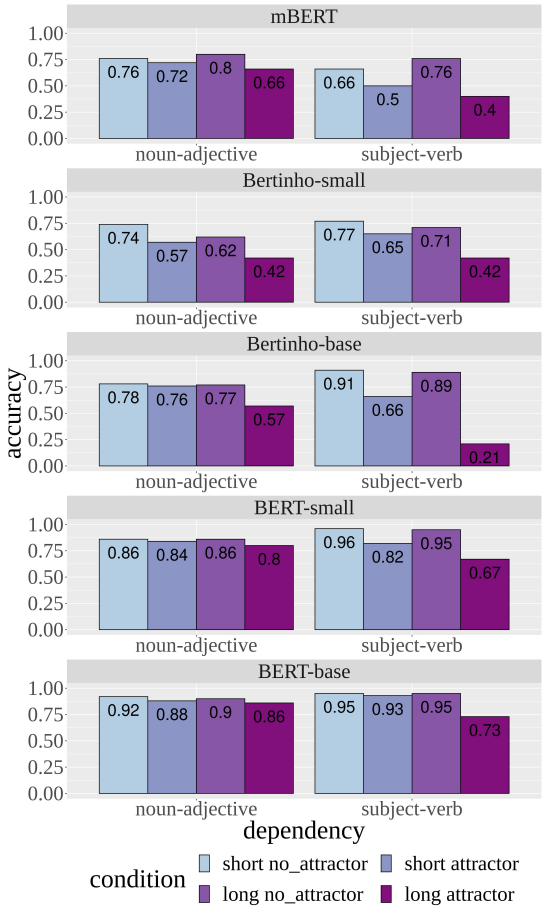


Figure 3: Mean accuracy by dependency type and experimental condition for the five models under investigation.

5.2 Learning curves

Overall accuracy: Moving now into the analyses by training checkpoints, Figure 4 shows the overall accuracy for Check-small and Check-base for the two dependencies at every checkpoint, from 25k to 425k steps.

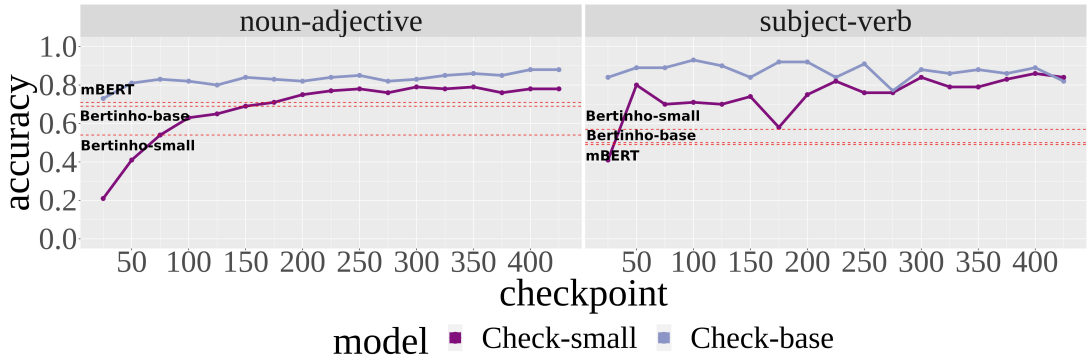


Figure 4: Mean accuracy by dependency type and checkpoint for Check-base and Check-small. The horizontal red lines indicate the overall accuracy mean for Bertinho-base, Bertinho-small and mBERT for ease of comparison (cf. Figure 2).

Check-small: Focusing on the small model (with 6 layers and trained from scratch) represented by the dark line, the results show that it needs relatively few checkpoints to surpass the average accuracy by Bertinho and mBERT. On noun-adjective dependencies, Bertinho-small is surpassed at checkpoint 75k, while it needs between 150k and 175k steps to outperform Bertinho-base and mBERT, both with 12 layers. This is even clearer on subject-verb dependencies, as the second checkpoint (75k) already shows better results than any of the three mentioned models. When comparing with the results of the published BERT-base and BERT-small (see Figure 2), it is worth noting that these models do not obtain notoriously better results even though they have been trained for a longer period of time. These results suggest that the amount of (monolingual) training data, rather than training time, is crucial to generalize the target dependencies, as Check-small obtains better results at checkpoint 75k than Bertinho-base at 1.5M training steps.

Check-base: Moving now into Check-base, represented by the dark line, it should be reminded that this model has been initialized with the weights of mBERT, and at the first checkpoint (25k) it already obtains comparable results to that of the final BERT-base model on subject-verb agreement (see Figure 2). On noun-adjective dependencies, the model keeps a more constant learning rate, but it achieves similar performance than BERT-base at around 400k steps. In this case, we may hypothesize that the model is taking advantage of the linguistic properties of other languages covered by mBERT, and it adapts the model to Galician on early steps. Contrarily to the constant learning rate observed for noun-adjective dependencies, the panorama learning curve for subject-verb dependencies seems to be much more unstable. Indeed, no improvement is observed for neither Check-base nor Check-small. Although accuracy improves with time, subject-verb agreement dependencies experience ups and downs in intermediate checkpoints.

Accuracy per condition: Finally, Figure 5 shows the curves for each experimental condition for Check-small and Check-base at the different training steps. As expected, short contexts and sentences without attractors are easily solved by both models on the two dependencies. As previously shown (Figure 4), Check-base overtakes the performance of mBERT on the first checkpoints, but then the raise of the learning curve is very small, namely for subject-verb agreement. The small variant, especially on noun-adjective dependencies, shows a constant learning process which seems to stabilize around checkpoint 300k. Interestingly, subject-verb agreement dependencies are easily solved by Check-small at 50k steps in the absence of attractors. However, when attractors are present, Check-small is sensitive to them even in short contexts, where it preserves the performance again about checkpoint 300k. Finally, on long-distance dependencies with attractors, none of the models seem to have a stable behaviour, as the performance of both of them varies unpredictably. This is more noticeable for Check-base, which solves most cases properly but struggles with the more complex structures (i.e. long sentences with attractors).

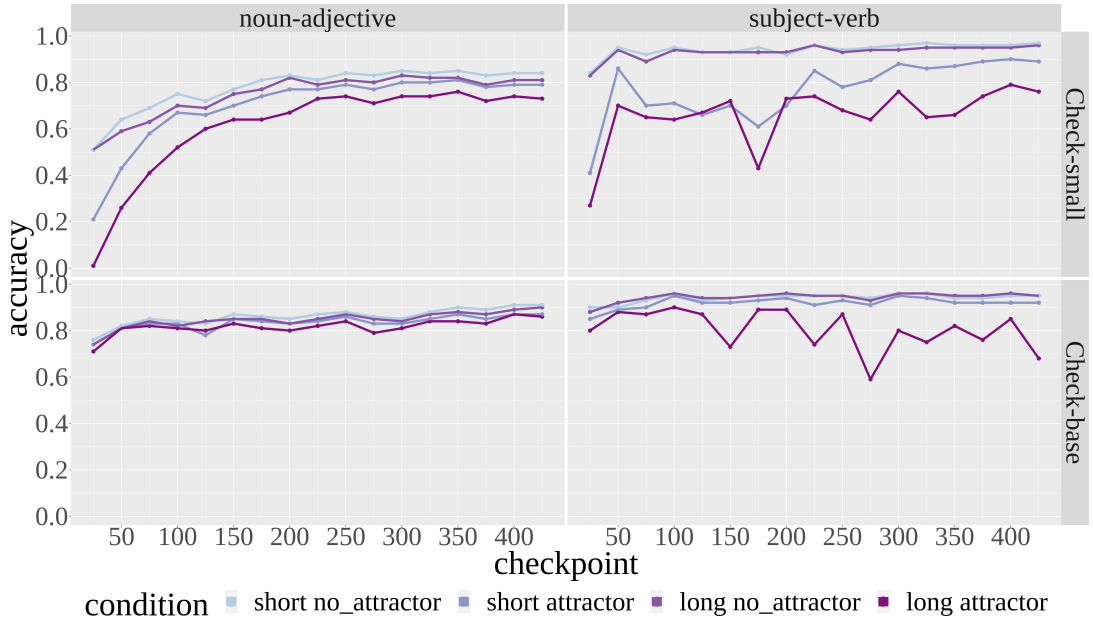


Figure 5: Mean accuracy of Check-small (top) and Check-base (bottom) by dependency type, experimental condition, and checkpoint.

6 Conclusions and future work

This paper has presented a multidimensional evaluation of a variety of BERT models for Galician on two types of agreement dependencies, noun-adjective and subject-verb. We compared the performance of multilingual and monolingual models with diverse properties, including 6 and 12 layers variants, different sizes of training data and vocabularies, and two initializations: training from scratch, and fine-tuning a multilingual BERT on Galician data.

Our results show a gradient in the ability of Galician BERT LMs models to resolve agreement dependencies, with BERT-base being the most accurate and Bertinho-small the least accurate. Furthermore, we observed that accuracy varied as a function of dependency type, with noun-adjective agreement dependencies being easier to handle than subject-verb agreement dependencies. Interestingly, BERT LMs are subject to interference effects from structurally-irrelevant attractor nouns, and the degree of fallibility to attraction effects is inverse to accuracy (i.e. less accurate models show more attraction effects). Last and most important, although training time does seem to have a small effect on the models' accuracy, this factor is far from being comparable with the influence of the size of the training corpus.

Besides the results and analyses of the performed experiments, we contribute with new 34 checkpoints of BERT models for an understudied language, Galician, which are freely released with this paper and can hopefully contribute to foster research on languages different from English.

This exploratory work has opened many lines of inquiry that we aim to explore in future research. On the one hand, we plan to create new datasets in Galician that do not only overcome some shortcomings observed in the one released by Garcia and Crespo-Otero (2022) but also incorporate new types of linguistic relations. On the other hand, we plan to compare the results obtained for Galician with other languages in order to observe cross-linguistic differences and similarities.

Acknowledgements

This research was funded by the project "Nós: Galician in the society and economy of artificial intelligence" (Xunta de Galicia/Universidade de Santiago de Compostela), by grant ED431G2019/04 (Galician Government and ERDF), by a *Ramón y Cajal* grant (RYC2019-028473-I), and by Grant ED431F 2021/01 (Galician Government).

References

Agerri, R., X. Gómez Guinovart, G. Rigau, and M. A. Solla Portela. 2018. Developing new linguistic resources and tools

for the Galician language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Bernardy, J.-P. and S. Lappin. 2017. Using deep neural networks to learn syntactic agreement. In *Linguistic Issues in Language Technology, Volume 15, 2017*. CSLI Publications.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Garcia, M. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3625–3640, Online, August. Association for Computational Linguistics.

Garcia, M. and A. Crespo-Otero. 2022. A Targeted Assessment of the Syntactic Abilities of Transformer Models for Galician-Portuguese. In *International Conference on Computational Processing of the Portuguese Language (PROPOR 2022)*, pages 46–56. Springer.

Gauthier, J., J. Hu, E. Wilcox, P. Qian, and R. Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online, July. Association for Computational Linguistics.

Goldberg, Y. 2019. Assessing BERT's Syntactic Abilities. arXiv preprint arXiv:1901.05287.

Gulordava, K., P. Bojanowski, E. Grave, T. Linzen, and M. Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana, June. Association for Computational Linguistics.

Henderson, J. 2020. The unstoppable rise of computational linguistics in deep learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6294–6306, Online, July. Association for Computational Linguistics.

Hewitt, J. and C. D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Kuncoro, A., C. Dyer, J. Hale, and P. Blunsom. 2018a. The perils of natural behaviour tests for unnatural models: the case of number agreement. *Learning Language in Humans and in Machines*, 5(6). https://osf.io/9usyt/.

Kuncoro, A., C. Dyer, J. Hale, D. Yogatama, S. Clark, and P. Blunsom. 2018b. LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Melbourne, Australia, July. Association for Computational Linguistics.

Lakretz, Y., G. Kruszewski, T. Desbordes, D. Hupkes, S. Dehaene, and M. Baroni. 2019. The emergence of number and syntax units in LSTM language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 11–20, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Lin, Y., Y. C. Tan, and R. Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy, August. Association for Computational Linguistics.

Linzen, T., E. Dupoux, and Y. Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Linzen, T. and B. Leonard. 2018. Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. arXiv preprint arXiv:1807.06882.

Marvin, R. and T. Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium, October-November. Association for Computational Linguistics.

Mueller, A., G. Nicolai, P. Petrou-Zeniou, N. Talmina, and T. Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online, July. Association for Computational Linguistics.

Newman, B., K.-S. Ang, J. Gong, and J. Hewitt. 2021. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online, June. Association for Computational Linguistics.

Pérez-Mayos, L., M. Ballesteros, and L. Wanner. 2021. How much pretraining data do language models need to learn syntax? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1571–1582, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Pérez-Mayos, L., A. Táboas García, S. Mille, and L. Wanner. 2021. Assessing the syntactic capabilities of transformer-based multilingual language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3799–3812, Online, August. Association for Computational Linguistics.

Sellam, T., S. Yadlowsky, J. Wei, N. Saphra, A. D'Amour, T. Linzen, J. Bastings, I. Turc, J. Eisenstein, D. Das, I. Tenney, and E. Pavlick. 2022. The MultiBERTs: BERT Reproductions for Robustness Analysis. In *The Tenth International Conference on Learning Representations (ICLR 2022)*. arXiv preprint arXiv:2106.16163.

Tran, K., A. Bisazza, and C. Monz. 2018. The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736, Brussels, Belgium, October-November. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. arXiv preprint arXiv:1706.03762.

Vilares, D., M. Garcia, and C. Gómez-Rodríguez. 2021. Bertinho: Galician BERT Representations. *Procesamiento del Lenguaje Natural*, 66:13–26.

Wei, J., D. Garrette, T. Linzen, and E. Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Wenzek, G., M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May. European Language Resources Association.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 27-38 recibido 30-03-2022 revisado 20-05-2022 aceptado 23-05-2022

Information fusion for mental disorders detection: multimodal BERT against fusioning multiple BERTs

Fusión de información para detección de transtornos mentales: BERT multimodal contra múltiples BERTs fusionados

Mario Ezra Aragón[1], A. Pastor López-Monroy[2], Luis C. González-Gurrola[3], Manuel Montes-y-Gómez[1]

[1]Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico
[2]Centro de Investigación en Matemáticas A.C., Guanajuato, Mexico
[3]Universidad Autónoma de Chihuahua, Chihuahua, Mexico
{mearagon,mmontesg}@inaoep.mx, pastor.lopez@cimat.mx, lcgonzalez@uach.mx

Abstract: Given the increasing number of modalities that modern classification problems provide, recently a multimodal BERT transformer (MMBT) was proposed. An interesting opportunity to evaluate the effectiveness of such model is posed by the problem of timely detection of mental disorders of social media users. For this problem, a multi-channel perspective involves extracting from each user post different types of information, such as thematic, emotional and stylistic content. This study evaluates the suitability of tackling this problem by the apparently ad-hoc MMBT, moreover, we further evaluate if regular BERT models could be combined or fused in such a way that could have a chance in a multi-channel arena. For the evaluation, we use recent public data sets for three important mental disorders: Depression, Anorexia, and Self-harm. Results suggest that BERT models can get on their own a data representation that could be later fusioned and boost the classification performance by at least 5% in F1 measure, even surpassing the MMBT.
Keywords: Multichannel information, Transformers, Mental disorders.

Resumen: Dado el creciente número de modalidades que ofrecen los problemas de clasificación modernos, recientemente se ha propuesto un transformer BERT multimodal (MMBT). Una oportunidad interesante para evaluar la eficacia de dicho modelo la plantea el problema de la detección oportuna de los trastornos mentales de usuarios de las redes sociales. Para este problema, una perspectiva multicanal implica extraer de cada post de los usuarios diferentes tipos de información, como su contenido temático, emocional y estilístico. Este estudio evalúa la idoneidad de abordar este problema mediante el aparentemente ad-hoc MMBT, además, evaluamos si los modelos BERT regulares podrían combinarse o fusionarse de tal manera que pudieran tener una oportunidad en un escenario multicanal. Para la evaluación, utilizamos conjuntos de datos públicos recientes para tres importantes trastornos mentales: Depresión, Anorexia y Autolesiones. Los resultados sugieren que los modelos BERT pueden obtener por sí solos una representación de los datos que podría fusionarse posteriormente y aumentar el rendimiento de la clasificación en al menos un 5% en la medida F1, superando incluso al MMBT.
Palabras clave: Multicanal, Transformers, Trastornos Mentales.

1 Introduction

Over the last few years, millions of people around the world have been affected by one or more mental disorders, for example, in 2018 a study of mental disorders in Mexico reveals that 17% of people in the country have at least one mental disorder and one in four will suffer a mental disorder at least

ISSN 1135-5948. DOI 10.26342/2022-69-2

once in their life (Renteria-Rodriguez, 2018). Unfortunately, this phenomenon causes interference in their daily life, especially affecting their behavior and thinking. Regularly, the self-awareness of having a mental disorder causes emotional and physical damage that could make people feel fear to the idea of being vulnerable to criticism, judgment, or opposing opinions. Mental disorders may be related to a particular event that generated excessive stress on the affected person or to a series of different stressful events (World Health Organization, 2019). For instance, some of the causes are environmental stress, genetic factors, or different difficult life situations. There are several common mental disorders such as depression, anorexia, or self-harm affecting people worldwide (Kessler et al., 2017).

A reality nowadays, is that for some people their social life does not occur in their surroundings or immediate environment, but takes place in a virtual world created by social media platforms like Facebook, Twitter, or Reddit (Baer, 2021). In other words, social media has become a vital link for some of us. This scenario presents opportunities to study and analyze, given the availability of data, how people communicate, and more specifically, how this communication could be associated to possible mental health issues that people are experimenting, contributing in this way to attract attention to silent disorders.

Previous studies have shown that texts shared by users in social networks have different evidence or types (channels or pseudo modalities) of information that may be relevant for the detection of mental disorders (Guntuku et al., 2017), for example, their topics of interest, emotional state, or their writing style. This has motivated us to propose a method that considers these pieces of information and to study how to combine or fusion them. With this in mind, we evaluate the plausibility of using the multimodal BERT (MMBT)(Kiela et al., 2019) for this task, being this the first time. In addition, we state the question: whether MMBT is the best option to handle this sensitive task or if multiple BERTs can better exploit the nature of the data. Thus, we compare the performance of MMBT against different architectures based on early and late fusion approaches of the popular BERT model. In this study we take the text modality and divide it into three channels[1] that focus on different aspects of the users' communication. The first channel captures the thematic information used for contextual analysis. The second channel indicates the manifested emotions, attempting to capture emotional topics related to mental disorders. Finally, the third channel focuses on the writing style, where we want to capture the use of personal expressions and verbs tense, among other aspects. As could be observed, our hypothesis is that people that present some mental disorder tend to express differently, at different dimensions, regarding the control group.

It is widely known that the BERT model (Devlin et al., 2019) has led to important improvements in representation learning for natural language processing and text classification problems. In a recent work (Kiela et al., 2019), the authors demonstrate that supervised bidirectional transformers with unimodal pre-trained components obtain good performance in multimodal fusion. They found that learning to map dense multimodal features to BERT's token embedding space is easy to extend to different modalities. Inspired by their findings, we adapt their Multimodal BERT (MMBT) module with our channels as modalities to create a multichannel contextualized representation. The main idea of our work is to find out the best way to combine the different types of information and see if using multimodal BERT is better than considering a fusion of multiple BERTs, each one specialized in a different channel.

We can summarize the contributions of our work as follows:

1. We adapt a Multimodal BERT (MMBT) for the detection of mental disorders, considering three channels of information: thematic, emotional and stylistic.
2. We explore different strategies to combine these information, considering early and late fusion approaches.
3. We analyze and evaluate in detail these three information channels and the importance of their fusion, then concluding about its feasibility of integration to boost classification performance.

[1]In this work, we define a *channel* as a different property or view from the same modality (Qianli et al., 2017).

2 Related work

2.1 Mental disorders detection

In the last few years, the study of public mental health through social media has increased. This is mainly because these media provide a source of support for those who suffer from a mental health disorder, like for example, a sense of community, relatedness, and understanding (Hilton, 2016; Dyson et al., 2016). In general, for the construction of corpora, researchers identify a group of users who expressed in one of their publications having been clinically diagnosed with a mental disorder and then download all or part of their posts (De Choudhury, Counts, and Horvitz, 2013; Wang et al., 2017).

Recent works (Trifan and Oliveira, 2019; Van Rijen et al., 2019), explored the analysis of the posts' content. In these works, the authors consider different features, such as word and char n-grams, and then apply a classification algorithm to make a decision. The shortcoming with that strategy is the high overlap in the vocabulary used by the control and positive users, which generates several missclassifications. Another well-known strategy consists of counting the number of occurrences of positive, negative, and neutral words in texts (Kang, Yoon, and Kim, 2016), or on measuring how similar are their words to some reference negative and positive lexicons (Htait, Fournier, and Bellot, 2017). On the other hand, analyzing sentiments has shown interesting results since it has been found that negative comments are more abundant in people with a declared mental health disorder than in comments generated from a control group (Coopersmith, Dredze, and Harman, 2014; Preotiuc-Pietro et al., 2015). Other works have used a LIWC-based representation (Tausczik and Pennebaker, 2010), which consists of a set of psychological categories that aim to represent users' posts by features of social relationships, thinking styles and individual differences (Coppersmith et al., 2015).

2.2 Fusion approaches for mental disorders detection

How to effectively combine information is challenging and has a long history in machine learning (Baltrusaitis, Ahuja, and Morency, 2019). In particular, some recent works on mental disorders detection have considered the use of ensemble approaches to combine bag of words representations, LIWC features, and different deep neural models (Trotzek, Koitka, and Friedrich, 2018). In (Ragheb et al., 2019), the authors combine the temporal mood variation and Bayesian inference for their detection. The first phase uses an attention-based deep model to construct a representation for the mood variation. Then, in the second phase, the model uses Bayesian inference to detect clear signs of mental disorders and then give a decision based on their combination. In (Ji et al., 2020), the authors apply an attention model combined with sentiment and topic analysis to detect suicidal ideation. A recent work (Uban, Chulvi, and Rosso, 2021), explored the evolution of emotion expression in relation to cognitive styles and found specific patterns in users with mental disorders.

The performance shown by ensemble approaches suggests the suitability of adapting advanced techniques to fusion information from different channels in a more effective way. For that reason, we decided to implement our multi-channel BERT-based approach as a new way to combine the thematic, emotional and stylistic views of the information shared by social media users. This proposed model takes inspiration from multimodal BERT (Kiela et al., 2019) in order to have an adapted version that is able to model individual channels. As our experimental evaluation will show, the proposed strategy improves the classification performance.

3 Information channels representation

The fusion strategies that we are going to explore are implemented on the basis of BERT, therefore, the three channels of information are captured by three different representations of the words. The main idea of this approach is to have different views of the content of the users' posts. We achieve this by generating *three embeddings* for each word in the posts, capturing or emphasizing the thematic, emotional and style information, respectively. In the following subsections, we briefly describe how to generate these three representations.

3.1 Thematic embeddings

In order to capture the thematic content related to each word, we consider vanilla GloVe embeddings (Pennington, Socher, and Man-

ning, 2014). However, since this type of embeddings does not take into account the context of the words, we decided to also use some contextualized word embeddings, in particular the BERT embeddings (Devlin et al., 2019). For our experiments, we used both separately and evaluated which one contributes the most to the final representation.

With contextualized embeddings, words that have similar meaning or show some semantic relation are closer to each other. For example, the words "insecure" and "worried" have similar embeddings, as do the words "therapy" and "treatment".

3.2 Emotion-based embeddings

For this work, we use the emotion-based word embeddings that were originally proposed in (Aragon et al., 2019). In short, to construct these vectors, first, we generated groups of fine-grained emotions for each general emotion that belong to the EmoLEX lexicon (Mohammad and Turney, 2013). We achieved this by representing each word of the lexicon with its FastText embedding (Bojanowski et al., 2016) and then applying a clustering algorithm on them. After obtaining the fine-grained emotions, which are groups of words that capture specific topics related to the same emotion, we represented each of them by means of the average vector of its words. Subsequently, and as the last step, for each word in the vocabulary we measured its cosine similarity with all fine-grained emotions, and assigned to each one of the embedding from its closest fine-grained emotion.

According to the process described above, the words "accident" and "crash" will have the same embedding because both belong to the same subgroup of the Surprise emotion, whereas the word "magician" will have a slightly different embedding since it corresponds to a different subgroup of the same emotion. On the other hand, the words "accomplish" and "achieve" will have a completely different embedding as they belong to the Joy emotion.

3.3 Style-based embeddings

The third representation of the words aims to capture particular characteristics of the writing style of social media users. Its idea is to capture how users with mental disorders tend to talk, for example, referring to past events or to uncertainties about the future. To capture the stylistic information, we propose a new word representation inspired by the successful use of character n-grams in author profiling tasks.

To define the style-based embedding of each word from the users' posts, we carried out the following process:

1. Divide the word into character 3-grams.
2. Compute, for each 3-gram, its embedding using FastText (Bojanowski et al., 2016), as well as its discriminative score according to its chi^2 distribution in the two given classes (positive and control users).
3. Obtain the embedding vector of the word by applying a weighted sum of the vectors of its character 3-grams, considering as weights their chi^2 values.

Take for example the word "depression", its style-based embedding is obtained by the weighted sum of the vectors corresponding to its character 3-grams "dep", "epr", ..., "ion". It is important to notice that the style-based embeddings are similar for words that have similar spelling rather than meaning. For example, words in superlative resemble each other, as well as regular verbs in past tense, or words with the same root. Take for instance the word "mental" some of their closest words would be "dental", "mentality" or "decremental".

4 On the fusion of the three channels

The objective of our work is to compare different ways of combining information from different channels. This is a key stage in the classification process, and have the intuitive idea of learning the relevance of each channel in an automatic way. We use two main strategies, firstly, one based on multimodal BERT whose idea is to learn a joint representation of the three types of information, and secondly, different architectures that treat each channel separately and apply different early and late fusion techniques.

4.1 Multimodal BERT (MMBT)

It is a recently proposed supervised multimodal bitransformer model for classifying images and text (Kiela et al., 2019). The MMBT model starts with pre-trained BERT

weights, and takes their contextual embeddings as input. These contextual embeddings are obtained as the sum of the segment, position, and token embeddings of each word. Then, the model weights them and project each of the embeddings to a token input. In Figure 1, we can appreciate the components of the architecture of the MMBT model. Although proposed for only two modalities, this architecture can be generalized to any number of modalities, assigning a different segment id to each of them.

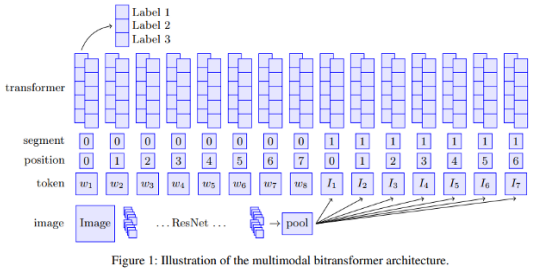


Figure 1: Illustration of the multimodal bitransformer architecture.

Figure 1: Multimodal bitransformer architecture proposed in (Kiela et al., 2019).

In the original work, the authors took contextual embeddings as input, learned the weights, and projected each image's embeddings to a dimensional token input. Instead of image embeddings, we used the sequence of the words for the fine-tuning with our different channels as embeddings. Once we fine-tuned the model with the channels, we took advantage of the contextual information learned for classifying the users' posts. For this purpose, we used the first output of the final MMBT layer as input to a Convolutional Neural Network (CNN) for feature extraction, and then add a dense layer for achieving the classification. Figure 2 presents the general architecture of this approach.

4.2 Combining information using multiple BERTs

The authors of MMBT (Kiela et al., 2019) noted that their method is compatible with scenarios where not every modality is present and can be generalized to an arbitrary number of modalities. Then, for the second model, we decided to train individual BERTs and fine-tuning them with each channel separately. After the training, similar to the first approach, we used the first output of the final layer of each BERT and concatenate them as input for a CNN layer, we will refer to

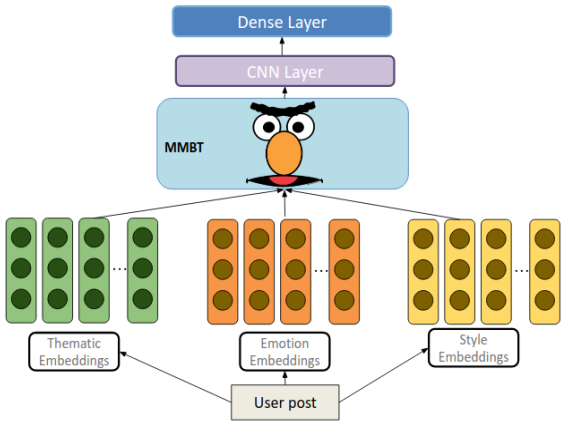


Figure 2: General diagram of the Model 1: Multimodal BERT with vectors of three channels, then a CNN layer, and a classification layer.

this model as BERT-CNN. In Figure 3, we present the general diagram for this process.

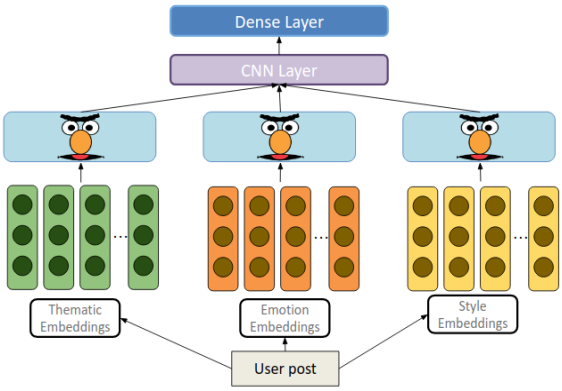


Figure 3: General diagram of the BERT-CNN model: Each channel separately enters to a BERT model, then join vectors feed a single CNN layer, and a classification layer.

For the third model, instead of concatenating the vectors and using a single CNN layer, we separate them into different convolutional layers and used the output for a dense layer. With this approach, the model obtains for each channel different feature maps of each region and concatenates them together to form a single feature vector. This can be interpreted as summarizing the local information to find patterns, and then combining the information. The hypothesis that the local information per channel is important and should be extracted before it is combined, we call this model BERT-3CNN. Figure 4 presents the general diagram for this model.

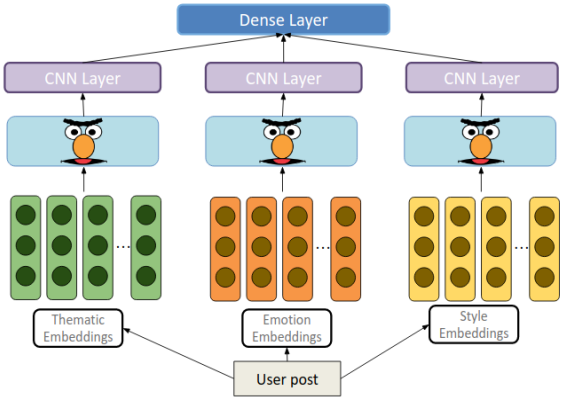


Figure 4: General diagram of the BERT-3CNN model: Each channel separately enters a BERT model and a CNN layer, then their outputs are concatenated and fed to a single classifier.

One of the challenges of this work is the problem of fusing information. A simple solution is to concatenate the representations of each channel into one vector or perform an operation like adding or taking the product. However, the use of these operations assumes that all channels have the same relevance, which is usually not the case. In recent work (Arevalo et al., 2019), the authors proposed a novel type of hidden unit called Gated Multimodal Unit (GMU). This unit works similarly to the control flow mechanism in gated recurrent units. The gates in the unit let the model regulate the flow of information into the next one. Figure 5 presents a general overview of the GMU module we used, where the x_i inputs represent the feature vectors associated with each modality, and the z_i weights indicate their relevance.

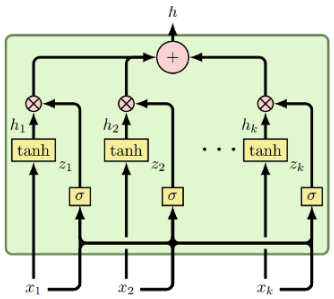


Figure 5: Overview of GMU module (Arevalo et al., 2019). Where x_i represents the ith input modality. The final fused representation of all modalities is represented by h at the top.

Motivated by the outstanding results of the GMU module in different multimodal tasks, our fourth model takes advantage of it. That is, after the feature extraction, we implement a Gated Multimodal Unit (GMU) module to learn the relations between each channel feature vector. Then, apply a dense layer to classify the final vector with the information of the three channels, we call this model BERT-GMU. Figure 6 describes the process for this model.

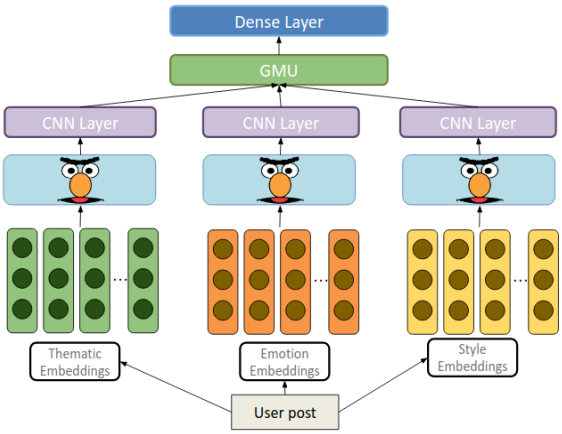


Figure 6: General diagram of the BERT-GMU model: Each channel separately enters a BERT model and a CNN layer, then their outputs are combined by a GMU module, and fed to a single classifier.

5 *Experimental settings*

5.1 Data sets

We performed experiments over data sets from the eRisk 2019 and 2020 evaluation tasks (Losada, Crestani, and Parapar, 2019; Losada, Crestani, and Parapar, 2020). These data sets consist of the detection of depression, anorexia, and self-harm, and contain the post history of several users from the Reddit platform. For each mental disorder, we have two types of users: 1) the control group, people collected who do not suffer from any mental disorder; and 2) positive users, a group composed of people affected by either depression, anorexia, or self-harm.

In the tasks of anorexia and self-harm, the positive class is composed of users who explicitly mentioned that they were diagnosed by a medical specialist or that they had committed self-harm. On the other hand, the control class for both tasks is composed of random users from the Reddit platform. However, to add realism to the data sets and make the

detection of positive users challenging, the control group also contains users who often interact in the threads of anorexia and self-harm.

For depression, the organizers of the shared task asked the participants to predict, for each user, the possible answer to each input of the BDI questionnaire (Beck et al., 1961), which contains 21 questions that allow assessing the level of severity of the depression. In contrast to them, in this work we exclusively consider a binary prediction task, i.e., to distinguish between positive and control users. In particular, the positive class is composed of users that obtained 21 points or more in the final result of the questionnaire (presence of moderate or severe depression), whereas the control class is formed by the rest of the users, having 20 points or fewer in their final result.

Table 1 shows how classes distribute within these data sets as well as some general information regarding the collections. For the depression task, we used for training the data set from eRisk 2018 (Losada, Crestani, and Parapar, 2018), this data set was constructed similarly to anorexia and self-harm data sets.

Data set	**Train**		**Test**	
	P	C	P	C
Anorexia'19	61	411	73	742
avg. num. posts	407.8	556.9	241.4	745.1
avg num. words	37.3	20.9	37.2	21.7
avg. days	800	650	510	930
Depression'20	214	1493	40	49
avg. num. posts	440.9	660.8	493.0	543.7
avg num. words	27.5	22.75	39.2	45.6
avg. days	686	663	642	1015
Self-harm'20	41	299	104	319
avg. num. posts	169.0	546.8	112.4	285.6
avg num. words	24.8	18.8	21.4	11.9
avg. days	495	500	270	426

Table 1: Data sets used for experimentation, where P indicates the positive users and C is used for control users.

5.2 Preprocessing

The texts were normalized by lowercasing all words and removing special characters like URLs, emoticons, and #; the stopwords were kept. Our decision to keep stopwords was completely experimental, we performed experiments removing them before masking the texts, but consistently we got slightly lower performances.

5.3 Classification

The main goal is to classify users into one of the two classes (Depressed / Control, Anorexia / Control, or Self-harm / Control). We separate each post history into N parts. We select the N value empirically, testing recommended sizes of sequences in the literature, i.e., $N = \{25, 35, 50, 100\}$. For training, we process each part of the post history as an individual input and train the model. For the test, each part receives a label of 1 or 0; then, if the majority of the posts are positive, the user is classified as showing a mental disorder. The main idea is to consistently detect the presence of major signs of depression, anorexia, or self-harm through all the user posts.

5.4 Baselines

The results are compared to the traditional Bag-of-Words representation combined with a SVM classifier. This representation was created using word unigrams and n-grams; these are common baseline approaches for text classification. For both approaches, we selected the same number of features using tf-idf representation and chi^2 distribution X_k^2. We also add some baselines based on deep learning approaches, using a CNN and a Bi-LSTM. The neural networks used 100 neurons, an adam optimizer, and word2vec and Glove embeddings with a dimension of 300. For the CNN we use 100 random filters of sizes 1, 2, and 3 (parameters recommended in literature). We also add a BERT model with a fine-tuning over the training data set. Additionally, the obtained results are compared against the top-three participants of the eRisk evaluation tasks. For all these comparisons, we considered the F_1 score, precision, and recall over the positive class (Losada, Crestani, and Parapar, 2018).

6 Evaluation and Analysis

For the evaluation, besides the baselines, we also performed experiments using independently the proposed thematic, emotion and style representations. Table 2 presents the results in terms of F_1 score, precision, and recall over the positive class to detect Anorexia (eRisk'19), Depression (eRisk'20) and Self-harm (eRisk'20). We organize the results in three groups: baseline methods, our proposal but limited to only one channel, and our proposal using all information channels

Method	Anor			Dep			SH		
Baselines									
	F1	P	R	F1	P	R	F1	P	R
BoW-unigrams	0.67	0.85	0.55	0.58	0.56	0.60	0.50	0.95	0.34
BoW-Ngrams	0.66	0.83	0.55	0.57	0.55	0.59	0.50	0.92	0.33
Bag of char 3grams	0.67	0.85	0.55	0.58	0.56	0.60	0.52	0.97	0.36
RNN-word2vec	0.65	0.95	0.49	0.57	0.62	0.53	0.55	0.60	0.51
CNN-word2vec	0.66	0.94	0.52	0.60	0.57	0.62	0.56	0.54	0.59
RNN-GloVe	0.65	0.92	0.51	0.58	0.59	0.57	0.57	0.62	0.53
CNN-GloVe	0.67	0.93	0.52	0.61	0.56	0.68	0.57	0.62	0.53
RNN-Attention	0.66	0.94	0.52	0.50	0.67	0.40	0.58	0.76	0.47
Best eRisk participants									
1st	0.71	0.64	0.79	-	-	-	**0.75**	0.82	0.69
2nd	0.68	0.77	0.60	-	-	-	0.62	0.62	0.62
3rd	0.68	0.67	0.68	-	-	-	0.62	0.59	0.65
Our methods: Single-channel									
Thematic	0.77	0.70	0.85	0.62	0.55	0.72	0.60	0.44	0.94
Emotion	0.70	0.85	0.60	0.61	0.62	0.61	0.63	0.68	0.59
Style	0.69	0.86	0.57	0.62	0.64	0.61	0.64	0.70	0.59
Our methods: Multi-channel Bi-transformers									
MMBT	0.76	0.72	0.84	0.65	0.51	0.91	0.65	0.75	0.58
BERT-CNN	**0.82**	0.81	0.82	**0.70**	0.54	0.96	0.70	0.73	0.68
BERT-3CNN	0.80	0.81	0.79	0.68	0.53	0.95	0.70	0.69	0.71
BERT-GMU	0.81	0.80	0.81	**0.70**	0.55	0.95	0.73	0.73	0.74

Table 2: F1, precision and recall results over the positive class in three eRisk's tasks.

combined by the multimodal transformer as well as by different architectures based on multiple BERTs.

From this evaluation, we observe that most of our proposals outperform the baseline results. Firstly, the single-channel representations obtain an improvement in comparison with baselines, in particular those based on style and emotion information. Unexpectedly, for the baselines, the performance of deep learning models applied over word-based representations is closer to traditional approaches using a Bag-of-Words. We think this could be due to the small size of the data set and the intersection of thematic content. Something interesting to notice is that CNN networks obtain better performance than RNN networks. The latter could be because CNN networks search for the presence of specific local information important for the detection of these disorders.

For our representations based on the fusion of information, their performance is higher in comparison with the other models, suggesting the relevance of combining the information from different channels. Something interesting to notice is that all models using multiple BERTs outperformed the multimodal BERT model in F1, indicating that, for this particular task, and with this way of representing the channels, it is better to represent each channel independently and combine them later. In general, the model that use the GMU module showed the best average performance, which suggest that weighting the information helps to create a better representation of the posts and the users.

From these experiments, we highlight the following observations:

1. Most single-channel representations obtain better results than the baselines.

2. The architectures using multiple BERTs obtained better performance than the multimodal BERT.

3. These results confirms our intuition that learning to combine different types of information is very relevant to capture signs of mental disorders in users.

4. In general, our models obtain an harmonic result between precision and recall deriving in a better F1 score.

6.1 Comparison against the eRisk participants

To expand our analysis and add context to the results, we also include a comparison against the original participants of the eRisk tasks. These evaluation forum considers a total of 54 models for the anorexia detection task and 57 for the self-harm detection task in eRisk-19 and 20 editions (Losada, Crestani, and Parapar, 2019; Losada, Crestani, and Parapar, 2020). It is important to mention that the participants focused on obtaining early and accurate predictions on the users, while our approach focuses exclusively on determining accurate classifications.

We observe that our models achieve competitive results in both tasks; they surpassed the first place results in Anorexia, and showed a slightly lower performance than the first place in Self-harm. For the depression task, organizers changed the evaluation strategy. While our approach focuses on binary classification, the eRisk task considered the assessment of the level of depression severity for each user. For this reason, we cannot directly compare our results against the participants.

6.2 Contribution of each information channel

To understand how each information channel contributes to the final decision we will utilize the GMU units in the fourth model (BERT-GMU) and analyze the weighting of the gates, where the gates in the unit let the model regulate the flow of information into the next one. The main idea in a GMU is that the unit learns to weigh the modalities (channels for us) and fuse them according to their relevance. A GMU works similar to a neural network layer and finds an intermediate representation based on the different modalities.

For this analysis, we obtained the gates' z_i values of the GMU module corresponding to the test set posts. Figure 7 presents the results for the three tasks, where each value already takes into account the average of all posts. For anorexia, we can appreciate that the thematic information contributes the most to the final decision, followed by emotion and style information. For depression, we can observe that the thematic information is also the most important and the value for the style information is higher than the emotional value. Finally, for the self-harm task, the thematic information obtains the lowest value and the style information the highest. In general, it can be noticed how the activation for each channel are different depending on the mental disorder. Something interesting is that the thematic channel presents the highest variation, with the lowest value in self-harm and the highest value in anorexia. We think that this variation indicates that the posts of users who suffer from anorexia are probably more homogeneous than those who suffer self-harm.

For a further analysis of the GMU, Table 3 presents the posts of the depression data set with the highest z_i value for each channel. We can notice that the posts are related to personal opinions, different topics, and in general express negative emotions even when they are not directly related to mental disorders. Take for example the emotion channel where the post is related to regrets in life and feelings, or the style channel where the post talks about a mental illness, but also contains words such as "don't" and "nothing".

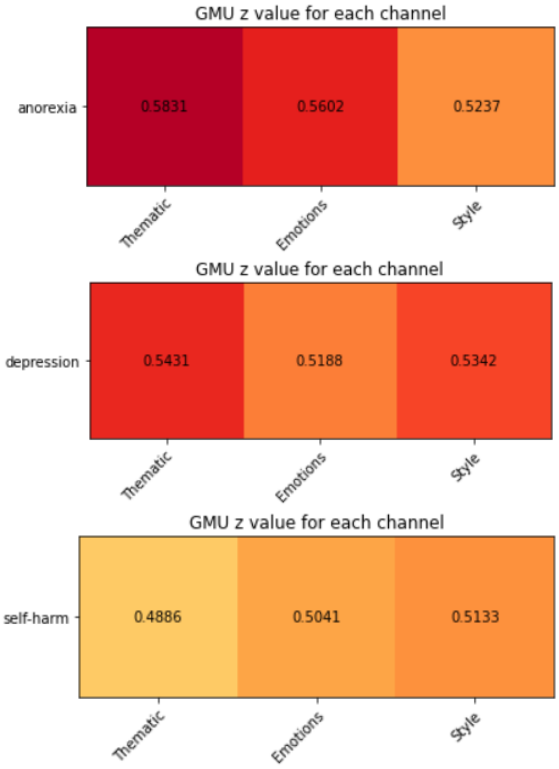


Figure 7: Average z_i value for the three mental disorders over the test set instances.

7 Conclusion and future work

In this work, we explored the detection of users that suffer anorexia, depression, or self-

Channel	Post
thematic	*"...I have no idea what either of them were trying to communicate tbh i was* ***having a really good day*** *and then you had to bring up hughes..."*
emotion	*"...take the chance and have no* ***regrets*** *in life, its always better to know if the other person* ***feels*** *something so that you are not wasting your time..."*
style	*"...these days parents* ***don't*** *know a whole lot about* ***mental illness*** *they were told it was* ***nothing*** *so that's all..."*

Table 3: Posts with highest z_i value for each channel over the depression task.

harm. For this task, we used the users' thematic interests, emotions and writing style. We tested different strategies to combine these information channels inspired by the usage of transformers to learn contextual knowledge. Our results suggest that enriching emotional and style data using a transformer improves the detection of users with mental disorders. Moreover, a striking result is the superiority of using a combination of multiple BERTs over the recent multimodal BERT transformer; this finding by its own opens an opportunity to explore models inspired by transformers to create new representations and continue improving the performance in the detection of people with mental disorders. The results outperform traditional and state-of-the-art baselines and are competitive with the performance of top eRisk participants. We believe that it is important to mention that these models, although they obtain better results, are extremely resource-consuming (processor, memory, energy, etc.) in comparison with simple models. For future work, we want to explore more sophisticated combination techniques that could improve the results and understanding of mental disorders detection, for example, multi-modal transformers. We note that most of the analysis of mental disorders has been made for the English language, then, one of our interests lies in the expansion of this study to Spanish language.

References

Aragon, M. v., A. Lopez-Monroy, L. Gonzalez-Gurrola, and M. Montes-y Gomez. 2019. Detecting depression in social media using fine-grained emotions. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).*

Arevalo, J., T. Solorio, M. Montes-y Gómez, and F. González. 2019. Gated multimodal networks. *Neural Computing and Applications.*

Baer, J. 2021. Multimodal machine learning: A survey and taxonomy. *https://www.convinceandconvert.com/social-media-research/social-media-usage-statistics/.*

Baltrusaitis, T., C. Ahuja, and L. Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2):423–443.*

Beck, A., C. Ward, M. Mendelson, J. Mock, and J. Erbaugh. 1961. An inventory for measuring depression. *JAMA Psychiatry 4(6), 561–571.*

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2016. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics.*

Coopersmith, G., M. Dredze, and C. Harman. 2014. Quantifying mental health signals in twitter. *Workshop on Computational Linguistics and Clinical Psychology.*

Coppersmith, G., M. Dredze, C. Harman, and K. Hollingshead. 2015. From adhd to sad: analyzing the language of mental health on twitter through self-reported diagnoses. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology*, pages 1–10.

De Choudhury, M., S. Counts, and E. Horvitz. 2013. Social media as a measurement tool of depression in populations. *In Proceedings of the 5th Annual ACM Web Science Conference.*

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HTL.*

Dyson, M., L. Hartling, J. Shulhan, A. Chisholm, A. Milne, P. Sundar, S. Scott, and A. Newton. 2016. A systematic review of social media use to discuss and view deliberate self-harm acts. *PLOS ONE.*

Guntuku, S., D. Yaden, M. Kern, L. Ungar, and J. Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, pages 43–49.

Hilton, C. 2016. Unveiling self-harm behaviour: what can social media site twitter tell us about self-harm? a qualitative exploration. *Journal of clinical nursing.*

Htait, A., S. Fournier, and P. Bellot. 2017. Lsis at semeval-2017 task 4: Using adapted sentiment similarity seed words for english and arabic tweet polarity classification. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).*

Ji, S., X. Li, Z. Huang, and E. Cambria. 2020. Suicidal ideation and mental disorder detection with attentive relation networks. *arXiv:2004.07601.*

Kang, K., C. Yoon, and E. Kim. 2016. Identifying depressive users in twitter using multimodal analysis. *In Big Data and Smart Computing (BigComp), 2016 International Conference on. IEEE, 231–238.*

Kessler, R., E. Bromet, P. Jonge, V. Shahly, and Marsha. 2017. The burden of depressive illness. *Public Health Perspectives on Depressive Disorders.*

Kiela, D., S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine. 2019. Supervised multimodal bitransformers for classifying images and text. *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop.*

Losada, D., F. Crestani, and J. Parapar. 2018. Overview of erisk 2018: Early risk prediction on the internet (extended lab overview). *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France.*

Losada, D., F. Crestani, and J. Parapar. 2020. Overview of eRisk 2020: Early Risk Prediction on the Internet. *Experimental IR Meets Multilinguality, Multimodality, and Interaction Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020).*

Losada, D. v., F. Crestani, and J. Parapar. 2019. Overview of erisk 2019: Early risk prediction on the internet. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland.*

Mohammad, S. and P. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence.*

Pennington, J., R. Socher, and C. Manning. 2014. Glove: global vectors for word representation. *In Proceedings of the Conference on Empirical Methods on Natural Language Processing.*

Preotiuc-Pietro, D., J. Eichstaedt, G. Park, M. Sap, L. Smith, V. Tobolsky, H. Schwartz, and L. Ungar. 2015. The role of personality, age and gender in tweeting about mental illnesses. *In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology.*

Qianli, M., S. Lifeng, C. Enhuan, T. Shuai, W. Jiabing, and C. Garrison. 2017. Walking walking walking: Action recognition from action echoes. *Twenty-Sixth International Joint Conference on Artificial Intelligence.*

Ragheb, W., J. Aze, S. Bringay, and M. Servajean. 2019. Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media. *Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland.*

Renteria-Rodriguez, M. 2018. Salud mental en mexico. *NOTA-INCyTU NÚMERO 007.*

Tausczik, Y. and J. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, pages 24–54.

Trifan, A. and J. Oliveira. 2019. Bioinfo@uavr at erisk 2019: delving into social media texts for the early detection of mental and food disorders. *Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland.*

Trotzek, M., S. Koitka, and C. Friedrich. 2018. Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia. *Proceedings of the 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France.*

Uban, A., B. Chulvi, and P. Rosso. 2021. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems.*

Van Rijen, P., D. Teodoro, N. Naderi, L. Mottin, J. Knafou, M. Jeffryes, and P. Ruch. 2019. A data-driven approach for measuring the severity of the signs of depression using reddit posts. *Proceedings of the 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland.*

Wang, T., M. Brede, A. Ianni, and E. Mentzakis. 2017. Detecting and characterizing eating-disorder communities on social media. *In Proceedings of the Tenth ACM International conference on web search and data mining.*

World Health Organization, W. 2019. Mental health: Fact sheet. *https://www.euro. who.int/en/health-topics/noncommunicable-diseases/mental-health.*

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 39-49 recibido 31-03-2022 revisado 20-05-2022 aceptado 30-05-2022

Un redactor asistido para adaptar textos administrativos a lenguaje claro

A writing assistant to adapt administrative texts into plain language

Iria da Cunha
Universidad Nacional de Educación a Distancia
iriad@flog.uned.es

Resumen: El lenguaje claro aboga por que los textos dirigidos a los ciudadanos estén redactados en un lenguaje más sencillo y transparente, para que estos puedan entender fácilmente el mensaje que se les quiere transmitir. En este contexto, nuestro objetivo es desarrollar un redactor asistido para el español que ayude al personal de la Administración pública a escribir en lenguaje claro los textos que dirige a la ciudadanía. El sistema, gratuito y en línea, integra diferentes herramientas de Procesamiento de Lenguaje Natural (PLN) para detectar en los textos escritos por los usuarios los rasgos lingüísticos que interfieren con las recomendaciones sobre lenguaje claro. Asimismo, ofrece al usuario información para hacer más sencillo su texto. Para evaluar los algoritmos se empleó un corpus anotado manualmente, y las medidas de precisión y cobertura. Los resultados son muy positivos, aunque también reflejan algunos aspectos que se pueden mejorar en el futuro.
Palabras clave: Lenguaje claro, redacción asistida, Procesamiento de Lenguaje Natural (PLN), Administración pública.

Abstract: Plain language advocates that texts addressed to citizens should be written in simpler and more transparent language, so that they can easily understand the message to be conveyed. In this context, our aim is to develop an assisted writing tool for Spanish to help Public Administration staff to write texts addressed to citizens in plain language. The system is free and online. It integrates different Natural Language Processing (NLP) tools to detect the linguistic features that interfere with plain language recommendations in the texts written by users. It also provides users with information to make their text clearer. A manually annotated corpus, and precision and recall measures were used to evaluate the algorithms. The results are very positive, although they also highlight some aspects that could be improved in the future.
Keywords: Plain Language, Assisted Writing, Natural Language Processing (NLP), Public Administration.

1 Introducción

El lenguaje claro es una corriente que aboga por que los textos que se dirijan a los ciudadanos estén redactados en un lenguaje más sencillo y transparente, para que estos puedan entender el mensaje que se les quiere transmitir y, así, contribuir a que ejerzan sus derechos y a que cumplan con sus obligaciones. Según el sitio web de la International Plain Language Federation,[1] la comunicación en lenguaje claro tiene unos rasgos concretos: "A communication is in plain language if its wording, structure, and design are so clear that the intended readers can easily find what they need, understand what they find, and use that information".

Esta corriente adquiere una relevancia especial en el marco de la modernización del lenguaje jurídico y administrativo, uno de los grandes retos de la Administración española (Cassany, 2005; Montolío, 2012; Carretero y Fuentes, 2019; Montolío y Tascón, 2020; Bayés, 2021). El reto es cambiar una tradición en la Administración que hace que los textos que recibe la ciudadanía (notificaciones, requerimientos, resoluciones, multas, etc.)

[1] https://www.iplfederation.org/plain-language/

ISSN 1135-5948. DOI 10.26342/2022-69-3

tengan unas características lingüísticas que "oscurecen" el texto, como párrafos larguísimos, oraciones con múltiples subordinadas e incisos, verbos en voz pasiva, en gerundio y en participio, formas verbales arcaicas, terminología y fraseología muy compleja, multitud de negaciones y formas subjetivas, entre otros rasgos (De Miguel, 2000; Alcaraz, Hugues y Gómez, 2014).

Por esta razón, en los últimos años se han elaborado guías y manuales para ayudar a redactar en lenguaje claro los textos que se dirigen a los ciudadanos. Algunos ejemplos son las publicaciones de la Comisión Europea (2015), Carretero et al. (2017), y Montolío y Tascón (2017). Sin embargo, a pesar de estas meritorias aportaciones, son escasos los trabajos que establecen sinergias entre el lenguaje claro en español y la tecnología.

En este contexto, el objetivo de la presente investigación es desarrollar el primer redactor asistido para el español que ayude al personal de la Administración pública a escribir en lenguaje claro los textos que dirige a la ciudadanía. Este redactor asistido se llama "arText claro", puede utilizarse gratuitamente y está disponible en línea desde la dirección: http://sistema-artext.com/. El sistema integra diferentes herramientas de Procesamiento de Lenguaje Natural (PLN), gracias a las cuales logra detectar en los textos escritos por los usuarios los rasgos lingüísticos que interfieren con las recomendaciones sobre lenguaje claro más habituales. Asimismo, ofrece al usuario información para hacer más claro y sencillo su texto. Las recomendaciones que integra el sistema están relacionadas con tres niveles de la lengua: discursivo, morfosintáctico y léxico.

En el apartado 2 se incluye un breve estado la cuestión sobre herramientas que tienen que ver con la revisión de textos, especialmente, en relación con el lenguaje claro. En el apartado 3 se muestran las fases de la metodología de la investigación. En el apartado 4 se detallan las funcionalidades y la implementación del redactor asistido. En el apartado 5 se explica la evaluación del sistema. Finalmente, en el apartado 6, se exponen las conclusiones y se plantean algunas líneas de trabajo futuro.

2 Estado de la cuestión

Actualmente, existen herramientas tecnológicas que ayudan a revisar y corregir textos en español. Estas herramientas incluyen diferentes funcionalidades relacionadas con diversos niveles de la lengua. Estos niveles pueden ir desde el más simple, como el ortográfico, hasta el más complejo, como el discursivo o pragmático, pasando por el léxico y el sintáctico. Además del clásico corrector de los procesadores de textos como Word y OpenOffice, hay otros sistemas destacables que se han desarrollado en los últimos años, como, por ejemplo, LanguageTool,[2] OutWrite,[3] Stilus[4] y Estilector.[5] Sin embargo, ninguno de estos correctores está diseñado específicamente para ayudar a redactar en lenguaje claro.

Para la lengua inglesa, sí existen algunos ejemplos muy recientes. Por ejemplo, los sistemas comerciales con demos gratuitas Hemingway Editor[6] y VisibleThread.[7] Estos sistemas otorgan una puntuación a la legibilidad o comprensibilidad del texto. Para asignar estas puntuaciones, tienen en cuenta diferentes cuestiones gramaticales y de estilo que hacen que los textos resulten más claros. Por ejemplo, recomiendan limitar el uso de los adverbios y de la voz pasiva; proponen alternativas más simples para algunas expresiones utilizadas en el texto; y señalan las oraciones que resultan difíciles de leer para que el usuario las acorte o divida. Cada una de estas cuestiones se marca en el texto con un color diferente.

En el ámbito de las tecnologías de la lengua en español, hay aún mucho camino por recorrer en relación con el lenguaje claro. Por ejemplo, es de destacar la herramienta en versión beta Clara,[8] gratuita y en línea, que permite realizar un test de claridad a textos en español, principalmente documentos administrativos y contratos de servicios. A partir de un texto escrito por el usuario, de un mínimo de 40 palabras y un máximo de 120, la herramienta ofrece un porcentaje global de claridad. Asimismo, indica un porcentaje específico de claridad para cada una de las métricas que incluye y ofrece propuestas de mejora. La herramienta incluye nueve métricas de evaluación (Torrijos y Oquendo, 2021: 123-124):

[2] https://www.languagetool.org
[3] https://es.outwrite.com/
[4] http://www.mystilus.com/
[5] http://www.estilector.com/
[6] https://hemingwayapp.com/
[7] https://www.visiblethread.com/
[8] https://clara.comunicacionclara.com/

- *Métrica 1: uso de palabras fuera del diccionario.*
- *Métrica 2: uso de conectores discursivos.*
- *Métrica 3: uso de la puntuación.*
- *Métrica 4: citas y referencias a leyes.*
- *Métrica 5: uso de la voz pasiva.*
- *Métrica 6: uso de nexos subordinados.*
- *Métrica 7: uso de tecnicismos financieros y administrativos.*
- *Métrica 8: uso de palabras del ranking de las 1000 más comunes en español.*
- *Métrica 9: número medio de palabras por frase.*

No obstante, Clara no marca en el texto escrito por el usuario las cuestiones específicas que interfieren con el lenguaje claro, ya que no es un redactor asistido ni un corrector, sino un sistema de medición de la claridad textual.

3 Metodología

La metodología de esta investigación incluye tres fases, que se detallan a continuación:

Fase 1. *Búsqueda de fuentes bibliográficas sobre lenguaje claro en el ámbito administrativo en español*. Para seleccionar las recomendaciones sobre lenguaje claro que integra el redactor asistido, partimos del trabajo de Da Cunha y Escobar (2021). En este estudio se analizan las principales fuentes sobre lenguaje claro en el ámbito jurídico-administrativo en español peninsular (Vilches y Sarmiento, 2010; Ministerio de Justicia, 2011; Comisión Europea, 2015; Jiménez Yáñez, 2016; Carretero *et al*., 2017; Montolío y Tascón, 2017; Carretero, 2019) y se cuantifican las recomendaciones que recogen, para seleccionar las más frecuentes. A continuación, se dividen las recomendaciones seleccionadas en función de los tres niveles de la lengua mencionados en el apartado 1:

- Nivel discursivo. Ejemplo de recomendación: "Se recomienda redactar oraciones cortas".
- Nivel morfosintáctico. Ejemplo de recomendación: "Se recomienda utilizar la voz activa en vez de la voz pasiva".
- Nivel léxico. Ejemplo de recomendación: "Se recomienda evitar los arcaísmos".

Fase 2. *Diseño e implementación del sistema*.

Fase 2a. Selección de herramientas de PLN en abierto para la lengua española que permitan un procesamiento lingüístico del texto escrito por el usuario, concretamente:

- Lematización.
- Análisis *Part of Speech* (POS).
- Segmentación oracional.
- Segmentación discursiva intraoracional.

Fase 2b. Diseño de algoritmos que detecten en el texto escrito por el usuario los rasgos lingüísticos que interfieren con el lenguaje claro. Para ello se tienen en cuenta las recomendaciones seleccionadas en la Fase 1 y el resultado del procesamiento lingüístico del texto de la Fase 2a. Por ejemplo, en el caso de la recomendación "Se recomienda redactar oraciones cortas", el algoritmo detecta todas las oraciones del texto y contabiliza las palabras que incluye cada una. Si una oración supera las 25 palabras, la subraya en amarillo en el texto escrito por el usuario y le ofrece la recomendación correspondiente para adaptarla a lenguaje claro. En el caso de la recomendación "Se recomienda utilizar la voz activa en vez de la voz pasiva", el algoritmo marca en el texto del usuario los verbos en voz pasiva obtenidos gracias al analizador POS y le ofrece la recomendación correspondiente.

Fase 2c. Redacción de las recomendaciones sobre lenguaje claro ofrecidas al usuario. En cada recomendación se incluye la siguiente información:

- Título de la recomendación.
- Sugerencia para la adaptación a lenguaje claro.
- Ejemplo (en caso de que se considere necesario para que el usuario entienda la recomendación).

En el Anexo 1 se detallan todas las recomendaciones que ofrece el sistema.

Fase 2d. Implementación del redactor asistido (los detalles técnicos se incluyen en el apartado 4).

Fase 3. *Evaluación del sistema*. Para evaluar el sistema, se compila un corpus de textos del ámbito de la Administración, concretamente, resoluciones del BOAM (Boletín Oficial del Ayuntamiento de Madrid). Una persona con formación en lingüística y experiencia en

anotación de corpus anota manualmente en cada texto los diferentes rasgos lingüísticos correspondientes a las recomendaciones que se quieren evaluar (detallados en el apartado 5). A continuación, se aplica el sistema sobre los textos del corpus para obtener automáticamente las recomendaciones sobre lenguaje claro. Finalmente, se calcula la precisión y cobertura de los resultados del sistema en contraposición con la anotación manual.

4 Funcionalidades e implementación del redactor asistido

El sistema se desarrolló en un entorno Linux con un servidor Apache. También se usaron distintos recursos en el *back-end* (Bash, Perl y PHP, con un entorno de trabajo Laravel) y en el *front-end* (HTML, CSS, JavaScript, con AJAX y jQuery). El sistema está optimizado para utilizarse con el navegador Google Chrome.

El redactor integra dos herramientas de PLN existentes para el español que permiten procesar lingüísticamente el texto escrito por el usuario:

- El analizador morfosintáctico de Freeling (Atserias et al., 2006), mediante el cual se lematizan todas las unidades léxicas del texto y se asigna una categoría gramatical a cada una de ellas. Este analizador permite detectar rasgos lingüísticos que son utilizados por los algoritmos del sistema en las recomendaciones que tienen que ver principalmente con el nivel morfosintáctico, como, por ejemplo, los verbos en voz pasiva, los gerundios, los participios, los verbos en futuro de subjuntivo, o los verbos en 1.ª persona del singular y del plural.
- Un segmentador discursivo (Da Cunha et al., 2010, 2012b), que permite dividir el texto en oraciones y, además, en segmentos discursivos intraoracionales, a partir de la definición de Tofiloski et al. (2009, p.77): "Discourse segmentation is the process of decomposing discourse into elementary discourse units (EDUs), which may be simple sentences or clauses in a complex sentence, and from which discourse trees are constructed". Este segmentador permite detectar los rasgos utilizados por los algoritmos en las recomendaciones relativas al nivel discursivo, como la segmentación discursiva de las oraciones largas y la sugerencia de conectores alternativos.

Además de integrar estas dos herramientas, el sistema incluye diversos algoritmos desarrollados en el marco de nuestra investigación. Estos algoritmos toman como entrada el texto procesado lingüísticamente por las dos herramientas de PLN mencionadas y detectan en el texto del usuario los rasgos lingüísticos necesarios para poder ofrecer las recomendaciones asociadas a cada uno de ellos. Estos rasgos son:

- Párrafos-oración.
- Párrafos largos, con un umbral de 135 palabras (teniendo en cuenta las fuentes recopiladas en la Fase 1).
- Oraciones largas, con un umbral de 25 palabras (teniendo en cuenta las fuentes recopiladas en la Fase 1).
- Conectores discursivos interoracionales e interoracionales que evidencian ocho relaciones discursivas: antítesis, causa, concesión, condición, contraste, propósito, reformulación y resumen (Da Cunha et al., 2012a).
- Nominalizaciones verbales. Concretamente, el algoritmo detecta los sustantivos acabados en *-ción* (en singular y plural) que comienzan por minúscula, excepto los incluidos en una lista de exclusión predefinida que integra términos del ámbito de la Administración, como "licitación" y "notificación" (Da Cunha, 2022).
- Unidades que expresan negación de una lista predefinida, que incluye unidades como "no", "ni" y "ninguno".
- Unidades léxicas que indican subjetividad, como ciertos adjetivos (ej. "bueno"), adverbios (ej. "evidentemente") y frases (ej. "sin ninguna duda"), extraídas de Otaola (1988).
- Siglas propias (Giraldo, 2008) y sus correspondientes términos desplegados. Para hacer la correlación entre la sigla y su término desplegado, se tiene en cuenta que la letra inicial de las unidades léxicas incluidas en el término (excepto las *stopwords*) se correspondan, en el mismo orden, con las mismas letras que incluye la sigla. Ej. "Plan de Emergencias Invernales

del Ayuntamiento de Madrid" > "PEIAM".

- Términos difíciles de entender que tienen una variante sinonímica más sencilla de una lista predefinida (Da Cunha, 2022). Por ejemplo, la variante más sencilla del verbo "adverar" es "certificar" y la variante más sencilla del sustantivo "aquiescencia" es "consentimiento".
- Expresiones difíciles de entender que tienen una variante sinonímica más sencilla de una lista predefinida (Da Cunha, 2022). Por ejemplo, el latinismo "ad valorem" tiene como variante en español "según en valor" y la expresión arcaica "a tenor de" tiene como variante más clara "según".
- Palabras poco precisas de una lista predefinida (Da Cunha, 2022), como "cosa", "varios", "alguno", "muy" y "poco".
- Expresiones redundantes de una lista predefinida (Da Cunha, 2022), como "está claro que", "mi opinión personal" y "como es bien sabido".
- Palabras largas que tienen una variante sinonímica más breve de una lista predefinida (Da Cunha, 2022), como "gratuitamente/gratis" y "encomendar/encargar".

En total, el redactor incluye 22 recomendaciones sobre lenguaje claro. Como se ha indicado, en el Anexo 1 se detallan todas ellas, divididas en función de los tres niveles de la lengua mencionados en el apartado 3.

En la Figura 1 se ofrece una captura de pantalla de "arText claro" en donde se muestra la recomendación sobre párrafos largos en un texto del corpus de evaluación.

En relación con la exportación e importación de documentos, por una cuestión de protección de datos, se decidió que el sistema no guardase en su servidor los textos escritos por los usuarios. Para ello, existen varias opciones de exportación de documentos en local: .doc, .pdf, .txt, .html, etc. Para poder importar un texto posteriormente en el redactor, debe utilizare un formato creado específicamente para este sistema: .artext.

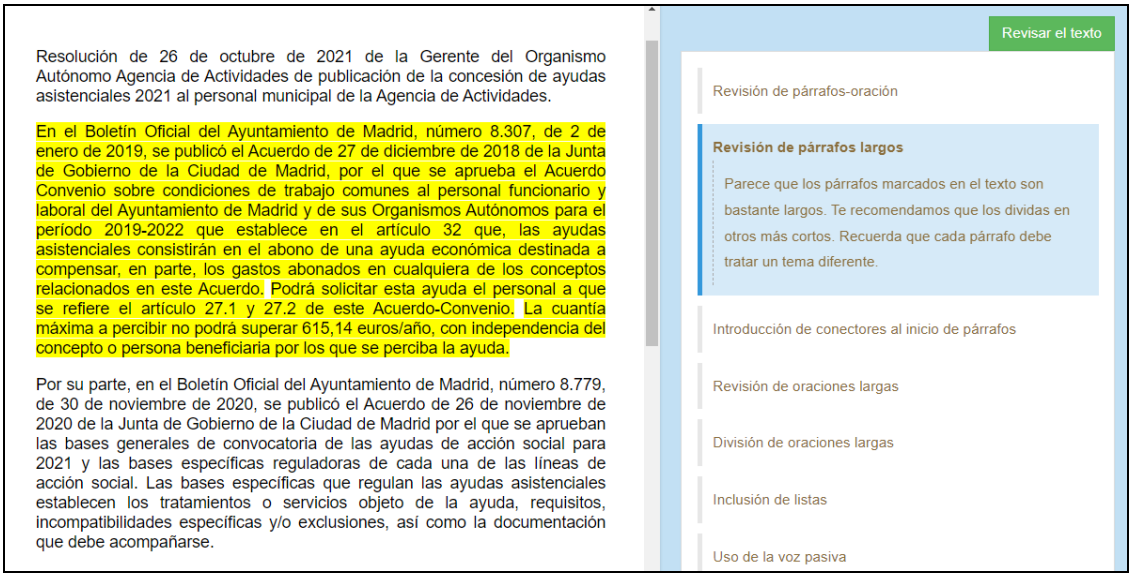


Figura 1: Captura de pantalla de "arText claro" en donde se muestra la recomendación sobre párrafos largos en un texto del corpus de evaluación.

5 *Evaluación*

Como se avanzaba en el apartado 3, una vez implementado el sistema, se llevó a cabo una evaluación *data-driven* utilizando un corpus formado por 10 resoluciones del BOAM publicadas en el año 2021, que en total suman 8.052 palabras. El texto más corto tiene 436 palabras y el texto más largo, 1398. En el Anexo 2 se recogen los títulos de las resoluciones empleadas.

Para comparar los resultados del sistema con los resultados de la anotación manual, se calculó la precisión y cobertura de los rasgos

lingüísticos anotados en los textos del corpus. Los rasgos anotados se incluyen en la Tabla 1.

Nivel de	Rasgos anotados
Discursivo	a1. Párrafos-oración
	a2. Párrafos largos
	a3. Párrafos que no comienzan por un conector discursivo
	a4. Oraciones largas
	a5. Oraciones largas que pueden dividirse en segmentos discursivos
	a6. Conectores discursivos que aparecen 3 veces o más
	a7. Listas
Morfo-sintáctico	b1. Verbos en voz pasiva
	b2. Gerundios
	b3. Participios
	b4. Verbos en futuro de subjuntivo
	b5. Verbos en 1.ª persona del plural y del singular
	b6. Nominalizaciones verbales
	b7. Unidades que expresan negación
Léxico	c1. Unidades que expresan subjetividad
	c2. Siglas que no aparecen con su correspondiente término desplegado la 1.ª vez que aparecen en el texto
	c3. Términos desplegados para los que se introdujo su sigla previamente en el texto
	c4. Términos difíciles de entender que tienen una variante sinonímica más sencilla
	c5. Expresiones difíciles de entender que tienen una variante más sencilla
	c6. Palabras poco precisas
	c7. Expresiones redundantes
	c8. Palabras largas que tienen una variante sinonímica más breve

Tabla 1: Rasgos lingüísticos anotados manualmente en el corpus de evaluación.

Los resultados de la evaluación de cada recomendación se muestran en la Tabla 2 ("ID" se refiere al identificador de la recomendación).

ID	Precisión	Cobertura
a1	0,99	0,74
a2	0,7	0,8
a3	0,84	0,71
a4	1	0,7
a5	1	0,89
a6	1	1
a7	1	0,83
b1	1	0,67
b2	1	1
b3	0,86	1
b4	1	1
b5	0,17	1
b6	1	1
b7	1	1
c1	Rasgo sin ocurrencias en el corpus	
c2	0,6	1
c3	1	1
c4	1	1
c5	1	1
c6	1	1
c7	Rasgo sin ocurrencias en el corpus	
c8	1	1

Tabla 2: Resultados de la evaluación del sistema.

Como puede apreciarse, los resultados obtenidos son, en general, positivos para la mayor parte de las recomendaciones evaluadas. Destaca especialmente que el sistema obtiene la máxima precisión y cobertura en relación con la detección de 10 rasgos: los conectores que aparecen más de tres veces en el texto, los gerundios, los verbos en futuro de subjuntivo, las nominalizaciones verbales, las unidades que expresan negación, los términos desplegados para los que se introdujo su sigla previamente en el texto, los términos difíciles de entender que tienen una variante sinonímica más sencilla, las expresiones difíciles de entender que tienen una variante más sencilla y las palabras poco precisas.

En cuanto a las recomendaciones relacionadas con los párrafos (a1, a2, a3), también se obtienen resultados positivos, aunque bajan ligeramente. Esto se debe, principalmente, a signos de puntuación que interfieren en la segmentación oracional, lo cual tiene consecuencias en la detección correcta de párrafos:

- Citas a artículos y leyes. Ej. "los artículos 8.1 y 46.1 de la Ley".
- Elementos numerados con puntuación Ej. "1.", "2.".
- Enlaces web. Ej. "en la intranet/extranet municipal ayre (https://ayre.madrid.es) y en la web https://jubilacion.madrid.es".

En cuanto a las recomendaciones relacionadas con la detección de oraciones largas y segmentación discusiva (a4, a5), la precisión es muy alta, pero en cambio baja ligeramente la cobertura. El motivo es que en este tipo de textos administrativos en ocasiones hay oraciones que acaban con una coma o directamente sin ningún signo de puntuación.

Por tanto, el sistema no logra recuperarlas. Por ejemplo:

(1) "En su virtud, de conformidad con el Acuerdo de 27 de junio de 2019 […] la competencia para la ejecución de los planes y programas de formación de los empleados y directivos del Ayuntamiento de Madrid,"
(2) "En virtud de las facultades que me han sido conferidas por [..], esta Gerencia"

En cuanto a la detección de rasgos morfosintácticos, los que no alcanzan la máxima precisión y cobertura son los siguientes:

- Verbos en voz pasiva (b1). En este caso, los errores en la cobertura (0,67) provienen del procesamiento lingüístico con Freeling. Por ejemplo, no se detectan las formas "hubiera sido cesado" o "han sido conferidas".
- Participios (b3). En esta ocasión el problema es de precisión (0,86) y se debe a la desambiguación de Freeling. Por ejemplo, se detecta "propuesta" y "puesto" como participios cuando en el texto tienen función de sustantivo ("la propuesta", "el puesto de trabajo").
- Verbos en 1.ª persona del plural y del singular (b5). Este ha sido el rasgo que ha obtenido peores resultados en la evaluación, con un 0,17 de precisión, debido también a problemas en la desambiguación. Por ejemplo, se detectan como 1.ª persona del singular las formas "sea" y "haga", cuando en realidad en el texto son formas de 3ª persona singular ("sea esta accidental o intencionada", "cuando la previsión meteorológica haga previsible").

En relación con el nivel léxico, como se ha visto, se obtienen muy buenos resultados. Únicamente baja la precisión (0,6) en la detección de siglas que no aparecen con su correspondiente término desplegado la primera vez que aparecen en el texto (c2). El motivo principal es que en estos documentos suele haber números romanos (ej. "III", "IV") y el sistema los marca erróneamente como siglas. También señala otras secuencias de letras en mayúscula (como los acrónimos) que sí aparecen con su término desplegado y que, por tanto, no sería pertinente marcar en el texto del usuario. Ej. "Plan Territorial Superior de la Comunidad de Madrid (PLATERCAM)".

Finalmente, como se observa en la Tabla 2, hay dos recomendaciones del nivel léxico que no se han podido evaluar (c1 y c7), porque no se han detectado en el corpus unidades que expresan subjetividad ni expresiones redundantes.

6 Conclusiones y líneas de trabajo futuro

Como se ha visto en este trabajo, el lenguaje claro es un tema de investigación en el que aún queda mucho por explorar, especialmente desde el punto de vista del PLN. El objetivo de este trabajo ha sido desarrollar el primer redactor asistido para el español para escribir textos administrativos en lenguaje claro. La herramienta desarrollada tiene un gran potencial de aplicación en las diferentes dependencias de la Administración pública, como ayuntamientos, diputaciones, consejerías, ministerios, etc. Los empleados públicos dispondrán, así, de una herramienta tecnológica que les ayudará a revisar sus textos y adaptarlos al lenguaje claro.

La evaluación de "arText claro" ofrece buenos resultados, aunque aún quedan cuestiones que se pueden mejorar, como la precisión en la detección de siglas, o las interferencias con los signos de puntuación que provocan errores en la detección de párrafos y oraciones largas. También sería interesante ampliar el corpus de evaluación con más textos anotados manualmente, de otros géneros textuales del ámbito administrativo, con el objetivo de validar los resultados obtenidos en esta investigación.

Con respecto a la metodología, una posible línea de trabajo futuro sería la aplicación de estrategias de aprendizaje automático. Este enfoque ya está siendo utilizado en algunas investigaciones, como es el caso del sistema de medición de la claridad textual Clara (Torrijos y Oquendo, 2021), mencionado en el apartado 2. La principal dificultad de este enfoque, sin embargo, es la necesidad de contar con corpus muy extensos de textos originales y sus correspondientes textos clarificados.

Otra de nuestras líneas de investigación será llevar a cabo estudios de evaluación de la percepción de la claridad y de la comprensión de los textos escritos con la herramienta por parte de los destinatarios. Finalmente, nuestro

objetivo es implementar el sistema en las dependencias de la Administración pública española.

Agradecimientos

Este trabajo se deriva del proyecto de investigación titulado "Tecnologías de la Información y la Comunicación para la e-Administración: hacia la mejora de la comunicación entre Administración y ciudadanía a través del lenguaje claro (TIC-eADMIN)", financiado por el Ministerio de Ciencia, Innovación e Universidades en la convocatoria 2018 de Proyectos I+D del Subprograma Estatal de Generación de Conocimiento (referencia PGC2018-099694-A-I00), y desarrollado en el Departamento de Filologías Extranjeras y sus Lingüísticas de la Facultad de Filología de la Universidad Nacional de Educación a Distancia (UNED), en el marco del grupo de investigación ACTUALing y en colaboración con el grupo IULATERM (IULA-UPF).

Bibliografía

Alcaraz, E., B. Hugues, y A. Gómez. 2014. *El español jurídico*. 3.ª edición. Ariel, Barcelona.

Atserias, J., B. Casas, E. Comelles, M. González, Ll. Padró, y M. Padró. 2006. FreeLing 1.3. Syntactic and semantic services in an open-source NLP library. En *LREC 2006 Proceedings. 5th Edition of the International Conference on Language Resources and Evaluation*, páginas 48-55, European Language Resources Association (París).

Da Cunha, I. (Ed.). 2022. *Lenguaje claro y tecnología en la Administración*. Granada, Comares. En prensa.

Da Cunha, I. y M. Á. Escobar. 2021. Recomendaciones sobre lenguaje claro en español en el ámbito jurídico-administrativo: análisis y clasificación. *Pragmalingüística*, 29:129-148.

Da Cunha, I., E. SanJuan, J-M. Torres-Moreno, M. T. Cabré, y G. Sierra. 2012a. A Symbolic Approach for Automatic Detection of Nuclearity and Rhetorical Relations among Intra-sentence Discourse Segments in Spanish. *Lecture Notes in Computer Science*, 7181:462-474.

Da Cunha, I., E. SanJuan, J-M. Torres-Moreno, M. Lloberes, e I. Castellón. 2012b. DiSeg 1.0: The First System for Spanish Discourse Segmentation. *Expert Systems with Applications*, 39(2):1671-1678.

Da Cunha, I., E. SanJuan, J-M. Torres-Moreno, M. Lloberes, e I. Castellón. 2010. DiSeg: Un segmentador discursivo automático para el español. *Procesamiento del Lenguaje Natural*, 45:145-152.

Bayés, M. 2021. Análisis del impacto de una selección de (meta)indicaciones de redacción clara en la percepción de claridad de un documento administrativo: estudio de caso. Tesis doctoral, Universitat de Barcelona.

Carretero, C. 2019. *Comunicación para juristas*. Tirant lo Blanch, Valencia.

Carretero, C., y J. C. Fuentes. 2019. La claridad del lenguaje jurídico. *Revista del Ministerio Fiscal*, 8:7-40.

Carretero, C., J. M. Pérez, L. Lanne-Lenne, y G. de los Reyes. 2017. *Lenguaje Claro. Comprender y hacernos entender*. Instituto de Lectura Fácil y Clarity, Sevilla, Madrid.

Cassany, D. 2005. Plain Language in Spain. *Clarity*, 53:41-44.

Comisión Europea. 2015. *Cómo escribir con claridad*. Oficina de Publicaciones de la Unión Europea, Luxemburgo.

De Miguel, E. 2000. El texto jurídico-administrativo: análisis de una orden ministerial. *Círculo de lingüística aplicada al a comunicación*, 4: en línea. Disponible en: https://webs.ucm.es/info/circulo/no4/demiguel.htm

Giraldo, J. J. 2008. Análisis y descripción de las siglas en el discurso especializado de genoma humano y medio ambiente. Tesis doctoral. Universitat Pompeu Fabra, Institut de Lingüística Aplicada (IULA), Barcelona.

Jiménez Yáñez, R. M. 2016. *Escribir bien es de justicia*. Aranzadi, Cizur Menor.

Ministerio de Justicia. 2011. *Informe de la Comisión de modernización del lenguaje jurídico*. Ministerio de Justicia, Madrid. En línea.

Montolío, E. 2012. La modernización del discurso jurídico español impulsada por el Ministerio de Justicia. Presentación y principales aportaciones del Informe sobre el lenguaje escrito. *Revista de Llengua i Dret*, 57:95-121.

Montolío, E. y M. Tascón. 2020. *El derecho a entender: la comunicación clara, la mejor defensa de la ciudadanía*. La Catarata, Madrid.

Montolío, E. y M. Tascón. 2017. *Comunicación Clara. Guía Práctica*. Ayuntamiento de Madrid, Madrid.

Otaola, C. 1988. La modalidad (con especial referencia a la lengua española). *Revista de Filología Española*, 68(1):97-117.

Tofiloski, M., J. Brooke, y M. Taboada. 2009. A syntactic and lexical-based discourse segmenter. En *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, páginas 77-80, Association for Computational Linguistics (Singapur).

Torrijos, C. y S. Oquendo. 2021. ¡Hola! Soy clara y mido la claridad de tu texto" *Architetras Científica*, Vol. VI:119-133.

Vilches, F. y R. Sarmiento. 2010. *Manual de lenguaje Jurídico-Administrativo*. Dykinson, Madrid.

A Anexo 1: Recomendaciones que ofrece "arText claro"

a) Recomendaciones del nivel discursivo

a1. Título de la recomendación: *Revisión de párrafos-oración*.

Texto de la recomendación: *Parece que los párrafos marcados en el texto solo incluyen una oración. Ten en cuenta que suele recomendarse que cada párrafo incluya al menos dos oraciones.*

a2. Título de la recomendación: *Revisión de párrafos largos*.

Texto de la recomendación: *Parece que los párrafos marcados en el texto son bastante largos. Te recomendamos que los dividas en otros más cortos. Recuerda que cada párrafo debe tratar un tema diferente.*

a3. Título de la recomendación: *Introducción de conectores al inicio de párrafos*.

Texto de la recomendación: *Parece que los párrafos marcados en el texto no comienzan con una marca explícita que los relacione con su párrafo anterior. Te recomendamos que enlaces los diferentes párrafos a través de conectores discursivos. Haz clic en cada acción para ver sugerencias de conectores que pueden servirte de ayuda para introducir párrafos:*

Introducir un tema nuevo
Marcar un orden
Distinguir
Seguir el mismo tema
Enfatizar y reformular
Detallar
Resumir
Terminar
Indicar causa
Indicar consecuencia
Indicar oposición
Indicar objeción
Mostrar elementos en forma de lista

a4. Título de la recomendación: *Revisión de oraciones largas*.

Texto de la recomendación: *Las oraciones marcadas son muy largas. Te recomendamos que las revises. Por ejemplo, podrías dividir la oración en otras más cortas o eliminar la información poco relevante.*

a5. Título de la recomendación: *División de oraciones largas*.

Texto de la recomendación: *Parece que las oraciones marcadas podrían dividirse en otras más cortas. Te recomendamos que lo hagas. Haz clic en cada oración para ver dónde podrías segmentarla. Si decides dividir la oración en otras más cortas, te recomendamos que utilices conectores. Haz clic en las unidades marcadas en rojo en el texto para ver sugerencias de conectores alternativos.*

a6. Título de la recomendación: *Variación de conectores.*

Texto de la recomendación: *Los conectores de la lista siguiente se repiten varias veces en el texto. Haz clic en cada conector para ver sugerencias de conectores alternativos.*

a7. Título de la recomendación: *Inclusión de listas*.

Texto de la recomendación: *Parece que no has incluido ninguna lista en tu texto utilizando las opciones disponibles en la barra superior de herramientas (numeración o viñetas). Recuerda que las listas, si están bien construidas, son muy eficaces para transmitir información de manera clara. Te recomendamos que añadas alguna lista a tu texto.*

Ejemplo:

Para presentar su solicitud, debe enviar cuatro documentos:

1. *El formulario de solicitud cumplimentado.*
2. *Una fotocopia de su DNI.*

3. *Su acta de nacimiento.*
4. *Su certificado de empadronamiento.*

b) Recomendaciones del nivel morfosintáctico

b1. Título de la recomendación: *Uso de la voz pasiva.*

Texto de la recomendación: *Las unidades marcadas parecen verbos en voz pasiva. Ten en cuenta que en este tipo de textos es más habitual la voz activa.*

Ejemplo:

La oración "El suministro que se contrata fue aprobado por este organismo" podría sustituirse por "Este organismo aprobó el suministro que se contrata".

b2. Título de la recomendación: *Revisión de gerundios.*

Texto de la recomendación: *Las unidades marcadas en el texto parecen verbos en gerundio. Te recomendamos que evites estas formas verbales, ya que pueden causar ambigüedades, alargar demasiado la oración y hacer difícil la comprensión.*

Ejemplo:

En la siguiente oración, no queda claro quién es el sujeto del gerundio: "El aspirante a la plaza dio su documentación al empleado del registro solicitando una copia."

Si el sujeto es el aspirante, sería más conveniente decir: "El aspirante a la plaza dio su documentación al empleado del registro, a quien solicitó una copia."

En cambio, si el sujeto es el empleado del registro, sería más adecuado decir, por ejemplo: "El aspirante a la plaza dio su documentación al empleado del registro y solicitó una copia."

b3. Título de la recomendación: *Revisión de participios.*

Texto de la recomendación: *Las unidades marcadas en el texto parecen verbos en participio. A no ser que sea imprescindible utilizarlos, te recomendamos que evites estas formas verbales, ya que pueden causar ambigüedades, alargar demasiado la oración y hacer difícil la comprensión.*

Ejemplos:

En las siguientes oraciones, no queda claro quién es el sujeto del participio:
"Enviada la resolución, el aspirante tiene diez días para reclamar."
"Presentado el documento, se cerró el expediente."

Podría especificarse de la siguiente manera, por ejemplo:
"Una vez el comité evaluador envíe la resolución, el aspirante tiene diez días para reclamar."
"Cuando el solicitante presentó el documento, se cerró el expediente".

b4. Título de la recomendación: *Eliminación de formas verbales arcaicas.*

Texto de la recomendación: *Las formas verbales marcadas en el texto están en desuso y son innecesarias. Te recomendamos que las evites y que utilices en su lugar otras formas más sencillas y actuales.*

Ejemplos:

En vez de "si resultare necesario", podría decirse "si resulta necesario", "si resultara necesario" o "si resultase necesario".

En vez de "si hubiere sido necesario", podría decirse "si hubiera sido necesario" o "si hubiese sido necesario".

b5. Título de la recomendación: *Sistematicidad en el uso de verbos en 1.ª persona.*

Texto de la recomendación: *Las unidades marcadas en verde parecen verbos en 1.ª persona del singular y las marcadas en azul parecen verbos en 1.ª persona del plural. Si estas formas verbales se refieren al emisor del texto, te recomendamos que optes por el singular o el plural para que el texto sea sistemático.*

b6. Título de la recomendación: *Revisión de nominalizaciones verbales.*

Texto de la recomendación: *Parece que algunas de las palabras marcadas en el texto son nombres derivados de verbos y, por tanto, pueden resultar difíciles de entender. A no ser que sean términos de este ámbito que no se puedan cambiar, te recomendamos que las sustituyas por sus correspondientes verbos para hacer el texto más claro y dinámico.*

Ejemplos:

La oración "Se llevará a cabo una evaluación de los riesgos" podría sustituirse por "Se evaluarán los riesgos".

La oración "Se efectuó la instalación de los programas informáticos" podría sustituirse por "Se instalaron los programas informáticos".

b7. Título de la recomendación: *Reformulación de ideas expresadas en negativo.*

Texto de la recomendación: *Parece que las palabras marcadas en el texto indican negación. Te recomendamos que optes por la formulación afirmativa de tus ideas en caso de ser posible, puesto que así se favorece la legibilidad y la interpretación del mensaje.*

Ejemplos:

La oración "No es infrecuente que se acepten nuevos proyectos" podría formularse en positivo de la siguiente manera: "Es habitual que se acepten nuevos proyectos".

c) Recomendaciones del nivel léxico

c1. Título de la recomendación: *Uso de indicadores de subjetividad.*

Texto de la recomendación: *Las unidades marcadas podrían indicar subjetividad. Ten en cuenta que este tipo de textos suelen ser objetivos. Te recomendamos que revises estas unidades para confirmar que son adecuadas en tu texto.*

c2. Título de la recomendación: *Introducción de siglas.*

Texto de la recomendación: *Las unidades marcadas parecen siglas. Si es así, ten en cuenta que la primera vez que se utiliza una sigla en un texto suele ir acompañada del término desplegado.*

Ejemplos:

Universidad Nacional de Educación a Distancia (UNED)

EPOC (enfermedad pulmonar obstructiva crónica)

c3. Título de la recomendación: *Sistematicidad en el uso de siglas.*

Texto de la recomendación: *Las unidades marcadas parecen el término desplegado de siglas que utilizas en el texto. Si es así, ten en cuenta que, una vez se introduce una sigla en un texto, se suele seguir utilizando la sigla y no el término desplegado.*

c4. Título de la recomendación: *Utilización de términos más transparentes.*

Texto de la recomendación: *Los términos de la lista siguiente aparecen en tu texto y pueden resultar difíciles de entender. Te recomendamos que los sustituyas por otros más transparentes o que los aclares entre paréntesis la primera vez que aparecen. Haz clic en cada uno de ellos para ver sugerencias de términos alternativos más claros.*

c5. Título de la recomendación: *Sustitución de expresiones difíciles de entender.*

Texto de la recomendación: *Las expresiones de la lista siguiente aparecen en tu texto y pueden resultar difíciles de entender, por ser formas arcaicas, en desuso o latinismos. Te recomendamos que las sustituyas por otras más claras o, si no es posible, que las expliques en el texto la primera vez que aparecen (entre paréntesis o en una nota, por ejemplo). Haz clic en cada una de ellas para ver sugerencias de variantes alternativas más comprensibles o una explicación de su significado.*

c6. Título de la recomendación: *Sustitución de palabras poco precisas.*

Texto de la recomendación: *Parece que las palabras marcadas en el texto son poco precisas. Te recomendamos que las elimines o las sustituyas por otras palabras más precisas.*

Ejemplos:

En vez de "Se trataron algunos temas en la reunión", sería más preciso decir "Se trataron cuatro temas en la reunión".

En vez de "Dijeron que había que hacer una reunión", sería más preciso decir "Dijeron que había que convocar una reunión".

c7. Título de la recomendación: *Eliminación de expresiones redundantes.*

Texto de la recomendación: *Las expresiones marcadas en el texto incluyen información redundante. Te recomendamos que las elimines para hacer el texto más breve.*

c8. Título de la recomendación: *Revisión de palabras largas.*

Texto de la recomendación: *Las palabras de la lista siguiente aparecen en tu texto y tienen variantes alternativas más cortas. Te recomendamos que las utilices para favorecer la legibilidad. Haz clic en cada palabra para ver cuál es su alternativa más corta.*

B Anexo 2: Resoluciones del BOAM incluidas en el corpus de evaluación

1. BOAM 8933 (19/07/2021). Resolución 1927.
2. BOAM 8980 (22/09/2021). Resolución 2393.
3. BOAM 9001 (22/10/2021). Resolución 2773.
4. BOAM 9030 (07/12/2021). Resolución 3277.
5. BOAM 8919 (29/06/2021). Resolución 1691.
6. BOAM 9027 (01/12/2021). Resolución 3241.
7. BOAM 9023 (25/11/2021). Resolución 3167.
8. BOAM 9005 (28/10/2021). Resolución 2831.
9. BOAM 8997 (18/10/2021). Resolución 2728.
10. BOAM 9021 (23/11/2021). Resolución 3120.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 51-62 recibido 08-03-2022 revisado 20-05-2022 aceptado 23-05-2022

Exploiting user-frequency information for mining regionalisms in Argentinian Spanish from Twitter

Explotando información de frecuencia de usuarios para minar regionalismos del español de Argentina en Twitter

Juan Manuel Pérez,[1,2] **Damián E. Aleman,**[1]
Santiago N. Kalinowski,[3] **Agustín Gravano**[4,2]

[1]Universidad de Buenos Aires, Argentina
[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina
[3]Academia Argentina de Letras, Buenos Aires, Argentina
[4]Universidad Torcuato Di Tella, Buenos Aires, Argentina

{jmperez,daleman}@dc.uba.ar, s.kalinowski@aal.edu.ar, agravano@utdt.edu

Abstract: The task of detecting regionalisms (expressions or words used in certain regions) has traditionally relied on the use of questionnaires and surveys, heavily depending on the expertise and intuition of the surveyor. The emergence of social media and microblogging services has produced an unprecedented wealth of content (mainly informal text generated by users), opening new opportunities for linguists to extend their studies of language variation. Previous work on the automatic detection of regionalisms depended mostly on word frequencies. In this work, we present a novel metric based on Information Theory that incorporates user frequency. We tested this metric on a corpus of Argentinian Spanish tweets in two ways: via manual annotation of the relevance of the retrieved terms, and also as a feature selection method for geolocation of users. In either case, our metric outperformed other techniques based on word frequency, suggesting that measuring the amount of users that use a word is an informative feature. This tool has helped lexicographers discover several unregistered words of Argentinian Spanish, as well as different meanings assigned to registered words.
Keywords: Lexical dialectology, Social media, Spanish variants, Entropy.

Resumen: La tarea de detectar regionalismos (expresiones o palabras utilizadas en determinadas regiones) se ha basado tradicionalmente en el uso de cuestionarios y encuestas, dependiendo en gran medida de la pericia e intuición del investigador. El surgimiento de las redes sociales y los servicios de microblogging ha producido una riqueza de contenido sin precedentes (principalmente textos informales generados por usuarios), lo cual ha abierto nuevas oportunidades para el estudio de la variación lingüística. Estudios previos de la detección automática de regionalismos dependen sobre todo de la frecuencia de palabras. En este trabajo presentamos una métrica novedosa basada en la Teoría de la Información, que incorpora la frecuencia de usuarios. Ponemos a prueba esta métrica en un corpus de Tweets en español argentino de dos maneras: a través de la anotación manual de la relevancia de los términos recuperados, y también usándola como un método de selección de características para la geolocalización automática de usuarios. En ambos casos, nuestra métrica superó otras técnicas basadas en la frecuencia de palabras, lo que sugiere que medir la cantidad de usuarios que usan una palabra es una característica informativa. Esta herramienta ha ayudado a lexicógrafos a descubrir varias palabras no registradas del español argentino, así como significados nuevos de palabras ya registradas.
Palabras clave: Dialectología léxica, Redes sociales, Variantes del español, Entropía.

ISSN 1135-5948. DOI 10.26342/2022-69-4

1 Introduction

Lexicography has been aided and enriched in the past 30 years by tools and resources from Computational Linguistics, mainly in the form of corpora of selected texts (Atkins and Rundell, 2008). Statistical analyses of corpora usually result in evidence to support the addition of a word to a dictionary, its removal, or its marking as dated or as unused or as regional, among other decisions.

In the process of compiling dictionaries, differences emerge between dialects, where frequently certain words or meanings do not span all speakers. Since languages are ideal constructs based on the observation of dialects, it is of paramount importance to establish which words are likely shared by an entire linguistic community and which are used only by smaller groups. In the latter case, word usage descriptions can profit considerably from information as precise as possible, about geographical extension (region, province, district, city, even neighborhood), registry (colloquial, neutral, formal), frequency (current, past or a combination of both depending on the chronological span of the corpus), and other such variables.

Regionalisms (words used mainly in a particular subregion, such as *che* or *metegol* in Argentinian Spanish[1]) are commonly detected through surveys or transcriptions, using methods that depend more or less on the intuition and expertise of linguists (Almeida and Vidal, 1995; Labov, Ash, and Boberg, 2005). The results of this methodology are of great value to lexicographers, who need evidence to support the addition of a word into a regional dictionary, as well as the indication of where it is used. Information gathered with such methods has been used as lexical variables to compute similarities between dialects (Kessler, 1995; Nerbonne et al., 1996).

The emergence of social media and microblogging services has produced an unprecedented wealth of content, with a clear tendency towards informal or colloquial text generated by users. This fact has opened many opportunities for linguists due to the possibility of accessing geotagged contents, which provide valuable information about the location of users. In this sense, social media texts have been used to aid *lexical dialectology*, for example to establish "continuous" isoglosses (Gonçalves and Sánchez, 2014; Huang et al., 2016) or to study the diffusion of lexical change (Eisenstein et al., 2014), inter alia.

A problem closely related to lexical dialectology is *geolocation*, which maps words into regions or locations (Eisenstein, 2014). A possible way to evaluate dialectological models is to use them in geolocation algorithms; regionalisms can be seen as *location-indicative words* (Han, Cook, and Baldwin, 2012). Most previous work in word-centric geolocation algorithms (and lexical dialectology) relies on the observation of the frequency of a certain word, ignoring the number of users producing them. Also, to our knowledge very little work has been performed in Spanish on these topics.

In this work, we present an information-theoretic measure to detect regionalisms in social media texts, particularly on Twitter, and we test it against a dataset of tweets in Argentinian Spanish. Our contributions are twofold: a) we introduce a new metric based on Information Theory which can be seen as a mixture of *TF-IDF* and Information Gain; and b) we show that measuring the dispersion of users is a strong indicator of relevance, for both lexical dialectology and geolocation. We conduct our experiments on a dataset of tweets in Argentinian Spanish, with 81M tweets, 56K users, all balanced across the country's 23 provinces.

2 Previous Work

Most previous work in lexical dialectology consists in measuring the usage of words that are known a priori to be regional variants. These studies typically use information gathered from sources such as web searches (Grieve, Asnaghi, and Ruette, 2013) and manually-collected regionalisms (Ueda and Ruiz Tinoco, 2003; Kessler, 1995). Even papers that analyze data from Twitter (Huang et al., 2016; Gonçalves and Sánchez, 2014) still rely on words already known for the discovery of dialectal patterns.

Language evolves so quickly that it is important to detect these contrastive words automatically – or at least, to alleviate the efforts needed to detect them. Two types of approaches exist for this problem: *model-based* approaches and *metric-based* approaches (Rahimi, Baldwin, and Cohn,

[1] *Che*: interjection for getting the interlocutor's attention; *metegol*: mechanic game that emulates football (*futbolín*) (Academia Argentina de Letras, 2008).

	Total	Mean	SD
Words	647M	28.14M	6.64M
Tweets	80.9M	3.51M	0.91M
Users	56.2K	2.44K	0.04K
Vocabulary	7.5M	0.32M	0.04M

Table 1: Dataset summary. Total figures, along with province-level means and standard deviations.

2017). Model-based approaches use generative models to detect topics and regional variants (Eisenstein et al., 2010; Ahmed, Hong, and Smola, 2013). Typically, these are computationally expensive, which limits the amount of data that may be processed. Metric-based approaches compute statistics for each word or expression, and use them to create rankings (Cook, Han, and Baldwin, 2014; Chang et al., 2012; Jimenez et al., 2018; Monroe, Colaresi, and Quinn, 2008). These rankings are subsequently evaluated by checking external sources of regionalisms, such as dictionaries. In the following section, we compare our metrics to those proposed by Han, Cook, and Baldwin (2012): Term-Frequency Inverse Location Frequency (TF-ILF) and Information-Gain Ratio.

Text-based geolocation can be seen as the inverse problem of lexical dialectology: while dialectology maps regions into text, geolocation maps text into regions (Eisenstein, 2014). Thus, a reasonable way of assessing the performance of a method for discovering regional words is to use it as a feature-selection method for a geolocation classifier, as proposed by Han, Cook, and Baldwin (2012). In the present work, we use provinces as our unit of study (see Section 3), but finer grained geolocation could be performed by using an adaptive grid (Roller et al., 2012).

Rahimi, Cohn, and Baldwin (2017) propose a different approach to this problem. They train a multilayer perceptron with a bag-of-words as input to geolocate users. Intermediate layers serve as vector representations to perform lexical analysis by analyzing proximities in the embedding space.

Information Theory is the basis for many of these methods (Han, Cook, and Baldwin, 2012; Roller et al., 2012; Chang et al., 2012). Other uses of information theoretic measures include telling whether a hashtag is promoted by spammers by analyzing its dispersion in time and in users (Cui et al., 2012; Ghosh, Surachawala, and Lerman, 2011), and also to discover valuable features from user messages on Twitter for sentiment analysis and opinion mining (Pak and Paroubek, 2010). The metrics discussed in the next section use this concept of measuring the entropy of the users of a particular word.

3 Materials

The territory of Argentina is divided into 23 *provinces* and the autonomous city of Buenos Aires, with populations ranging from 127,000 (Tierra del Fuego Province) to 15 million (Buenos Aires Province), according to the 2010 National Census.[2] Provinces are further subdivided into *departments*, which in some cases are called *partidos* or *comunas*.

To gather our data, we first collected information of all departments in Argentina from the 2010 National Census and conducted a lookup through the Twitter API for users with location matching those departments. Even though location fields in Twitter are not very reliable (Hecht et al., 2011), given that we restrict our search to a fixed number of department names, we observe that most of the potential noise is reduced. We used the Python library *tweepy* to interact with the Twitter API.[3]

For each of the retrieved users, we successfully downloaded their entire tweetlines. Tweets were tokenized using *NLTK* (Bird, Klein, and Loper, 2009). Hashtags and mentions to users were removed; the remaining words were downcased; and identical consecutive vowels were normalized up to three repetitions ("woaaa" instead of "woaaaaaa"). Table 1 summarizes the collected dataset, and Figure 1 shows the distributions of tweets per user and tweet length.

It is well known that the Twitter vocabulary tends to be very noisy with lots of contractions, non-normal spellings (e.g., vocalizations), typos, etc. (Kaufmann and Kalita, 2010). For this reason, we decided to take into account only words occurring more than 40 times and used by more than 25 users (these values were chosen empirically). This removes about 1% of the total words and shrinks the vocabulary from 2.3 million words to around 135,000 words.

[2] `https://www.indec.gov.ar`

[3] `https://www.tweepy.org`

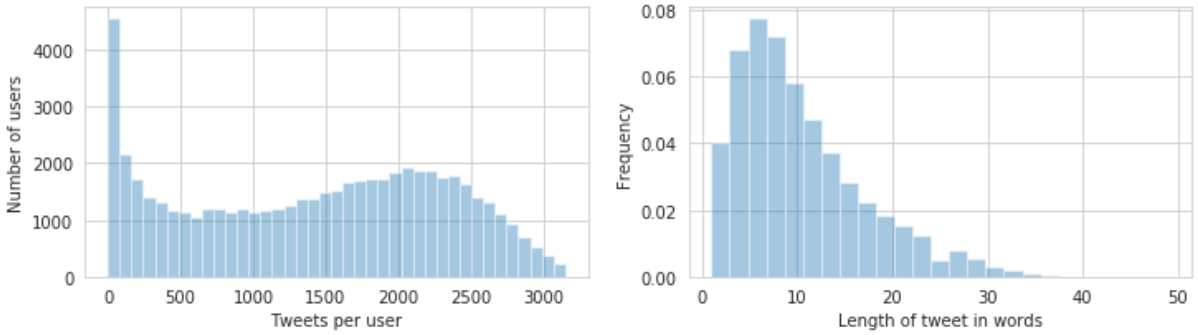


Figure 1: Dataset distributions: Number of tweets per user (left) and words per tweet (right).

4 Method

We can think of a regionalism as a word whose usage is not uniform across the territory – i.e., whose concentration is higher in a specific region. With this in mind, we aim to measure these *disorders* in word usage – or, more precisely, the *entropy* of words (Shannon, 1948).

In general, words with high entropy are more likely to be pronouns, connectors and other closed-class words, whereas their low-entropy counterparts are usually nouns, verbs, adjectives and adverbs with fuller semantic content (Montemurro and Zanette, 2002; Montemurro and Zanette, 2010). Also, words with high entropy (i.e., high disorder) can be regarded as used evenly across the country. On the other hand, low-entropy words are used with higher frequency in a few specific locations.

Let $l_1, l_2, \ldots l_N$ be our locations, and ω_1, ω_2, $\ldots \omega_M$ our vocabulary. If O_j refers to the event of occurrence of word ω_j, then $p(l_i|O_j)$ denotes the probability that word w_j occurred in location l_i.

We next define the *word-count entropy* as

$$H_{\text{words}}(\omega_j) = -\sum_{i=1}^{N} p(l_i|O_j) \cdot \log p(l_i|O_j). \quad (1)$$

Note that this measure does not take into account the actual frequency of words. For instance, if two words ω_1 and ω_2 occur only in one particular location, but ω_1 is much more frequent than ω_2, both words will still have the same entropy according to Equation 1.

In a similar fashion to *tf-idf* and inspired by Montemurro and Zanette (2010) and Han, Cook, and Baldwin (2012), we define measure $I_{\text{words}}(\omega)$ for word ω as follows:

$$I_{\text{words}}(\omega) = p(\omega) \cdot (\log N - H_{\text{words}}(\omega)), \quad (2)$$

where $\log N$ is the maximum possible value of $H_{\text{words}}(\omega)$ (Shannon, 1948), and $p(\omega)$ is the relative frequency of ω in the corpus ($0 \leq p(\omega) \leq 1$). In this way, $I_{\text{words}}(\omega)$ will be high for frequent words that accumulate in just a few locations.

Another important aspect of a word is the amount of people that use it (Cui et al., 2012). Assuming we now sample Twitter users, let U_j be the event that a particular user uses word ω_j. Then $p(l_i|U_j)$ denotes the probability that the location of a user is l_i given the fact that s/he uses word ω_j. We define the *user-count entropy* as

$$H_{\text{users}}(\omega_j) = -\sum_{i=1}^{N} p(l_i|U_j) \cdot \log p(l_i|U_j) \quad (3)$$

and the following metric of ω,

$$I_{\text{users}}(\omega) = q(\omega) \cdot (\log N - H_{\text{users}}(\omega)), \quad (4)$$

where $q(\omega)$ is the proportion of users who mentioned ω in the corpus ($0 \leq q(\omega) \leq 1$). Note that $I_{\text{users}}(\omega)$ will be high for words mentioned by several users who accumulate in just a few locations.

According to Zipf's Law, the counts of the most frequent words are orders of magnitude higher than the counts of the remaining words – a phenomenon that is also true when counting users of words. So the $p(\omega)$ and $q(\omega)$ terms in Equations (2) and (4) become a problem as words with high frequencies overcome their low entropies. To alleviate this, we performed a normalization on the word frequency as follows. Let M_ω be the most frequent word, that is,

$$M_\omega = \arg\max_{\omega \in W} \#\omega, \quad (5)$$

where $\#\omega$ denotes the total number of occurrences of ω in our dataset. Then, the *Nor-*

malized log-frequency of word occurrences is defined as

$$n_{\text{words}}(\omega) = \frac{\log(\#\omega)}{\log(\#M_w)}. \quad (6)$$

Words with very high frequency differ little in their values of $n_{\text{words}}(\omega)$. We define analogously the *Normalized log-frequency* of user mentions n_{users}. Hence, we rewrite Equations (2) and (4) and arrive at the final definition of our two metrics as

$$I_{\text{words}}(\omega) = n_{\text{words}}(\omega)(\log(n) - H_{\text{words}}(\omega)) \quad (7)$$

$$I_{\text{users}}(\omega) = n_{\text{users}}(\omega)(\log(n) - H_{\text{users}}(\omega)) \quad (8)$$

We call the first metric *Log-Term Frequency Information Gain (LTF-IG)* and the second one *Log-User Frequency Information Gain (LUF-IG)*. Summing up, **words with high values of LTF-IG or LUF-IG are cantidates for being regionalisms** – words that occur much more often in a certain region than in the rest of the country.

We subsequently sort all words in our dataset relative to these metrics, thus obtaining two word rankings: *Word-Count Ranking* and *User-Count Ranking*. The words that appear in the first positions of a ranking are those with high values for the metric, and thus more likely to be regionalisms.

4.1 Lexicographic Validation

With these rankings, a team of lexicographers from Academia Argentina de Letras performed a linguistic validation of the first thousand words according to each metric. This qualitative analysis consisted in a detailed study, word by word, to determine if the word in question is part of the lexical repertoire of a community of speakers.

Proper and place names (toponyms) were excluded –as is usual in lexicography– although many words in this class had high values for our metrics. Potential toponyms were automatically highlighted to facilitate their manual exclusion by lexicographers.

To perform the linguistic validation, lexicographers were provided with tables containing counts for each word and province: number of users, number of occurrences and normalized frequency (occurrences per million words). Also, samples of tweets containing these words were provided when necessary. The goal of this manual validation was to identify not only words used exclusively or mainly in a region, but also words used there with a different meaning.

As a result of this process, every word in the top-1000 of each ranking was annotated with '1' if it had lexical relevance as a regionalism, or '0' if it had not. Lastly, lexicographers performed a characterization of the words marked as regionalisms, according to the linguistic phenomenon they represent. The outcome of these procedures is described in Section 5.

4.2 Feature Selection for Geolocation

To indirectly assess the usefulness of our metrics, we used each as a feature-selection method to train geolocation classifiers. This means that, instead of using the entire bag-of-words as input for a geolocation algorithm, we consider a smaller subset of the vocabulary. This dimensionality reduction of the feature space is aimed at boosting the classifier performance.

This approach to geolocation can be described as "word-centric", as it uses lexical information from tweets to predict a location (Zheng, Han, and Sun, 2018). But we emphasize that we are interested in *user* geolocation, not tweet geolocation. Thus, the units considered here are all the tweets from individual users. We randomly selected 10,000 users from our dataset – 7,500 for training and 2,500 for testing.

For reference, we compare our results to those obtained using the *Information Gain Ratio (IGR)* metric (Han, Cook, and Baldwin, 2012; Cook, Han, and Baldwin, 2014): if L is a random variable denoting the location of a given occurrence of word ω_i, then the *Information Gain* of ω_i is

$$\begin{aligned} IG(\omega_i) &= H(L) - H(L|\omega_i) \\ &\propto P(\omega_i) \sum_{j=1}^{m} P(c_j|w_i) \log P(c_j|w_i) \\ &\quad + P(\overline{w_i}) \sum_{j=1}^{m} P(c_j|\overline{w_i}) \log P(c_j|\overline{w_i}) \end{aligned}$$

where $P(\overline{\omega_i})$ denotes the probability that ω_i does not occur. Then, $IGR(\omega_i)$ is defined as

$$IGR(\omega_i) = \frac{IG(\omega_i)}{IV(\omega_i)} \quad (9)$$

Rank	Word	User
1	ushuaia	chivil
2	rioja	ush
3	chivilcoy	poec
4	bragado	**malpegue**
5	viedma	**aijue**
6	logroño	tolhuin
7	chepes	vallerga
8	oberá	**yarca**
9	**cldo**	blv
10	tdf	portho
11	riojanos	jumeal
12	breñas	**sinf**
13	choele	plottier
14	gallegos	kraka
15	tiemposur	fsa
16	fueguinos	**bombola**
17	chilecito	**yarco**
18	blv	sanagasta
19	ush	wika
20	merlo	obera

Table 2: Top 20 words for the two metrics. Words in bold have lexicographic interest as regionalisms.

where IG is normalized by

$$IV(\omega) = -P(\omega)\log P(\omega) - P(\overline{\omega})\log P(\overline{\omega})).$$

We also calculate IGR with respect to the user frequencies of a word (which we abbreviate "user frequencies" for the sake of simplicity), in a similar way to Equation 4. As a baseline for our feature selection methods, we also calculate *Term-Frequency Inverse Location Frequency (TF-ILF)*, which consists in sorting our terms first by Location Frequency (in ascending order) and then by Term-Frequency (in descending order).

Summing up, five feature selection methods are tested as feature selection for geolocation: *TF-ILF*, *LTF-IG*, *LUF-IG*, basic *IGR*, and *User IGR*. We train Multinomial Logistic Regressions using the top N% words as features, and test against the 2.5K held out users. Performance is assessed using accuracy and mean distance between capital cities of each province – a fairly good estimate, since most of the population concentrates around those cities.

5 Results

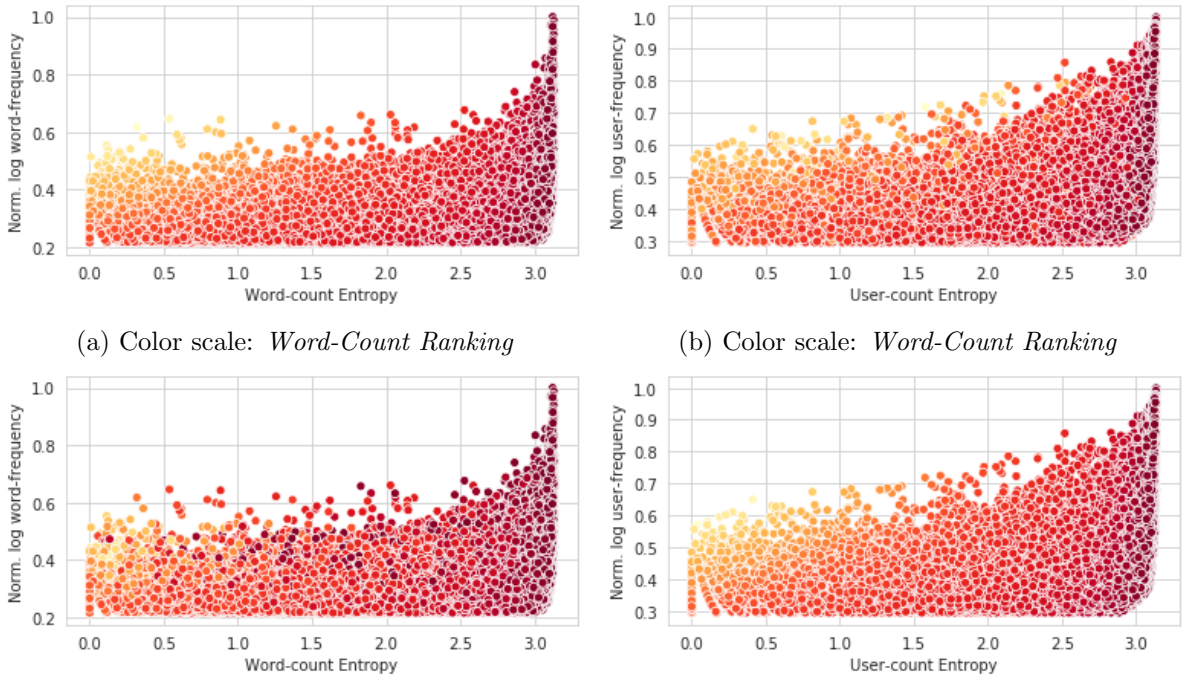
Table 2 shows the top-20 words calculated with each metric. Many are toponyms: *chivil, ush, blv, tolhuin, kraka, sanagasta, wika* refer to towns, cities and local clubs. Also, some words refer to gentilics (*riojanos, fueguinos*), or local institutions (*POEC*). Some of these words emerge as regionalisms: *yarca/yarco, aijue, sinf, cldo, bombola, malpegue*. We observe that the two rankings even share many words: *User-Count* and *Word-Count* have an overlap of 63% in the top thousand words.

Figure 2 shows four three-dimensional scatter plots. A dot in these plots corresponds to an individual word in our corpus, and is placed along the horizontal axes according to its word- or user-count entropy ($H_{\text{words}}(\omega)$ and $H_{\text{users}}(\omega)$, respectively). Along the vertical axes, each dot is located following its corresponding word or user frequency ($n_{\text{words}}(\omega)$ and $n_{\text{users}}(\omega)$). Additionally, each dot is colored according to the position of the word in one of our rankings using a chromatic scale, such that the lighter the dot, the higher the word's rank. For clearer visualization, word rankings are also shown in logarithmic scale.

Figure 2a shows that words higher in the *Word-Count Ranking* (in lighter color) tend to appear closer to the upper-left corner of the plot – that is, such words are more frequent and their mentions are concentrated in fewer regions. Figure 2d shows a very similar thing, now with respect to the number of users that mention the words: words higher in the *User-Count Ranking* are mentioned by a larger number of users from fewer regions. These two figures display a gradient from the upper-left corner (words ranked higher, in lighter color) to the lower-right corner (words ranked lower, in darker color).

Figure 2b uses horizontal and vertical axes corresponding to users (H_{users} and n_{users}), but colors each word with respect to the *Word-Count Ranking*. Here we can observe a slight perturbation in the gradient: there are words far from the left-corner that have light colors. From this, we understand that there are words with high *Word-Count Ranking* that have low *User-Count Ranking*.

Likewise, Figure 2c uses *User-Count Ranking* to color the points, and word axes H_{user} and n_{user}. The perturbation in the gradient is clearer in this plot; many words appear high in the *Word-Count Ranking* (closer to the top-left corner, see Figure 2a) but low in *User-Count Ranking* (darker color).



(a) Color scale: *Word-Count Ranking*

(b) Color scale: *Word-Count Ranking*

(c) Color scale: *User-Count Ranking*

(d) Color scale: *User-Count Ranking*

Figure 2: Scatter plots showing words (dots) along three dimensions. Horizontal axes: word-count entropy H_{words} (left plots) or user-count entropy H_{users} (right plots). Vertical axes: normalized log word frequencies n_{words} (left plots) or user frequencies n_{users} (right plots). Color: log word rank according to *Word-Count* (top plots) or to *User-Count* (bottom plots); lighter color means higher rank.

To further inspect this phenomenon, we searched for words that have large differences in the logarithm of *Word-Count Ranking* and *User-Count Ranking*. The logarithm reduces the difference between words ranked very high (e.g., between the word at position 10,000 and another in position 20,000) and amplifies the difference when one of the ranks is low and the other is high. A close examination of these words and their tweets showed that they were produced by bots (news and metheorological accounts, or accounts using tools to gain more followers) or in small niches of fans of some celebrity. From the top-100 words sorted by this difference, only one ranks higher in users than in words.

Summing up, when a word has a high *User-Count Ranking*, it also tends to have a high *Word-Count Ranking*. The reverse is not true, however, as words produced by a small number of accounts would not rank well with respect to users. Thus, the *User-Count Ranking* successfully discards words coming from automatic agents, as already done in Cui et al. (2012).

Word	Word Rank	User Rank
rioja	2	2499
vto	27	28179
hoa	81	83717
contextos	88	71290
cardi	32	23756
agraden	107	75042
hemmings	59	40227
ushuaia	1	565
tweeted	43	21342
precipitación	66	31042

Table 3: Top 10 words with the largest gaps between log word rank and log user rank.

5.1 Lexicographic Validation

The first thousand words in the *Word-Count Ranking* were manually analyzed by the lexicographers, who marked 21.9% as likely regionalisms. Likewise, from the first thousand words in the *User-Count Ranking*, 30.2% were marked as being lexicographically relevant. **This validation suggests that considering user-frequency dispersion is more relevant when assessing a word as a regionalism.**

Lexical characterization is illustrated in Table 4, which displays a few examples of groups of regionalisms found thanks to this methodology. A special note is reserved for the group of *indigenisms*, where a number of words were found coming from the *Guaraní* language (for instance, *mitaí, angá, angaú, nderakore*) and also from *Quechua* (*ura*). It is worth mentioning that the regions of the words derived from *Guaraní* – spoken in Northeastern Argentina, Paraguay, Bolivia and Southwest of Brazil – coincide with the region delimited by Vidal de Battini (1964).

Word	Region	Meaning
Colloquialisms		
culiado	Córdoba	asshole
chombi	Mendoza	poor in quality
carnasas	Neuquén	not classy, inelegant
bolasear	Cuyo	to bullshit
aprontar	E. Ríos	to get ready
Indigenisms		
ura	Northwest	vagina (quechua)
mitaí	Guaranitic	boy
angá	Guaranitic	unfortunate
Regional realities		
piadinas	San Juan	roll (food)
tarefero	Misiones	yerba mate worker
POEC	Neuquén	high School exam
Interjections		
aijue	Formosa	surprise
yirr	Corrientes	joy
aiss	Formosa	annoy
jiaa	Corrientes	yeehay
Ortographic variations		
pesao	Northwest	pesado
ql	Northwest	culiado
uaso	Córdoba	guaso
Regional Morpheme		
raraso	Córdoba	very strange (raro)
tardaso	Córdoba	very late (tarde)

Table 4: Examples of regionalisms found in the manual analysis. Each group corresponds to a subjective category found by the lexicographers during the annotation process.

5.2 Feature Selection for Geolocation

Moving on to the results of our second validation procedure, Figure 3 displays the performance of the different feature selection methods when used to train a discriminative classifier. Horizontal axes represent the percentage of top words selected, and the vertical axes represent the mean distance error in 3a and the accuracy in the case of 3b.

LUF-IG obtains the best performance in the user geolocation task, and stabilizes in a plateau at roughly 3.75% of top words used. It outperforms its word-frequency version LTF-IG and both IGR metrics. Table 5 displays the results of using the full bag of words (baseline) versus using the different feature selection methods with 5,000 top words.

When comparing our metrics, we note that the ones based on user-frequencies obtain a better performance than their word-frequency counterparts. This is more apparent in the case of LTF-IG and LUF-IG, but can also be observed for IGR metrics.

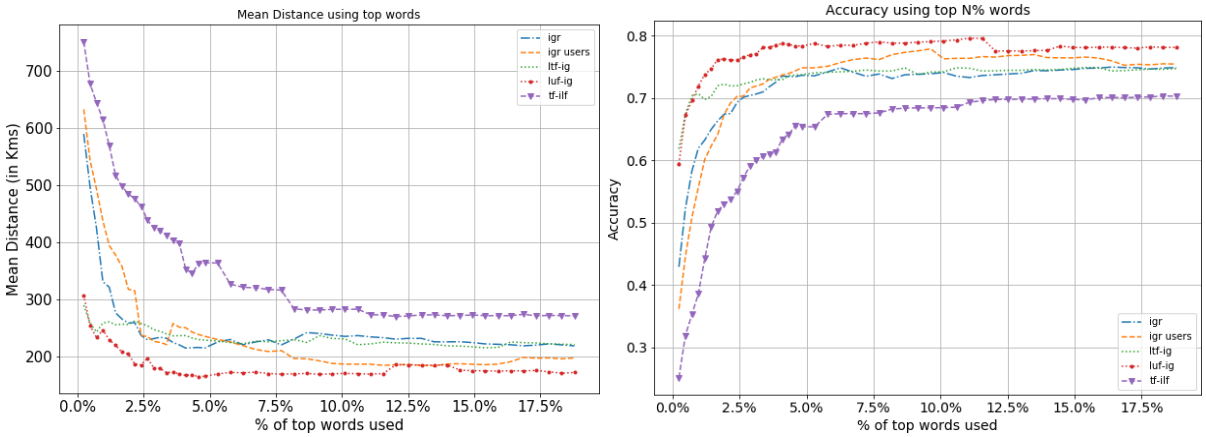
6 Discussion

Of the proposed metrics, *User-Count Metric* proved to be the most promising one. It successfully removed from the top of the ranking words likely to come from automatic agents or from small niches of users, and a manual lexicographic validation confirmed that this ranking contained more regionalisms than the *Word-Count Metric*. Further, using this metric as a feature selection method for geolocating users also showed a significative improvement over other metrics – both its word-frequency counterpart and IGR metrics from Han, Cook, and Baldwin (2012). This strongly suggests that measuring the dispersion of users of a certain word is a very informative indicator – both in lexicographic and in geolocation terms – backing what was already proposed in previous work to detect spam on Twitter (Cui et al., 2012).

The proposed metric was developed in the context of analyzing regional colloquialisms.

Features	Accuracy	Mean Distance
All	0.383	599.8
TF-ILF	0.654	363.3
IGR-Words	0.736	214.2
IGR-Users	0.748	234.7
LTF-IG	0.737	227.9
LUF-IG	**0.784**	**164.9**

Table 5: Performance of the different feature selection methods when using the top-5000 words.



(a) Mean distance error in user geolocation

(b) Accuracy in user geolocation

Figure 3: Comparison of the metrics when used as feature selection methods for geolocation. Vertical axes show the percentage of the top words used as features to train a Multinomial Logistic Regresion, and vertical axes display the performance of each respective classifier. Figure a uses mean distance error as y-axis (less is better) and Figure b uses accuracy (more is better)

This area of the lexicon is most elusive, since its impact on any printed medium arrives noticeably late – and in many cases it never reaches it at all. Colloquialisms are a class of words hardly found in any other media. Our best performing metric marked as relevant several words that were already listed in the *Diccionario del Habla de los Argentinos* (Academia Argentina de Letras, 2008), a fact that confirms the usefulness of both our metric and Twitter data in general for this task.

An outstanding subgroup of words found in the analysis are those coming from the *Guaranitic* region, in Northeastern Argentina. In particular, three words have already been proposed for addition to the aforementioned dictionary: *angá, angaú, mitaí*. This case is emblematic because it shows how this type of approach can help overcome the intrinsic limitations of doing regional lexicography. When lexicographers are native to only one of the different dialects of the region included in a projected dictionary, the probability of properly detecting and defining words of other dialects is slim or depends on mere chance. As the team of lexicographers expressed when confronted with these three words related to Guaraní heritage, those very robust normalized frequencies across a significant portion of the territory of Argentina would have otherwise remained unknown. Instead of including them in the next edition of the dictionary that attempts to describe all regional lexical items in the country, they would have remained unregistered, thus perpetuating a serious omission.

As our focus was in detecting lexical variations within provinces, we paid no attention to spatial granularity. If a better granularity were necessary in the analysis, adaptive partitioning could be used (Roller et al., 2012) to improve geolocation and to find localisms within provinces. Although previous work (Vidal de Battini, 1964) indicates that most provinces do not have large dialectal variations within them, this is something that would need to be explored and confirmed in future work.

Also, these techniques should be tested against other datasets, such as those used in (Roller et al., 2012; Han, Cook, and Baldwin, 2012), to further confirm that they outperform other feature selection methods.

7 Conclusions

In this work, we developed and compared two novel metrics useful for detecting regionalisms in Twitter based on Information Theory. One was based on the word frequency (*Log Term Frequency-Information Gain, LTF-IG*) and the other on the user frequency of a word (*Log user frequency-Information Gain, LUF-IG*). These metrics may be seen as a mixture of previous information-theoretic measures and classic *TF-IDF*.

We evaluated their performance in two ways. First, a team of lexicographers man-

ually assessed the presence of regionalisms in the first thousand words ranked by each metric. Second, we tested the metrics as feature-selection methods for geolocation algorithms, for which we also tested against metrics from previous works (Han, Cook, and Baldwin, 2012; Cook, Han, and Baldwin, 2014). In both evaluation types, the metric built upon user frequencies (*LUF-IG*) yielded the best results, suggesting that the number of users of a word is very informative – perhaps even more than simple word frequency.

This method has aided lexicographers in their task, allowing them to propose the addition of a number of words into the *Diccionario del Habla de los Argentinos*. The work behind this particular dictionary relies on a collaborative effort based on the intuition of scholars and lexicographers that identify regionalisms used mainly (seldom exclusively) within Argentina's borders by carefully parsing over a diversity of sources. Therefore, using Twitter to automatically detect regionalisms does not limit itself to avoiding most of this manual work, which, in and of itself, would already be a sizeable contribution. Since a considerable portion of the lexical repertoire of a community does not make its way across to published materials (which make most of the 300 millions words included to date in, for example, CORPES XXI (Real Academia Española, 2013)), the possibility of creating lists of words that are likely to be regional, based on actual utterances written by users, opens a way of shedding light onto entire pockets of lexical items that would remain otherwise chronically underrepresented in dictionaries. Even when a regional word is published, and then included in corpora, the task of appropriately isolating it remains largely unchanged, given that the word has to be previously identified in order to then take advantage of the statistical information available.

This work defines Argentinian provinces as the regional units of analysis, but this could be changed in order to repeat the analysis at different granularity levels. In this way, it might be possible to study intra-provincial dialectal differences (e.g., at the department level, see Section 3), although the limited precision of the geolocation of Twitter users may complicate this task. And it would definitely be possible to detect contrastive words across larger regions, for example to study Spanish in all its geographical variants.

A further challenge triggered by this work is the detection of regions with different dialectal uses (Gonçalves and Sánchez, 2014) but using features obtained in a semisupervised fashion with these metrics. This would allow to assess the validity of the dialectal regions of Argentina proposed by Vidal de Battini in 1964 (Vidal de Battini, 1964). Spatial and temporal information could be also explored, particularly finer-grained locations. Regarding geolocation, the proposed metrics should also be tested against other datasets to evaluate its performance as a feature selection method.

Acknowledgments

This work was funded in part by CONICET, Universidad de Buenos Aires, and Universidad Torcuato Di Tella. We thank Edgar Altszyler, Mariela Sued, and Federico Plager for helpful discussions, and our anonymous reviewers for valuable suggestions.

References

Academia Argentina de Letras. 2008. *Diccionario del habla de los argentinos*. Emecé Editores.

Ahmed, A., L. Hong, and A. J. Smola. 2013. Hierarchical geographical modeling of user locations from social media posts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 25–36. ACM.

Almeida, M. and C. Vidal. 1995. Variación socioestilística del léxico: un estudio contrastivo. *Boletín de filología*, 35(1):50.

Atkins, B. S. and M. Rundell. 2008. *The Oxford guide to practical lexicography*. Oxford University Press.

Bird, S., E. Klein, and E. Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Chang, H.-w., D. Lee, M. Eltaher, and J. Lee. 2012. @Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 111–118. IEEE Computer Society.

Cook, P., B. Han, and T. Baldwin. 2014. Statistical methods for identifying local dialectal terms from GPS-tagged documents. *Dictionaries: Journal of the Dictionary Society of North America*, 35(35):248–271.

Cui, A., M. Zhang, Y. Liu, S. Ma, and K. Zhang. 2012. Discover breaking events with popular hashtags in twitter. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM 12, pages 1794–1798, New York, NY, USA. ACM.

Eisenstein, J. 2014. Identifying regional dialects in online social media. In *School of Interactive Computing Faculty Publications*. Georgia Institute of Technology.

Eisenstein, J., B. O'Connor, N. A. Smith, and E. P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1277–1287. Association for Computational Linguistics.

Eisenstein, J., B. O'Connor, N. A. Smith, and E. P. Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Ghosh, R., T. Surachawala, and K. Lerman. 2011. Entropy-based classification of retweeting activity on Twitter. *arXiv preprint arXiv:1106.0346*.

Gonçalves, B. and D. Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PloS one*, 9(11):e112074.

Grieve, J., C. Asnaghi, and T. Ruette. 2013. Site-restricted web searches for data collection in regional dialectology. *American speech*, 88(4):413–440.

Han, B., P. Cook, and T. Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. *Proceedings of COLING 2012*, pages 1045–1062.

Hecht, B., L. Hong, B. Suh, and E. H. Chi. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 237–246. ACM.

Huang, Y., D. Guo, A. Kasakoff, and J. Grieve. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255.

Jimenez, S., G. Dueñas, A. Gelbukh, C. A. Rodriguez-Diaz, and S. Mancera. 2018. Automatic Detection of Regional Words for Pan-Hispanic Spanish on Twitter. In *Ibero-American Conference on Artificial Intelligence*, pages 404–416. Springer.

Kaufmann, M. and J. Kalita. 2010. Syntactic normalization of Twitter messages. In *International conference on natural language processing, Kharagpur, India*.

Kessler, B. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*, pages 60–66. Morgan Kaufmann Publishers Inc.

Labov, W., S. Ash, and C. Boberg. 2005. *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter.

Monroe, B. L., M. P. Colaresi, and K. M. Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Montemurro, M. A. and D. H. Zanette. 2002. Entropic analysis of the role of words in literary texts. *Advances in complex systems*, 5(01):7–17.

Montemurro, M. A. and D. H. Zanette. 2010. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153.

Nerbonne, J., W. Heeringa, E. Van den Hout, P. Van der Kooi, S. Otten, W. Van de Vis, et al. 1996. Phonetic distance between Dutch dialects. In *CLIN VI: proceedings of the sixth CLIN meeting*, pages 185–202.

Pak, A. and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Rahimi, A., T. Baldwin, and T. Cohn. 2017. Continuous representation of location for geolocation and lexical dialectology using

mixture density networks. *arXiv preprint arXiv:1708.04358*.

Rahimi, A., T. Cohn, and T. Baldwin. 2017. A neural model for user geolocation and lexical dialectology. *arXiv preprint arXiv:1704.04008*.

Real Academia Española. 2013. Banco de datos (CORPES XXI) [online]. *Corpus del español del siglo XXI (CORPES)*.

Roller, S., M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Ueda, H. and A. Ruiz Tinoco. 2003. Varilex, variación léxica del español en el mundo: Proyecto internacional de investigación léxica. In *Pautas y pistas en el análisis del léxico hispano (americano)*. Iberoamericana Vervuert, pages 141–278.

Vidal de Battini, B. E. 1964. El español en la Argentina. Technical report, Argentina.

Zheng, X., J. Han, and A. Sun. 2018. A survey of location prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1652–1671.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 63-72 recibido 31-03-2022 revisado 17-05-2022 aceptado 29-05-2022

Reflexive pronouns in Spanish Universal Dependencies: from annotation to automatic morphosyntactic analysis

Los pronombres reflexivos en las Universal Dependencies en español: desde la anotación hacia el análisis morfosintáctico automático

Jasper Degraeuwe, Patrick Goethals
Ghent University (Belgium)

Jasper.Degraeuwe@UGent.be
Patrick.Goethals@UGent.be

Abstract: In this follow-up article of Degraeuwe and Goethals (2020), we present the annotation scheme used to reannotate the 7298 potentially reflexive pronouns included in the Universal Dependencies Spanish AnCora v2.6 treebank, which resulted in significant modifications for the "Case" feature (100% changed) and dependency relations (87% changed). Next, we evaluate the performance of spaCy v3.2.2 and Stanza v1.3.0 (both trained on AnCora v2.8, and thus based on our reannotations) on the AnCora v2.8 test set, which yielded weighted F1 scores up to 0.88 and 0.98 for the "Case" and "Reflex" features, respectively, and up to 0.71 for the dependency relations. Finally, the error analysis of the spaCy results underlines the (generalisation) potential of the model, but also reveals some of the remaining issues in the automatic morphosyntactic analysis of reflexive pronouns in Spanish, such as determining if expletive relations denote an impersonal, passive or inherently reflexive use.
Keywords: reflexive pronouns, *se*, Universal Dependencies, morphosyntactic tagging and parsing.

Resumen: En este artículo de seguimiento de Degraeuwe y Goethals (2020), presentamos el esquema de anotación utilizado para reanotar los 7298 pronombres potencialmente reflexivos incluidos en el Universal Dependencies Spanish AnCora v2.6 *treebank*, lo cual resultó en un significativo número de modificaciones para la característica (*feature*) de "Case" (el 100% cambiado) y las relaciones de dependencia (el 87% cambiado). A continuación, evaluamos el desempeño de spaCy v3.2.2 y Stanza v1.3.0 (ambos entrenados en AnCora v2.8, y, por tanto, basados en nuestras reanotaciones) en el *set* de prueba de AnCora v2.8, lo cual dio como resultado puntuaciones de F1 ponderado de hasta 0,88 y 0,98 para las características de "Case" y "Reflex", respectivamente, y de hasta 0,71 para las relaciones de dependencia. Por último, el análisis de errores de los resultados de spaCy subraya el potencial (generalizador) del modelo, pero también desvela algunos de los problemas pendientes en el análisis morfosintáctico automático de los pronombres reflexivos en español, como por ejemplo determinar si las relaciones de dependencia expletivas son de carácter impersonal, pasivo o inherentemente reflexivo.
Palabras clave: pronombres reflexivos, *se*, Universal Dependencies, etiquetado y análisis gramatical morfosintáctico.

1 Introduction

As Natural Language Processing (NLP) tools such as spaCy (spacy.io) and Stanza (stanfordnlp.github.io/stanza/) are becoming more and more accessible (also for a non-expert audience, see e.g. Altinok (2021) and Vasiliev (2020)), automatic morphosyntactic analysis has been integrated into a wide range of text-based applications. By means of a simple programming script, for example, raw corpora

ISSN 1135-5948. DOI 10.26342/2022-69-5

can be transformed into intelligent resources containing morphosyntactic information. These "enriched corpora" can then be used as input for corpus query tools or language learning environments, enabling their users to perform much more fine-grained queries.

To train their (morphosyntactic) taggers and (syntactic) parsers, NLP tools usually make use of treebanks as reference data. One of the most well-known initiatives concerned with the construction of such treebanks is the Universal Dependencies (UD) project, established in 2014. In a nutshell, UD aims at developing "cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective" (https://universaldependencies.org/introduction, retrieved 22 February 2022; see also Nivre et al. (2016)). Together with the growing number of languages included in the project (v2.9 contains 217 treebanks in 122 languages), the standardised, cross-linguistically consistent approach of UD has led to increasing usage of UD treebanks not only for the development of NLP tools, but also for research purposes (see de Marneffe et al. (2021, p. 304) for an overview).

However, the UD initiative is also a "constantly improving effort" (Martínez Alonso and Zeman, 2016), implying that the annotation guidelines are regularly updated and fine-tuned over the successive releases of the treebanks. Furthermore, within UD there remain several annotation issues, which may be problematic from both a cross-linguistic and an intra-linguistic perspective. In a previous article (Degraeuwe and Goethals, 2020), we addressed one of those pending issues, namely the annotation of the potentially reflexive pronouns *me*, *te*, *nos*, *os* and *se* (see also Marković and Zeman, 2018) in the UD Spanish AnCora treebank (Martínez Alonso and Zeman, 2016; Taulé, Martí and Recasens, 2008). These are very frequent items in Spanish, with *se* alone occurring in almost 30% of the sentences in the AnCora treebank and being ranked eleventh in the list of most common lemmas in CORPES XXI (Real Academia Española de la Lengua, 2022).

The annotation proposal described in Degraeuwe and Goethals (2020) was revised by the UD contributor responsible for the AnCora treebank (see sections 2.1.1 and 2.1.2 for the details of the original and revised annotation schemes), after which all potentially reflexive pronouns in the treebank were reannotated and the resulting 7298 changes were pushed to the UD project (visible from v2.7 onwards). In 2021, both spaCy (with v3.0) and Stanza (with v1.2.0) released thoroughly updated versions of their tools trained on UD v2.7 or higher (thus including the reannotated reflexive pronouns). In this follow-up contribution, we will carry out both a quantitative and qualitative analysis of how well the tools perform on morphosyntactically analysing Spanish reflexive pronouns, and we will identify the key remaining issues (see section 3).

2 Literature overview

2.1 Reflexives in Spanish Universal Dependencies

The UD framework provides three main annotation layers by which linguistic constructions can be progressively defined and differentiated: a morphosyntactic Part-Of-Speech (POS) tag (limited to a universal set of seventeen tags), a syntactic dependency relation such as subject or (in)direct object, and a feature set containing additional lexical and grammatical properties (e.g. number or person in the case of pronouns).

2.1.1 Annotation scheme as proposed in Degraeuwe and Goethals (2020)

The original proposal (see Table 1) arose from the notion that, as NLP tools become more accessible, more theoretical linguists will use them and evaluate their linguistic accuracy and granularity. Consequently, we not only focused on improving annotation consistency in order to increase tagger/parser accuracy, but also took into account the (cross-)linguistic analyses made in non-computational linguistics (Croft et al., 2017; Maldonado, 2008; Mendikoetxea, 1999; Peregrín Otero, 1999).

First, the pronouns were disambiguated according to their general reflexive character, distinguishing between *me veo* ('I see myself') and *me ven* ('they see me'). In the latter group, the dependency relation is defined as "obj" or "iobj" (*me dieron algo*, 'they gave me something').

Secondly, the reflexive uses were assigned one of the dependency labels "obj", "iobj", "expl:impers", "expl:pass" and "expl:pv". This

means that reflexive and non-reflexive "obj" and "iobj" have the same dependency label but are distinguished by the "Reflex" feature, which is absent in the case of non-reflexives. Reflexive "obj" and "iobj" are further subdivided according to their genuine reflexive versus reciprocal use.

Thirdly, the umbrella category "expl:pv" consists of three subgroups, namely constructions with corresponding transitive verbs, constructions which show an alternation with intransitive verbs, and constructions without corresponding (in)transitive verbs. The first group of "transitivity-based" reflexive constructions is then further subdivided by assigning different combinations of feature sets. These feature sets overlap with other "non-expl:pv" constructions, showing their shared characteristics.

With this annotation proposal, many annotation inconsistencies were resolved (e.g. up to 30% and 60% of false positives of "expl:pass" and "iobj", respectively). Furthermore, the proposal also provided a more fine-grained and informative categorisation, as the previous taxonomy (AnCora ≤ v2.6) did not allow distinguishing between, for example, passive (*en este volumen se ofrecen textos sobre*, 'in this volume texts are provided about') and reflexive uses (*María se ofrece para hacerse cargo del bebé*, 'María offers herself to take care of the baby') of the same verb, or between passive (*se incautaron las armas*, 'the guns were seized') and inherently reflexive constructions (*la policía se incauta de las armas*, 'the police seized the guns'). In all these cases, *se* was labelled as "obj", and no differences were to be found between the feature sets of the *se* instances, nor between the feature sets of their verbal heads.

	Features			
	Pronoun		Verb	
	Case	Reflex	Voice	
Reflexive uses				
expl:pass	Acc	Reflex	Pass	(a) *la noticia se publicó*
obj	Acc	Reflex	Act	(b) *Pedro se ve en el espejo*
	Acc	Rcp	Act	(c) *Pedro y Juan se vieron en la calle*
iobj	Dat	Reflex	Act	(d) *Pedro se quita la ropa*
	Dat	Rcp	Act	(e) *Pedro y Juan se dieron la mano*
expl:impers	-	Reflex	Act	(f) *se trabaja mucho*
expl:pv	***With corresponding non-reflexive transitive verb***			
	Acc	Reflex	Pass	(g) *el fenómeno se manifiesta*
	Acc	Reflex	Act	(h) *la gente se manifiesta*
	Acc	Rcp	Act	(i) *Pedro y Juan se ponen de acuerdo*
	Dat	Reflex	Act	(j) *Pedro se da cuenta de que …*
	Dat	Rcp	Act	[does not occur in Spanish]
	Com	Reflex	Act	(k) *Pedro se llevó el regalo*
	With corresponding non-reflexive intransitive verb			
	-	Reflex	Act	(l) *Pedro se muere*
	Without corresponding non-reflexive verb			
	Acc	Reflex	Act	(m) *Pedro se atreve a …*
Non-reflexive uses				
obj	Acc	-	-	(n) *me/te/nos/os ven*
iobj	Dat	-	-	(o) *me/te/nos/os/se lo dijeron*

Table 1: Overview of the annotation scheme for potentially reflexive pronouns in Spanish as proposed in Degraeuwe and Goethals (2020). Translations: (a) 'the news was published', (b) 'Pedro sees himself in the mirror', (c) 'Pedro and Juan see each other on the street', (d) 'Pedro takes off his clothes', (e) 'Pedro and Juan shake hands', (f) 'a lot of work is being done', (g) 'the phenomenon becomes clear', (h) 'people are demonstrating', (i) 'Pedro and Juan agree', (j) 'Pedro realises that …', (k) 'Pedro took the present with him', (l) 'Pedro dies', (m) 'Pedro dares to …', (n) 'they see me/you/us/you', (o) 'they told it to me/you/us/you/him/her'.

2.1.2 Annotation scheme used for AnCora ≥ v2.7 annotations

As the annotation scheme presented in section 2.1.1 included some drastic modifications, the proposed changes were first revised by the UD contributor responsible for the AnCora treebank before applying any reannotations. Based on this feedback, the reflexive – reciprocal distinction was dropped: since "Reflex" is currently a Boolean feature in all UD treebanks, the addition of a "Rcp" value would lower cross-linguistic consistency. Moreover, the distinction also showed to be a too subtle one to make for machine learning methods (tested with custom models built on spaCy v3.2.2 and Stanza v1.3.0 architectures).

Secondly, changing the "Voice" feature of the verbal head of the potentially reflexive pronoun was also discarded, again to give full priority to cross-linguistic consistency. Although these characteristics do seem to be recognisable for machine learning models (average weighted F1 test set scores of 0.78 for custom model trained with spaCy v3.2.2 architecture and 0.61 with Stanza v1.3.0), the "Voice=Pass" feature was primarily designed for annotating verbal paradigms which distinguish active from passive voice morphologically, which is not the case in Spanish.

Even though the modifications presented above slightly decrease granularity, the annotation proposal (see Table 2) remains very informative (five different dependency labels, accusative/dative/comitative case distinction and reflexive/non-reflexive use distinction). Moreover, the new annotation scheme now adheres very strictly to the UD principles and guidelines.

The reannotation of the potentially reflexive pronouns was implemented from AnCora v2.7 onwards. Since all "Case" values of pronouns were labelled as "{Acc, Dat}" in the v2.6 treebank, all of the 7298 pronouns present in the development, test and training sets received a new "Case" value: 5933 instances were reannotated as "Acc", 893 instances as "Dat" and 472 instances as "NA" (non-applicable, for non-cased instances). The comitative case ("Com") did not occur in the data.

	Features		
	Case	Reflex	
Reflexive uses			
expl:pass	Acc	Yes	(a) *la noticia se publicó*
obj	Acc	Yes	(b) *Pedro se ve en el espejo* (c) *Pedro y Juan se vieron en la calle*
iobj	Dat	Yes	(d) *Pedro se quita la ropa* (e) *Pedro y Juan se dieron la mano*
expl:impers	-	Yes	(f) *se trabaja mucho*
expl:pv	***With corresponding non-reflexive transitive verb***		
	Acc	Yes	(g) *el fenómeno se manifiesta* (h) *la gente se manifiesta* (i) *Pedro y Juan se ponen de acuerdo*
	Dat	Yes	(j) *Pedro se da cuenta de que …*
	Com	Yes	(k) *Pedro se llevó el regalo*
	With corresponding non-reflexive intransitive verb		
	-	Yes	(l) *Pedro se muere*
	Without corresponding non-reflexive verb		
	Acc	Yes	(m) *Pedro se atreve a …*
Non-reflexive uses			
obj	Acc	-	(n) *me/te/nos/os ven*
iobj	Dat	-	(o) *me/te/nos/os/se lo dijeron*

Table 2: Overview of the annotation scheme for potentially reflexive pronouns in Spanish used in AnCora ≥ v2.7.

AnCora ≥ v2.7 / Ancora v2.6	expl:impers	expl:pass	expl:pv	iobj	obj	**Total (Ancora v2.6)**
expl:pass	301	159	35	1	6	**502 (6.9%)**
iobj	1	17	253	152	43	**466 (6.4%)**
obj	54	2052	2927	628	665	**6326 (86.7%)**
other	0	3	0	1	0	**4 (0,1%)**
Total (AnCora ≥ v2.7)	**356 (4.9%)**	**2231 (30.6%)**	**3215 (44.1%)**	**782 (10.7%)**	**714 (9.8%)**	**7298**

Table 3: Overview of the dependency relation changes in Spanish UD AnCora ≥ v2.7 compared to v2.6 (dev + test + train).

Next, the "Reflex" value of 7108 pronouns (97%) remained unaltered: 6483 instances maintained their reflexive annotation ("Yes") and 625 instances their non-reflexive character ("NA"). However, 49 pronouns were changed from reflexive to non-reflexive, while the value of 141 instances was modified the other way around from non-reflexive to reflexive.

Finally, 6322 of the 7298 potentially reflexive pronouns (almost 87%) received a new dependency label. A detailed, quantitative overview of the corresponding changes is presented in Table 3 (note that "expl:impers" and "expl:pv" do not occur in the v2.6 treebank). The statistics show a fundamental shift from "obj" as the predominant label to a more dispersed distribution, with "expl:pv" and "expl:pass" being the most important labels. In other words, the reannotation shows that reflexive pronouns usually express an expletive use, more specifically an inherently reflexive use ("expl:pv") or a passive one which blurs the subject role ("expl:pass").

2.2 Reflexives in machine learning

To our knowledge, to date no studies have been performed which focus on the performance of machine learning models at tagging potentially reflexive pronouns in Spanish based on UD treebank data. On non-UD data, however, some experiments have been carried out. In Aldama García and Barbero Jiménez (2021), for example, a machine learning approach is adopted to predict the dependency label of *se* in a one-per-sentence setup, for which a custom "*se* corpus" was compiled containing 2140 sentences from CORPES XXI (Real Academia Española de la Lengua, 2022). The corpus was annotated according to a four-category annotation scheme, containing the "se-mark" (for cases of valency reduction, such as passive and impersonal constructions), "expl" (for pure pronominal predicates or emphatic contexts), "iobj" (for indirect objects) and "obj" (for direct objects) labels. Next, nine different machine learning classifiers were applied to the test set of the corpus, with pre-trained language models based on a transformers architecture obtaining the best performance (macro F1 score of 0.7).

Results such as these indicate that, to a certain extent, recent machine learning methods are able to successfully distinguish different uses of the potentially reflexive pronoun *se*. To implement them in real-life scenarios, approaches as in Aldama García and Barbero Jiménez (2021), which are based on a language-specific setup and require a self-compiled and annotated set of training and test data, can be integrated as a custom component for that specific language in NLP tools such as spaCy and Stanza. This way, the morphosyntactic information offered by the tool's tagger and parser can be complemented by the output of the task-specific model.

However, the creation of such models is a very time-consuming operation, especially for non-computational linguists. Therefore, in section 3, we will study the potential of the default taggers and parsers included in spaCy and Stanza (which are trained on UD data), and analyse if they would need to be complemented by a task-specific model and where exactly (i.e. for which labels) issues arise.

3 Automatic morphosyntactic analysis of potentially reflexives pronouns

From section 2.1.2 it can be concluded that, in theory, any NLP tool trained on the reannotated treebank as input data should be able to perform a more fine-grained morphosyntactic analysis of potentially reflexive pronouns in Spanish. To evaluate the validity of this claim, we apply the large pretrained Spanish model of spaCy v3.2.2 ("es_core_news_lg") and the default pretrained

		Case			Reflex		Dependency relation				
		Acc	Dat	NA	Yes	NA	expl:impers	expl:pass	expl:pv	iobj	obj
#instances		**452**	**47**	**42**	**504**	**37**	**30**	**171**	**254**	**36**	**50**
spaCy	F1	0.94	0.62	0.57	0.99	0.88	0.51	0.75	0.8	0.6	0.37
	macro avg	0.71			0.93		0.6				
	weighted avg	0.88			0.98		0.71				
Stanza	F1	0.9	0.46	0	0.98	0.78	0.5	0.75	0.76	0.56	0.23
	macro avg	0.45			0.88		0.56				
	weighted avg	0.79			0.97		0.68				

Table 4: Results (macro and weighted F1) of automatic morphosyntactic analysis of potentially reflexive pronouns in the AnCora v2.8 test set using spaCy v3.2.2 and Stanza v1.3.0.

Spanish model of Stanza v1.3.0, which are both trained on the UD Spanish AnCora v2.8 training and development sets, to the corresponding AnCora v2.8 test set. This test data includes 668 potentially reflexive pronouns, of which 127 instances are clitic forms such as *se* in *la gente va a la calle a manifestarse* ('people go to the streets to demonstrate'). As spaCy does not include a multiword tokeniser (which is required to split words with clitics into so-called "subword tokens" and then analyse these separate tokens instead of the entire word form), clitic forms will be excluded from the evaluation in order to obtain comparable results. Table 4 presents a detailed overview of the morphosyntactic analysis, with F1 as the evaluation metric and the number of instances for each label included in the "#instances" row.

Finally, some architectural characteristics of the NLP tools should be highlighted: in the spaCy pipeline, the tagger and parser components listen to the same word embedding component but do not share any information between them, implying that the features and dependency relations are predicted independently of each other. Stanza, however, does take into account information from the tagger when training its dependency parser, which means that the Stanza dependency relation predictions partially depend on the feature predictions.

For the "Case" and "Reflex" features, satisfying results are obtained, especially with spaCy (weighted F1 scores of 0.88 for "Case" and 0.98 for "Reflex"). Stanza, however, does not seem to be able to recognise non-cased uses (see "NA" column), which correspond to reflexive pronouns with "expl:impers" as the dependency label and to "expl:pv" relations with verbs for which a corresponding non-reflexive intransitive counterpart exists (see also Table 2).

As far as the dependency relations are concerned, the automatic morphosyntactic analysis achieves relatively good results as well, with weighted F1 scores of 0.71 for spaCy and 0.68 for Stanza. Compared to the top macro F1 score of 0.7 reached in Aldama García and Barbero Jiménez (2021), both spaCy (0.6) and Stanza (0.56) perform worse, although it should be observed that Aldama García and Barbero Jiménez (2021) distinguish only four instead of five categories and exclusively focus on *se* as potentially reflexive pronoun (and not on *me*, *te*, *nos* and *os*). Next, the low scores for the "expl:impers" and especially "obj" category have to be highlighted. A first possible explanation for this lower performance could be the limited number of training instances in the training and developments sets: 268 for "expl:impers" (5.07%) and 444 for "obj" (8.4%). To gain more in-depth insights into this matter, and into the errors made by NLP tools in general, an additional analysis is performed based on the contingency tables included in Table 5, which zero in on the performance of spaCy, the best-performing tool. The error analysis will also include a qualitative component, with special attention to the generalisation potential of the tool (i.e. if it has learnt to make predictions based on patterns, not just to predict the most frequent label for each word form).

For the "Case" errors, three main findings can be extracted from the results:

1. Predicting the correct case of *me*, *te*, *nos* and *os* when they are used in accusative case is challenging (see (p) and (s) in Table 6 for some examples): together, these pronouns account for 29 of the 452 accusative instances, and 13

Case

correct \ predicted	Acc	Dat	NA	Total
Acc	435	14	3	452
Dat	19	28	0	47
NA	23	1	18	42
Total	477	43	21	541

Reflex

correct \ predicted	Yes	NA	Total
Yes	496	8	504
NA	2	35	37
Total	498	43	541

Dependency relation

correct \ predicted	expl:impers	expl:pass	expl:pv	iobj	obj	other	Total
expl:impers	14	9	5	1	1	0	30
expl:pass	7	138	24	1	0	1	171
expl:pv	4	39	201	7	3	0	254
iobj	0	5	3	25	3	0	36
obj	0	7	16	14	13	0	50
Total	25	198	249	47	21	1	541

Table 5: Results (contingency tables) of automatic morphosyntactic analysis of potentially reflexive pronouns in the AnCora v2.8 test set using spaCy v3.2.2.

of those 29 cases received the wrong "Dat" prediction (which corresponds to 13 of the 17 errors made for the accusative label). Importantly, 12 of those instances had also received a wrong dependency label (namely "iobj" instead of "obj" or "expl:pv").

2. Predicting the correct case of *se* when it is used in dative case also entails challenges (see (q) in Table 6 for an example): *se* accounts for 24 of the 47 dative instances, and 15 of those 24 cases were labelled wrongly as accusative (corresponding to 15 of the 19 errors made for this label).
3. As for the non-cased instances, the errors seem to indicate that the model does not just naïvely link labels to verbal heads, since in sentences with *irse*/*marcharse* ('to leave'), which frequently occur in the training data and under all circumstances receive the "NA" label, 5 incorrect but also 4 correct predictions were to be found. This finding can be considered evidence that the model has developed a kind of generalisation procedure, although it thus results in the introduction of some errors in the case of *irse*/*marcharse*. In this regard, it also appears that the generalisation as such is not entirely successful either, since several of the "Case" errors correspond to reflexive pronoun – verbal head combinations which (almost) do not occur in the training (and development) data, as was the case with *advertir*. For this verb, which only occurs once as a verbal head (i.e. the verbal form on which the potentially reflexive pronoun is syntactically dependent) in the training set, the test sentence (r) (see Table 6) was wrongly predicted as "Acc", meaning that the model was not able to generalise, in this particular case at least, from similar examples with other verbal heads (e.g. *contratar* as in *se contrata a alguien* 'someone was given a contract', which occurs three times in the training data).

Next, the (few) errors made for the "Reflex" feature (see (s) in Table 6) usually correspond to instances of *me*, *te*, *nos* and *os* for which the

model also wrongly predicted both the case (usually dative instead of accusative case) and the dependency relation (usually "iobj" instead of "expl:pv" or "obj"). Especially the wrong "iobj" prediction provides a plausible explanation for the error in the "Reflex" label, as in the training data non-reflexive "iobj" instances are twice as frequent as reflexive "iobj" instances.

Thirdly, the error analysis of the dependency relation predictions led to again three main findings:

1. Errors in one of the "expl" categories almost always correspond to one of the two other "expl" labels (14 of the 16 errors in "expl:impers", 31/33 in "expl:pass" and 43/53 in "expl:pv"). In other words, the model has no problem in identifying expletive uses of potentially reflexive pronouns, but assigning the right expletive subcategory seems to be a less straightforward operation from a machine learning point of view (see sentences (q) and (r) in Table 6).
2. For "iobj", 5 of the 11 errors are "expl:pass" predictions. Looking at the sentences, it appears that the model is not able to predict the "iobj" label for reflexive pronouns which co-occur with an explicit subject and direct object, as in the examples (t) and (u) in Table 6.
3. Predicting the correct dependency label of *me*, *te*, *nos* and *os* when they are used as direct object also poses challenges to spaCy's machine learning model (see (p) and (s) in Table 6): together, these pronouns account for 19 of the 50 "obj" instances, and 17 of those 19 cases received a wrong dependency relation (which corresponds to 17 of the 37 errors made for the "obj" label). Moreover, all of the 14 cases where an "iobj" instance was predicted instead of "obj" were to be found amongst those 17 errors, highlighting the sometimes fuzzy boundary between *me*, *te*, *nos* and *os* acting as direct or indirect object.

Finally, as a general, overarching observation it should be noted that the generalisation potential of the model, which was already briefly addressed in the discussion of the "Case" errors, also comes to the fore with pronoun – verbal head combinations which have multiple possible uses. A good case in point is the combination with the verbal head *tratar*, which occurs 111 times as "expl:impers" and 2 times as "expl:pass" in the training and development data. Despite the imbalanced distribution in the training data, the test sentence containing *los temas se tratarán* ('the topics will be treated') still got labelled correctly as "expl:pass", which hints at the fact that the model has leveraged "knowledge" from other "expl:pass" training examples to arrive at this correct prediction. Still other evidence of

Sentence	Case		Reflex		Dependency relation	
	correct	predicted	correct	predicted	correct	predicted
(p) […] nos trataron muy mal […]	Acc	Dat	NA	NA	obj	iobj
(q) […] las carreteras catalanas se cobraron 16 vidas […]	Dat	Acc	Yes	Yes	expl:pv	expl:pass
(r) Si se hubiera advertido a la gente […]	NA	Acc	Yes	Yes	expl:impers	expl:pass
(s) […] no me engaño a creer en la existencia de […]	Acc	Dat	Yes	NA	obj	iobj
(t) […] Beckenbauer se permite bromear […]	Dat	Acc	Yes	Yes	iobj	expl:pass
(u) […] el affaire Cristo-Rey se tomaba un respiro […]	Dat	Acc	Yes	Yes	iobj	expl:pass

Table 6: Selection of errors in automatic morphosyntactic analysis of potentially reflexive pronouns in the AnCora v2.8 test set using spaCy v3.2.2. Translations: (p) '[…] they treated us really badly […]', (q) '[…] Catalan roads claimed 16 lives […]', (r) 'If people had been warned […]', (s) '[…] I don't delude myself into believing in the existence of […]', (t) '[…] Beckenbauer affords himself to make jokes […]', (u) '[…] the Cristo-Rey affair took a breather […]'.

the generalisation potential can be found in the accuracy rates the model obtains for the 35 pronoun – verbal head combinations which do not occur at all in the training data: 73% for "Case", 97% for "Reflex" and 54% for the dependency relations.

4 Conclusion

In this article, we built upon Degraeuwe and Goethals (2020), in which a proposal was formulated to reannotate the potentially reflexive pronouns (*me*, *te*, *nos*, *os* and *se*) in the Universal Dependencies Spanish AnCora treebank. These items, and in particular *se*, occur very frequently in Spanish, and have also received much attention in non-computational linguistics (Croft et al., 2017; Maldonado, 2008; Mendikoetxea, 1999; Peregrín Otero, 1999). Taking into account that treebanks are used as reference data to train the models offered by state-of-the-art NLP tools such as spaCy and Stanza, we aim to contribute to improving the NLP-driven morphosyntactic analysis of potentially reflexive pronouns, and in doing so, also help creating higher-quality "enriched resources" which can be used as input for, amongst other applications, corpus query tools and language learning environments.

We presented the slightly modified annotation scheme used to reannotate the potentially reflexive pronouns included in the AnCora v2.6 treebank (7298 items in total; changes visible from v2.7 onwards), which resulted in label changes for all "Case" features, 3% of the "Reflex" features and 87% of the dependency relations. The application of spaCy v3.2.2 and Stanza v1.3.0 (both trained on AnCora v2.8, and thus based on our reannotations) to the AnCora v2.8 test set yielded promising results, hinting at the potential of using NLP-driven methods to perform fine-grained morphosyntactic analyses.

Finally, the error analysis on the spaCy results revealed some of the remaining issues in the automatic morphosyntactic analysis of potentially reflexive pronouns in Spanish (e.g. determining the right subcategory of expletive dependency labels), but also underlined the (generalisation) potential of the underlying model.

Although more than satisfactory performance levels were achieved (weighted F1 up to 0.88 for "Case", 0.98 for "Reflex" and 0.71 for dependency relations), there is still room for improvement, especially for the prediction of dependency relations. Therefore, future work could consist in studying if a task-specific model (as in Aldama García and Barbero Jiménez (2021)) can complement the default taggers and parsers of NLP tools in order to push performance. Furthermore, it is worth considering to implement rule-based predictions for a fixed set of verbs which always yield the same labels when functioning as verbal head of a reflexive pronoun (e.g. for *irse* and *marcharse*), and to define rules which determine the feature values for a given dependency relation (e.g. *if* "expl:pv" is predicted as the dependency relation *then* the "Reflex" feature should always be "Yes"). In spaCy, for instance, such specific rules can be easily implemented thanks to the "attribute ruler" component, which manages mappings and exceptions at token level.

Acknowledgements

This research has been carried out as part of a PhD fellowship on the IVESS project (file number 11D3921N), funded by the Research Foundation – Flanders (FWO).

References

Altinok, D. 2021. *Mastering spaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem*. Packt.

Croft, W., D. Nordquist, K. Looney, and M. Regan. 2017. Linguistic Typology meets Universal Dependencies. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 63-75, Indiana University (Bloomington, Indiana).

de Marneffe, M-C., C.D. Manning, J. Nivre and D. Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2): 255-308.

Degraeuwe, J., and P. Goethals. 2020. Reflexive pronouns in Spanish Universal Dependencies. *Procesamiento del Lenguaje Natural*, 64: 77-84.

Maldonado, R. 2008. Spanish middle syntax: A usage-based proposal for grammar teaching. In S. De Knop and T. De Rycker (eds.) *Cognitive Approaches to Pedagogical*

Grammar, 155-196. Mouton De Gruyter, Berlin.

Marković, S. and D. Zeman. 2018. Reflexives in Universal Dependencies. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 131-146, Oslo University (Oslo).

Martínez Alonso, H. and D. Zeman. 2016. Universal Dependencies for the AnCora treebanks. *Procesamiento del Lenguaje Natural*, 57: 91-98.

Mendikoetxea, A. 1999. Construcciones inacusativas y pasivas. In *Gramática descriptiva de la lengua española*, 2: 1575-1629. Espasa Calpe.

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659-1666, European Language Resources Association (Portorož).

Peregrín Otero, C. 1999. Pronombres reflexivos y recíprocos. In *Gramática descriptiva de la lengua española*, 1: 1427-1518. Espasa Calpe.

Real Academia Española de la Lengua. 2022. Banco de datos (CORPES XXI) [online]. Corpus del Español del Siglo XXI (CORPES) https://www.rae.es/recursos/banco-de-datos/corpes-xxi. Accessed date: 22/02/2022

Taulé, M., M. A. Martí, and M. Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (Marrakech).

Vasiliev, Y. 2020. *Natural Language Processing with Python and spaCy: A Practical Introduction.* No Starch Press.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 73-82 recibido 31-03-2022 revisado 18-05-2022 aceptado 20-05-2022

Multi-label Text Classification for Public Procurement in Spanish

Clasificación multi-etiqueta de textos de licitaciones públicas en español

María Navas-Loro, Daniel Garijo, Oscar Corcho
Ontology Engineering Group, AI.nnovation Space, Universidad Politécnica de Madrid
mnavas@fi.upm.es, daniel.garijo@upm.es, ocorcho@fi.upm.es

Abstract: Public procurement accounts for a 14% of the annual budget of the different governments of the European Union. In Europe, contracting processes are classified using Common Procurement Vocabulary codes (CPVs), a taxonomy designed to facilitate statistical reporting, search and the creation of alerts that can be used by potential bidders. CPVs are commonly assigned manually by public employees in charge of contracting processes. However, CPV classification is not a trivial task, as there are more than 9,000 different CPV categories, which are often assigned following heterogeneous criteria. In this paper we have created a CPV classifier that uses as an input the textual description of the contracting process, and assigns CPVs from the 45 top-level CPV categories. We work only with texts in Spanish, although our approach may be easily extended to other languages. Our results improve the state of the art (10% F1-score improvement) and are available online.
Keywords: CPV, Multi-label Classification, Public Procurement, Hierarchical Classification.

Resumen: Las licitaciones públicas suponen el 14% del presupuesto anual de la Unión Europea. En Europa, los procesos de contratación se clasifican usando la taxonomía Common Procurement Vocabulary (CPVs), diseñada para facilitar la generación de estadísticas, las búsquedas y la creación de alertas que puedan utilizar los posibles licitadores. Los códigos CPV suelen ser asignados manualmente por los empleados públicos encargados del proceso de contratación. Sin embargo, la clasificación de textos de acuerdo con estos códigos no es trivial, pues existen más de 9000 CPVs y no siempre se siguen los mismos criterios para su asignación. En este artículo se propone un clasificador que utiliza como entrada la descripción textual del proceso de contratación, y produce códigos de entre las 45 categorías de CPV más generales de la jerarquía. Trabajamos sólo con textos en español, aunque nuestro enfoque puede extenderse fácilmente a otros idiomas. Los resultados obtenidos superan el estado del arte (10% de mejora en F1), y se encuentran disponibles online.
Palabras clave: CPV, Clasificación Multi-etiqueta, Licitaciones Públicas, Clasificación Jerárquica.

1 Introduction

Public authorities in the European Union spend around 14% of the yearly Gross Domestic Product (around 2 trillion euros) purchasing services, utilities and supplies.[1] Access to this data is crucial for enabling a single digital market in Europe, as well as for accountability and transparency. Hence many governments provide this data in their open data portals as well as in `data.europa.eu`, and a number of platforms have been developed to improve both the efficiency and transparency in public procurement[2] (Soylu et al., 2022).

Common Procurement Vocabulary codes (CPVs)[3] help classify public procurement processes in the European Union across dif-

[1] `https://ec.europa.eu/growth/single-market/public-procurement_en`

[2] `https://opentender.eu/es/about/about-opentender`

[3] `https://simap.ted.europa.eu/web/simap/cpv`

ISSN 1135-5948. DOI 10.26342/2022-69-6

ferent languages. Thanks to CPVs, decision makers can easily explore contracting processes across Europe, and companies from different countries may use them to detect procurement processes of interest, independently of the country of origin.

Each public procurement process must be classified with at least one CPV. However, manual CPV classification presents three main challenges. First, there are thousands of possible codes (more than 9000), some of them with similar purposes, making it difficult for those assigning or curating them to decide which codes better suit a specific process. Second, countries with different official languages and countries with more than one official language, such as Spain or Belgium, often have offers in different languages (e.g., Catalan, Basque, Castilian, etc.). Offices from different regions therefore follow different classification guidelines. Third, CPVs are organized in a hierarchy, and thus annotated at different levels of granularity according to the annotator's or department's criteria. For example, the CPV "Pharmaceutical products" (3360000) shown in Figure 1 is often overgeneralized, instead of using more specific codes that shed more light in the type of purchase. This issue is in fact reflected in the European Union Policy Handbook, where the need of suggesting users to select more specific CPV codes is stressed (European Commission, 2020).

In order to address these issues and ease the assignment of CPV codes to procurement processes, this paper presents an approach to automatically assign high-level CPV codes (i.e., the 45 most general categories) to a procurement process. In this paper, we assume that we have the textual description of the process and that the text is in Spanish. Different methods have been tested to this end, outperforming the previous available results for the Spanish language. We expect this research line will help public procurement practitioners in assigning CPV codes in a more homogeneous manner by providing suggestions that humans can use in their decision process.

The rest of the paper is organized as follows. Section 2 introduces the CPV classification problem in detail, explaining the rationale behind each part of the codes. Section 3 summarizes the related work done in the context of multi-label text classification, as well as existing approaches for CPV classification in Spanish. Section 4 describes how the corpus used to train our classifier was developed, while in Section 5 we outline our approach. Finally, Section 6 details the results obtained by the different classification techniques used, and Section 7 concludes our work.

2 Background

The Common Procurement Vocabulary (CPV) allows classifying public procurement processes with a homogeneous code that represents the need and main object of the requested contract. Several CPV codes may be used to describe a single offer. The format of these CPV codes follows a five-level tree structure comprising the following digits:

- The first two digits identify the divisions (XX000000)
- The first three digits identify the groups (XXX00000)
- The first four digits identify the classes (XXXX0000)
- The first five digits identify the categories (XXXXX000)
- The following three digits give a greater degree of precision within each category (00000XXX)

A ninth check digit serves to verify the previous digits, and has no meaning by itself (00000000-Y).

Therefore, the task of automatically classifying CPVs increases in complexity the more digits we aim to predict. The current official list of CPVs has 9454 possible codes, grouped into 45 different divisions, 317 groups, 1321 classes and 3704 categories. In this paper we focus in classifying CPVs at the division level.

3 Related Work

While text classification has been widely explored in the literature (Aggarwal and Zhai, 2012; Minaee et al., 2021), multi-label classification for the Spanish language has received less attention so far. The main difference between the multi-label text classification case presented in this paper and other popular problems like sentiment analysis is the amount of possible labels. Sentiment analysis labels correspond to certain degrees

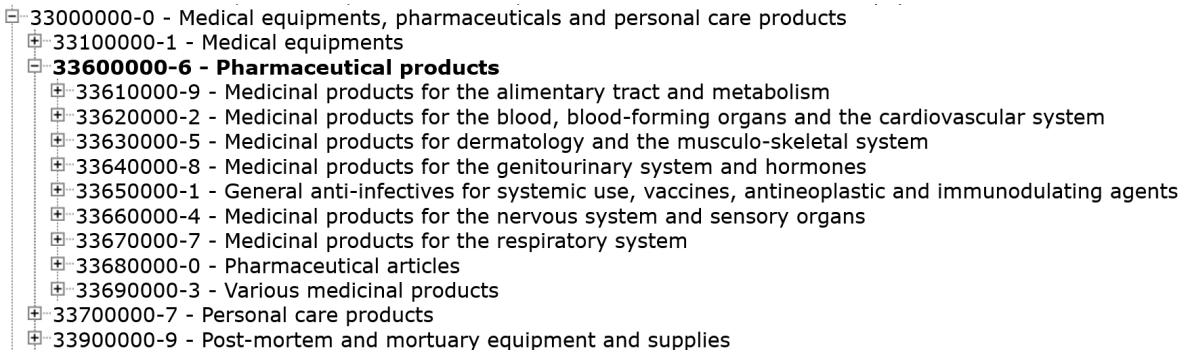


Figure 1: Excerpt of the tree-structure of CPV code 33600000, "Pharmaceutical products", extracted from `http://www.cpv.enem.pl/en/33600000-6`.

of positive and negative emotions, or to a taxonomy of emotions, whilst CPV labels may contain up to thousands of possible options. In order to target this kind of problems, a new subtask has been defined inside multi-label text classification: *extreme multi-label text classification* (XMTC) (Liu et al., 2017).

XMTC addresses the problem of assigning to a document its most relevant subset of class labels from an extremely large label collection (Liu et al., 2017). The work by Gargiulo et al. (2019) analyzes the impact of using different word embedding models in Deep Learning targeting extreme multi-label classification. Their approach uses Convolutional Neural Networks (CNN) to classify 27,775 hierarchical labels in the biomedical domain. Similarly, Liu et al. (2017) compared CNN to other approaches in XMTC, such as KNN-based approaches like SLEEC (Bhatia et al., 2015) or tree-based methods like FastXML (Prabhu and Varma, 2014). Finally, Chang et al. (2020) proposed a scalable framework to fine-tune Deep Transformer models that performed well in different XMTC datasets.

Regarding specific previous work on CPV classification, one of the main results was the multilingual model built by Kaan Görgün.[4] This model categorizes public procurement descriptions in multiple languages among 45 different division labels, with an F1 Score of 0.68. Industrial approaches have also targeted the CPV code classification problem, such as the solution developed by the data science consultancy uData (Deloitte, 2020), using a hierarchical nested approach consisting of one model to predict the first two digits of the CPV code, 50 models to predict the third code (depending on the first model results) and 250 additional models to predict the fourth digit. Other approaches in the literature include a deep learning sequence-processing regression algorithm (also containing several classifiers, considering different aspects of CPVs) (Suta, 2019), or the approach by Ahmia (2020), who used Linear SVMs in order to predict the first two digits of the CPV codes. SVMs were also used in Kayte and Schneider-Kamp (2019). Since the only model available for reuse and evaluation for the Spanish language is the one from Kaan Görgün, we use it as a baseline for comparison against our approach, making both training data and model results available to the community.

4 Creating a Spanish CPV Corpus

We created our training corpus with open data from historical public procurement from the Spanish Treasury's website (Hacienda[5]). We decided to use data from 2019, in order to avoid including later data that may have been influenced by public procurement related to COVID19 pandemics. Procurement processes' metadata were processed from their original format (Atom Syndication Format[6]) using different scripts available in our paper repository (Navas-Loro, Garijo, and Corcho, 2022).[7] Document pre-processing included the following stages:

1. **Information extraction** from all the

[4] `https://huggingface.co/MKaan/multilingual-cpv-sector-classifier`

[5] `https://www.hacienda.gob.es/es-ES/GobiernoAbierto/Datos%20Abiertos/Paginas/LicitacionesContratante.aspx`

[6] `https://www.w3.org/2005/Atom`

[7] `https://github.com/oeg-upm/cpv-classifier`

information contained in the Atom documents. We only retrieved the textual description of the offers and the different CPV codes assigned to them. This is represented as a CSV file in order to ease its further processing.

2. **Duplicate deletion** and trim of the descriptions. Additionally, we only keep texts in Spanish (to this aim we used fastText's language identification functionality[8]).
3. **Train/test dataset division**, in order to make the dataset more manageable, we split it into train and test sets (70/30) before uploading it to our public code repository.
4. **In-code preprocessing**. An additional set of scripts were used to remove rows with no CPV code assigned and generalize CPV codes to the division level, which is the one we use in our experiments.

The result of the first two steps are two csv files, available in our repository. The code used for all processing scripts can also be found in the same location. Figure 2 shows the distribution for each of the 45 division labels, which are clearly unbalanced. The most frequent label ('45', that represents the division 'works') is present in 16128 instances of the the training set, while label '76' is only present in 13 instances.

5 Approach

We addressed CPV classification in a hierarchical manner: instead of creating a classifier for nine thousand labels, we took advantage of the hierarchical structure of the CPVs and created a classifier for the 45 available divisions (first two digits). We believe this to be a good first step due to the training data available for most categories.

The only model openly available to perform this task is the model from Kaan Görgün (from now, MKaan) mentioned in the Related Work section. This model also targeted just the first two digits of the CPV code, so we use it as a baseline to compare the different approaches we have tested.

In order to perform multi-label classification, several approaches can be used. We can use algorithms adapted to the task, such as decision trees or random forests, or we can also use binary classifiers like Naïve Bayes or SVM and then apply different strategies so that they serve for multi-label classification. Another option is to fine-tune existing transformers, as done in the approach by MKaan. We briefly present below the different approaches we tested.

5.1 Classical Techniques

We tried the following classifiers:

Naïve Bayes (Minsky, 1961) has been widely used for text classification (İşgüder-Şahin, Zafer, and Adah, 2014), specially for sentiment analysis and SPAM classification. Although this algorithm relies on probability independence, it works very well even when this assumption is not met.

SVM Support Vector Machines (SVM) (Boser, Guyon, and Vapnik, 1992) are linear classifiers that define an hyperplane in order to discriminate among classes. SVM have been frequently used for multiclass classification tasks.

SVM with RBF kernel Besides testing the linear version of SVM, we also evaluated the performance of an SVM with the Radial Basis Function as kernel, that is:

$$rbf_\gamma = e^{-\gamma\|x-x'\|^2} \quad (1)$$

with parameter $\gamma \geq 0$.

Decision Trees (Quinlan, 1986) are an intuitive way to classify instances. In our implementation we used the sklearn optimized version of the CART algorithm.[9]

Random Forests (Breiman, 2001) are a tree-based ensemble approach to classification that overcomes most of the problems with decision trees, such as high variance. Due to this robustness they have been frequently used for Extreme Multi-label Classification (Siblini, Kuntz, and Meyer, 2018).

K-Nearest Neighbours (K-NN) (Hand, 2007) is widely used for multi-label classification (Zhang and Zhou, 2007). The idea behind K-NN is to check the K labeled instances that are the closest to the new instance and classify it with the most common label from these neighbours.

[8] https://fasttext.cc/docs/en/language-identification.html

[9] https://scikit-learn.org/stable/modules/tree.html#tree-algorithms-id3-c4-5-c5-0-and-cart

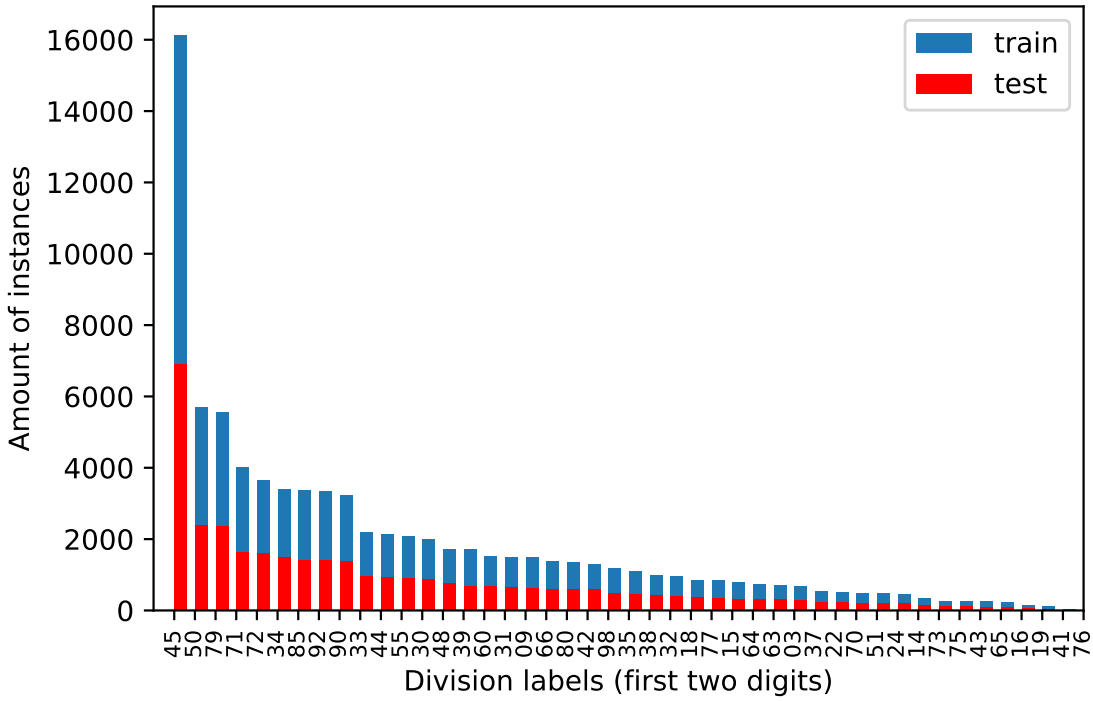


Figure 2: Bars (y axis) represent the amount of instances per division label (x axis). Blue bars represents the amount of labels in the training set, while red bars represent the number of instances in the evaluation set.

AdaBoost (Freund and Schapire, 1997) is a meta-estimator that fits different versions of models using boosting (i.e., different versions of the training dataset). We used the implementation defined in Hastie et al. (2009): AdaBoost-SAMME.

For all these approaches we used the Term Frequency - Inverse Document Frequency (TF-IDF) technique for vectorization, allowing n-grams with $n = 3$. For those algorithms that do not support multi-label classification, we decided to use the One-vs-the-rest (OvR) or One-vs-all strategy, frequently used for multiclass classification, where one binary classifier per label is built in order to decide if an instance should be classified with that label or not.

5.2 RoBERTa fine-tuned approach

In addition to the aforementioned classical approaches, we also decided to fine-tune a transformed-based model for the Spanish language, namely *RoBERTa-base-bne* (Gutiérrez-Fandiño et al., 2021), on a dataset derived from Spanish Public Procurement documents from 2019.

RoBERTa-base-bne is a transformer-based masked language model based on the RoBERTa model and pre-trained using the largest Spanish corpus known to date (570GB), compiled from the annual web crawlings performed by the National Library of Spain (Biblioteca Nacional de España) from 2009 to 2019.[10]

Table 1 summarizes the hyperparameters used in the fine-tuning process, performed using the HuggingFace transformers library. The whole training process can be reproduced using the notebook '`fine-tuned-roberta-for-spanish-cpv-codes.ipynb`' in our code repository.

6 *Evaluation*

This section describes how we evaluated the results obtained with the different approaches, and discusses them.

6.1 Metrics

We use two sets of metrics in our evaluation. First, we use *general* metrics such as the Area Under the ROC Curve (ROC AUC), F1-score and accuracy. Second, we use multi-label specific metrics, i.e., coverage error and

[10] `https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne`

Parameter	Value
learning rate	$2 * 10^{-5}$
train batch size	8
eval batch size	8
seed	42
optimizer	adam
epochs	10

Table 1: Summary of the hyperparameters used for training the RoBERTa fine-tuned model used in our analysis.

Label Ranking Average Precision. We briefly describe all these metrics below.

6.1.1 General Metrics

The metrics used that are not specific to multi-label classification are the following:

Area Under the ROC Curve (AUC): measures the capability of a classifier to distinguish between classes. The higher the AUC, the better the model can make the distinction among classes.

F1-score: harmonic mean between precision and recall, widely adopted to monitor both metrics at the same time.

Accuracy: fraction of predictions that the model classified correctly.

6.1.2 Coverage Error

The coverage error computes the average number of labels that have to be included in the final prediction such that all true labels are predicted. That is, the average amount of ranked labels to take into account to miss no true label.

$$coverage(y, \hat{f}) = \frac{1}{n_s} \sum_{i=0}^{n_s-1} \max_{j:y_{ij}=1} rank_{ij} \quad (2)$$

with n_l being the amount of labels, n_s being the amount of samples, $\hat{f} \in R^{n_s \times n_l}$ the score associated with each label, $y \in \{0,1\}^{n_s \times n_l}$ the ground truth labels, $rank_{ij} = \left\{k : \hat{f}_{ik} \geq \hat{f}_{ij}\right\}$.

6.1.3 Label Ranking Average Precision

Label Ranking Average Precision (LRAP) averages over the ground truth labels assigned to each sample, ranking true labels higher. This metric shows which ratio of higher-ranked labels were true labels.

$$LRAP(y, \hat{f}) = \frac{1}{n_s} \sum_{i=0}^{n_s-1} \frac{1}{||y_i||_0} \sum_{j:y_{ij}=1} \frac{|\mathcal{L}_{ij}|}{rank_{ij}} \quad (3)$$

with n_l being the amount of labels, n_s being the amount of samples, $\hat{f} \in R^{n_s \times n_l}$ the score associated with each label, $y \in \{0,1\}^{n_s \times n_l}$ the ground truth labels, $rank_{ij} = \left\{k : \hat{f}_{ik} \geq \hat{f}_{ij}\right\}$, $\mathcal{L}_{ij} = \left\{k : y_{ik} = 1, \hat{f}_{ik} \geq \hat{f}_{ij}\right\}$, $||\cdot||_0$ being the ℓ_0 norm (which computes the amount of nonzero elements in a vector), and $|\cdot|$ representing the cardinality of the set.

6.2 Results and Discussion

We compare our results against the model by MKaan, since it is the only available model that we have been able to find targeting the CPV code assignment problem in Spanish (besides other languages). Since no default threshold or function is provided, we tested different thresholds with the most common functions (softmax and sigmoid). Results are summarized in Table 2 (using only 10% of the dataset), and Table 3 (using the whole dataset).

The results clearly show that the RoBERTa fine-tuned model outperforms the rest of the approaches both when training using just a fraction of the dataset and the full dataset. The model by MKann shows a good performance taking into account its multilingual nature (not specific for the Spanish language). However, MKaan is matched and even outperformed by some of the traditional algorithms in both experiments.

In particular, classical approaches such as SVM, random forests and decision trees, produce remarkably good results (0.69, 0.64 and 0.63 F1 scores respectively on the full dataset). Given that these algorithms are usually less expensive to train, test and use than transformer-based solutions, they are reasonable candidates for assisting in CPV classification at a low cost. One possible explanation for this good performance is that, despite the presence of polysemous words that can be problematic, both the hyperplanes of SVM and the decisions of tree-based methods allow to effectively discriminate each label against all others (that is the strategy usually used to adapt the algorithms

Approach	ROC-AUC	F1	Accuracy	LRAP	Cov. Error
Multinomial NB	0.53	0.11	0.06	0.09	42.32
SVM	0.66	0.47	0.33	0.36	30.19
SVM (rbf)	0.66	0.47	0.33	0.36	30.19
KNN	0.70	0.54	0.41	0.45	26.54
Decision Tree	0.74	0.51	0.49	0.53	22.74
Random Forest	0.68	0.52	0.39	0.41	27.96
AdaBoost	0.75	0.56	0.41	0.49	22.10
RoBERTa fine-tuned (t=0.5)	0.84	**0.74**	**0.68**	**0.73**	14.13
RoBERTa fine-tuned (t=0.6)	0.83	0.73	0.67	0.71	14.86
RoBERTa fine-tuned (t=0.65)	0.82	0.73	0.67	0.70	15.41
RoBERTa fine-tuned (t=0.7)	0.81	0.72	0.64	0.68	16.54
MKaan (sigmoid, t=0.5)	0.80	0.13	0.0	0.07	17.38
MKaan (sigmoid, t=0.7)	0.85	0.19	0.0	0.11	13.31
MKaan (sigmoid, t=0.8)	0.86	0.24	0.0	0.15	12.21
MKaan (sigmoid, t=0.9)	0.87	0.32	0.01	0.23	11.49
MKaan (sigmoid, t=0.95)	0.87	0.42	0.06	0.34	11.64
MKaan (softmax, t=0.01)	**0.88**	0.37	0.25	0.44	**11.05**
MKaan (softmax, t=0.05)	0.86	0.55	0.43	0.59	12.48
MKaan (softmax, t=0.1)	0.85	0.61	0.51	0.64	13.64
MKaan (softmax, t=0.3)	0.81	0.65	0.61	0.66	16.63
MKaan (softmax, t=0.5)	0.79	0.65	0.60	0.63	18.71

Table 2: Results of the different approaches trained and tested on the 10% of the dataset (7243 training samples, 3104 test samples).

Approach	ROC-AUC	F1	Accuracy	LRAP	Cov. Error
Multinomial NB	0.56	0.22	0.14	0.16	39.07
SVM	0.78	0.69	0.58	0.62	18.89
SVM (rbf)	0.78	0.69	0.58	0.62	18.89
KNN	0.75	0.62	0.52	0.56	21.68
Decision Tree	0.80	0.63	0.60	0.64	17.68
Random Forest	0.74	0.64	0.51	0.54	22.32
AdaBoost	0.75	0.60	0.45	0.51	22.47
RoBERTa fine-tuned (t=0.5)	**0.89**	0.79	**0.74**	**0.80**	**10.32**
RoBERTa fine-tuned (t=0.6)	0.88	**0.80**	**0.74**	**0.80**	10.66
RoBERTa fine-tuned (t=0.65)	0.88	0.79	**0.74**	0.79	10.95
RoBERTa fine-tuned (t=0.7)	0.88	0.79	**0.74**	0.79	10.94
MKaan (sigmoid, t=0.5)	0.81	0.13	0.0	0.07	17.19
MKaan (sigmoid, t=0.7)	0.86	0.19	0.0	0.11	13.01
MKaan (sigmoid, t=0.8)	0.87	0.24	0.0	0.15	11.91
MKaan (sigmoid, t=0.9)	0.87	0.33	0.01	0.23	11.32
MKaan (sigmoid, t=0.95)	0.87	0.42	0.06	0.34	11.50
MKaan (softmax, t=0.01)	0.88	0.38	0.24	0.44	10.74
MKaan (softmax, t=0.05)	0.86	0.55	0.43	0.59	12.25
MKaan (softmax, t=0.1)	0.85	0.61	0.50	0.63	13.54
MKaan (softmax, t=0.3)	0.81	0.66	0.61	0.66	16.46
MKaan (softmax, t=0.5)	0.79	0.66	0.60	0.63	18.62

Table 3: Results of the different approaches trained and tested on the whole dataset (72429 training samples, 31042 test samples).

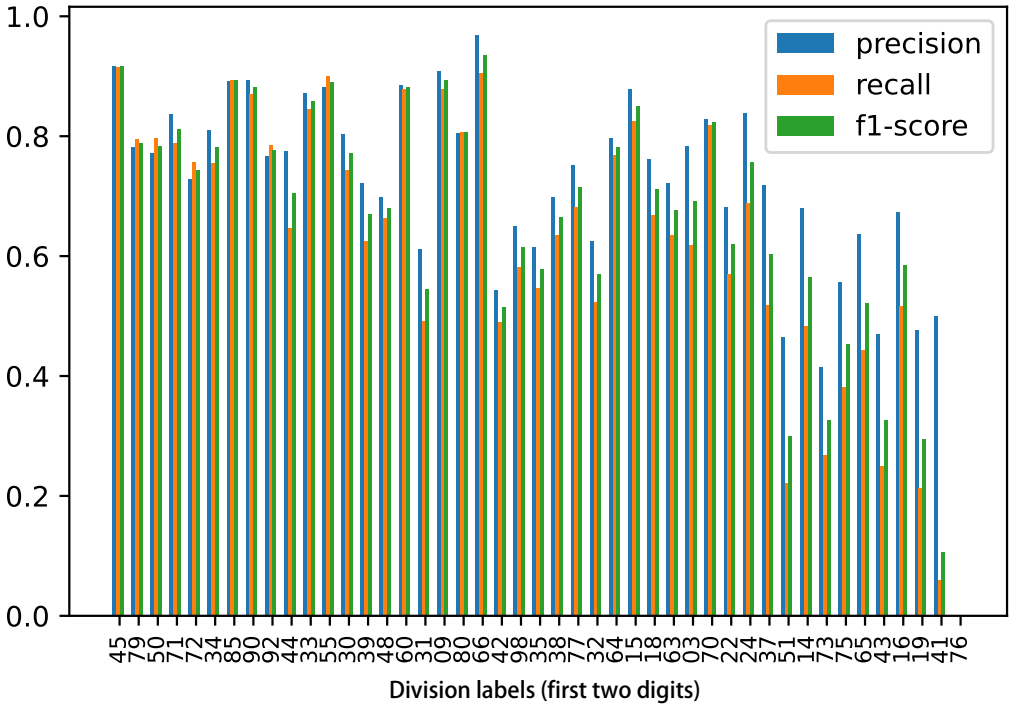


Figure 3: Results of the RoBERTa fine-tuned model (t=0.5) per label. We preserve the order presented in Figure 2, from more represented labels ('45') to less represented labels ('76').

to multiclass problems).

A limitation of our approach is the lack of measures for balancing input data. Typically, this would risk having our CPV classifier performing well only for the classes with more representation. However, as shown in Figure 3, our CPV classifier shows an excellent performance for most categories, and has an acceptable performance for classes with less data available (except fpr extremely rare categories '41' and '76'). We suspect that in addition to the number of training instances, the generality of the divisions and the overlap between them also play a role in the differences in performance. For example, divisions '42' and '43' represent "Industrial machinery" and "Machinery for mining, quarrying, construction equipment", respectively. Words similar to "machinery" will therefore appear frequently in descriptions of both divisions, leading to false positives/negatives. In Figure 3, we can in fact confirm that both divisions have worse performance than the immediate surrounding divisions having a similar amount of instances.

7 Conclusions and Future Work

This paper presents an approach to classify CPV code divisions for Spanish public procurement descriptions. Our work evaluated classical machine learning algorithms, showing that SVM had an excellent performance, surpassing the previous existing transformed-based approach for the task. Additionally, we fine-tuned the RoBERTa transformed-based model trained on a corpus of the BNE (Spanish National Library), that outperformed all the previous approaches. All data, data processing scripts and training notebooks have been made available through a public code repository, Zenodo (Navas-Loro, Garijo, and Corcho, 2022)[11] and a Research Object[12] for the sake of reproducibility. This material is also planned to be used in the AI4Gov international master.[13]

Our approach covers only CPV division classification, and therefore it does not yet address the CPV over-generalization problem when assigning CPVs to text (i.e., some codes

[11] https://zenodo.org/record/6554843
[12] https://w3id.org/dgarijo/ro/sepln2022
[13] https://ai4gov-master.eu/

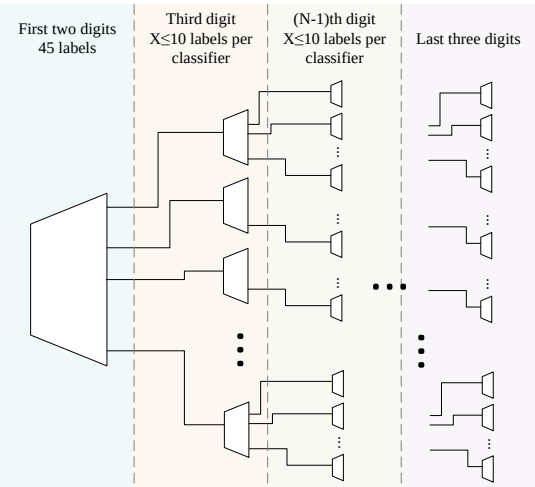


Figure 4: Hierarchical approach to the CPV classification problem. The first classifier would be responsible for categorizing the first two digits of the code, i.e., its division. The next level would attempt to predict the next digit based on the previous digits. For example, if the first classifier determined that a description corresponds to the labels '45' and '48', that description would be passed to the classifiers that determine the next digit trained with examples of those two codes.

are systematically not used in preference to more generic codes, even though the specific codes in disuse are much better suited to the topic of the description). Our future work includes designing a sequence of models that successively classify the digits of CPVs, as depicted in Figure 4, to be able to predict more specific CPVs. Alternatively, we plan on assessing techniques based on sentence embeddings against CPV descriptions, in order to suggest more specific CPVs despite the lack of training instances. Designing more specific classifiers will also require dealing with noise in data, e.g., when annotators assign different CPVs to the same contract description or incorrect CPVs. We also plan to increase the dataset, including contracting information from several years and also retrieving and making use of additional information from contracting processes. These include features such as the cost, that could help in the disambiguation of general words such as "service" or "work", that can be used in very different situations. Additionally, we will also enhance the preprocessing of the data in order to improve the quality in the dataset, a well-known problem in this kind of classification problem.

Overall, our positive results are a step forward towards the creation of a decision support system to help in CPV classification, allowing a more transparent and efficient public procurement in Spain and Europe.

Aknowledgments

This work has been supported by NextProcurement European Action (grant agreement INEA/CEF/ICT/A2020/2373713-Action 2020-ES-IA-0255) and the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with Universidad Politécnica de Madrid in the line Support for R&D projects for Beatriz Galindo researchers, in the context of the V PRICIT (Regional Programme of Research and Technological Innovation). We also acknowledge the participation of Jennifer Tabita for the preparation of the initial set of notebooks, and the AI4Gov master students from the first cohort for their validation of the approach. Source of the data: Ministerio de Hacienda.

References

Aggarwal, C. C. and C. Zhai. 2012. A survey of text classification algorithms. In *Mining text data*. Springer, pages 163–222.

Ahmia, O. 2020. *Assisted strategic monitoring on call for tender databases using natural language processing, text mining and deep learning*. Ph.D. thesis, Université de Bretagne Sud, 03.

Bhatia, K., H. Jain, P. Kar, M. Varma, and P. Jain. 2015. Sparse local embeddings for extreme multi-label classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Boser, B. E., I. M. Guyon, and V. N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1):5–32.

Chang, W.-C., H.-F. Yu, K. Zhong, Y. Yang, and I. S. Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3163–3171.

Deloitte. 2020. Study on up-take of emerging technologies in public procurement. Technical report, Deloitte.

European Commission. 2020. *eForms : policy implementation handbook*. Publications Office.

Freund, Y. and R. E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Gargiulo, F., S. Silvestri, M. Ciampi, and G. De Pietro. 2019. Deep neural network for hierarchical extreme multi-label text classification. *Applied Soft Computing*, 79:125–138.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. R. Penagos, and M. Villegas. 2021. Spanish language models. *CoRR*, abs/2107.07253.

Hand, D. J. 2007. Principles of data mining. *Drug safety*, 30(7):621–622.

Hastie, T., S. Rosset, J. Zhu, and H. Zou. 2009. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360.

İşgüder-Şahin, G. G., H. R. Zafer, and E. Adah. 2014. Polarity detection of turkish comments on technology companies. In *2014 International Conference on Asian Language Processing (IALP)*, pages 136–139. IEEE.

Kayte, S. and P. Schneider-Kamp. 2019. A mixed neural network and support vector machine model for tender creation in the european union ted database. In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 139–145. INSTICC, SciTePress.

Liu, J., W.-C. Chang, Y. Wu, and Y. Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 115–124.

Minaee, S., N. Kalchbrenner, E. Cambria, et al. 2021. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3), apr.

Minsky, M. 1961. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30.

Navas-Loro, M., D. Garijo, and O. Corcho. 2022. Code repository for multi-label text classification for public procurement in spanish, May.

Prabhu, Y. and M. Varma. 2014. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 263–272.

Quinlan, J. R. 1986. Induction of decision trees. *Machine learning*, 1(1):81–106.

Siblini, W., P. Kuntz, and F. Meyer. 2018. CRAFTML, an efficient clustering-based random forest for extreme multi-label learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4664–4673. PMLR.

Soylu, A., Corcho, B. Elvesæter, C. Badenes-Olmedo, F. Yedro-Martínez, et al. 2022. Data quality barriers for transparency in public procurement. *Information*, 13(2).

Suta, A. 2019. Multilabel text classification of public procurements using deep learning intent detection. Master's thesis, KTH, Mathematical Statistics.

Zhang, M.-L. and Z.-H. Zhou. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, 40(7):2038–2048.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 83-94 recibido 31-03-2022 revisado 20-05-2022 aceptado 23-05-2022

Selección de colocaciones académicas en español a través de un filtro de interdisciplinariedad

Selecting Spanish academic collocations using a filter of interdisciplinarity

Eleonora Guzzi, Margarita Alonso Ramos
Universidade da Coruña, CITIC
eleonora.guzzi@udc.es, margarita.alonso@udc.gal

Resumen: En este artículo se propone una metodología para compilar una lista de colocaciones académicas con base nominal que se integran en una herramienta léxica (Alonso-Ramos, García-Salido y Garcia, 2017). Para ello, establecemos un filtro que mide la interdisciplinariedad de los nombres académicos a partir de los cuales se extraen las colocaciones (García-Salido, 2021), con el fin de mantener los nombres frecuentes y bien distribuidos en distintas disciplinas académicas, y descartar aquellos que se adscriben a la terminología o que son más característicos de la lengua general. Utilizamos tres criterios: (1) el IDF (Jones, 1972); (2) el análisis de la distribución de colocaciones; (3) el contraste con listas de vocabulario académico inglés. Los resultados muestran que estos criterios son útiles para identificar los nombres prototípicos del discurso académico y permiten filtrar la lista de colocaciones académicas. No obstante, persiste el problema de cómo tratar la desambiguación semántica en relación con las diferentes disciplinas.

Palabras clave: discurso académico, interdisciplinariedad, colocaciones académicas.

Abstract: In this paper a methodology to compile a list of noun-based academic collocations that feed a lexical tool (Author, 2017) is proposed. To do so, a filter that measures the interdisciplinarity of academic nouns from which collocations are extracted (García-Salido, 2021) is established. This filter is applied to include nouns that are frequent and homogeneously distributed across different academic disciplines, and discard those ascribed to terminology or are more prototypical of general language. Three criteria were used: (1) the IDF (Jones, 1972); (2) an analysis of collocation distributions; (3) a contrast with vocabulary lists of academic English. Results show that these criteria are useful for identifying prototypical nouns of academic discourse and allow for filtering the list of academic collocations. However, the problem regarding how to deal with semantic disambiguation in different disciplines is still present.

Keywords: academic discourse, interdisciplinarity, academic collocations.

1 Introducción

Uno de los principales objetivos dentro del ámbito de las lenguas con fines académicos ha sido el de proporcionar listas de vocabulario académico para ser utilizadas como recursos pedagógicos e integradas en la enseñanza de la escritura académica. Este vocabulario incluye unidades y combinaciones léxicas que son específicas del género académico, pero no son terminológicas. A su vez, se caracterizan por ser más frecuentes en el discurso académico que en la lengua general o en otros géneros. Siguiendo a Tutin (2007a) podemos definir el vocabulario académico como aquel vocabulario que hace referencia a los procedimientos y actividades científicas del discurso científico, esenciales en la argumentación y en la estructuración de los textos académicos (Drouin, 2007; Paquot y Bestgen, 2009).

Hasta el momento, se han propuesto varias listas de unidades léxicas académicas especialmente para el inglés y el francés: *Academic Vocabulary Lis*t (AVL, Gardner y Davies, 2013), *Academic Word List* (AWL, Coxhead, 2000), *Academic Keyword List* (AKL, Paquot, 2007, *French Cross-disciplinary Scientific Lexicon* (Hatier et al., 2016), *Lexique*

ISSN 1135-5948. DOI 10.26342/2022-69-7

Scientifique Transdisciplinaire (LST, Drouin, 200)7, entre otras. Por otro lado, dentro de las listas de combinaciones léxicas destacan la *Academic Collocation List* (ACL, Ackermann y Chen, 2013) y la *Academic English Collocation List* (Lei y Liu, 2018), centradas específicamente en las colocaciones, así como la lista de palabras y colocaciones académicas empleadas para la herramienta lexicográfica *Collocaid* (Frankenberg-Garcia et al., 2019). En el ámbito del español, se ha propuesto recientemente una lista de unidades léxicas académicas, la *Spanish Academic Key Word List* (SpAKWL, García-Salido 2021), que incluye 1.239 lemas de nombres, adjetivos, verbos y adverbios. A partir de los nombres de esta lista, también se ha extraído una primera versión de colocaciones académicas con base nominal que se integran en la *Herramienta de Ayuda a la Redacción de Textos Académicos* (HARTA; Alonso-Ramos, García-Salido y Garcia, 2017; disponible en: http://www.dicesp.com:8083/).

En estos estudios se han aplicado varios criterios estadísticos para identificar automáticamente el vocabulario interdisciplinar en el discurso académico. Sin embargo, los criterios pueden diferir entre las listas de diferentes lenguas por los rasgos léxicos de las mismas. Por ejemplo, como apuntan Cobb y Horst (2004), existe un continuum entre el vocabulario académico y no académico en lenguas como el francés o el español debido a que, entre otras razones, el vocabulario greco-latino que es característico de los textos académicos, también se integra en dominios no académicos. Sin embargo, esto no sucede en inglés, donde los términos greco-latinos tienen mucha más presencia en el discurso académico. En el caso del español, además, aunque una palabra se utilice en la lengua general, se podría incluir en las listas académicas si también es frecuente en el discurso académico, debido a que, por un lado, los sentidos de las palabras académicas pueden diferir de los utilizados en los textos que no son académicos (Gilquin, Granger y Paquot, 2007), como, por ejemplo, *trabajo* (‘composición científica o literaria’ frente a *trabajo* como ‘empleo’, en la lengua general) y, por otro lado, porque dichas palabras pueden formar parte de combinaciones léxicas que son más exclusivas de los textos académicos (García-Salido, 2021), como la colocación *dato cuantitativo* con la palabra *dato,* frente a *dato personal.* El objetivo final a la hora de compilar las listas de vocabulario académico debería ser crear un balance entre la especificidad del vocabulario y su productividad. Esto es, se debería apuntar a la inclusión de unidades y combinaciones léxicas que son productivas para la redacción de textos académicos y que pueden estar en la intersección, por ejemplo, entre la lengua general y la lengua académica, pero que excluyen lo que es propio de otros géneros.

Otras dificultades a la hora de recoger este vocabulario argumentadas en Hyland y Tse (2007) son la falta de atención a la posible polisemia y a la variedad disciplinar. La polisemia implica que distintos sentidos sean utilizados en distintas disciplinas académicas: por ejemplo, la palabra *volumen* puede significar ‘capacidad’ en Física, ‘ejemplar de libro’ en Biblioteconomía o ‘frecuencia acústica’ en Ingeniería. Por otra parte, la variedad disciplinar se asocia a la posibilidad de que determinadas unidades y combinaciones léxicas pueden ser académicas, pero pueden utilizarse con más frecuencia en algunas disciplinas.

En definitiva, no existe un consenso común que apunte hacia una generalización de criterios sobre cómo determinar la interdisciplinariedad y obtener vocabulario característico de textos académicos. Como consecuencia, este trabajo se propone con un doble objetivo: por un lado, ofrecer una lista más refinada de nombres académicos en español y, por otro lado, utilizar la lista para filtrar las colocaciones académicas con base nominal que se integran en HARTA. Para alcanzar estos objetivos, seguimos una metodología basada en el refinamiento de los nombres de la *SpAKWL*, que implica el descarte de aquellos adscritos a una disciplina específica o que son significativamente más recurrentes en la lengua general.

A continuación, en la sección 2 presentamos el proceso de extracción de la *SpAKWL* y de las colocaciones académicas en español; en la

sección 3 presentamos la metodología empleada para identificar los nombres de la *SpAKWL* que son interdisciplinares y filtrar las colocaciones que contienen dichos nombres; en la sección 4 exponemos los resultados obtenidos a partir de los distintos análisis; y, en la sección 5, discutimos los resultados, seguidos de las conclusiones finales.

2 Descripción de los datos: SpAKWL y colocaciones académicas

Este estudio parte de la lista de palabras académicas del español *SpAKWL* (García-Salido, 2021), extraída siguiendo criterios de especificidad y de distribución a partir de un corpus académico, HARTA-Expertos (HE, Alonso-Ramos, García-Salido y Garcia, 2017). HE contiene 413 artículos científicos publicados, cuya mayoría proviene de la sección en español del corpus *Spanish-English Research Articles Corpus* (SERAC; Pérez-Llantada, 2014), y suma un total de 2.025.092 palabras. Está divido en 4 dominios principales (Artes y Humanidades, Biología y Medicina, Ciencias Sociales y Ciencias Físicas e Ingeniería) y 12 subdominios, siguiendo la estructura del SERAC. El proceso de tokenización y lematización se llevó a cabo mediante LinguaKit (Garcia y Gamallo, 2016) y el de etiquetación con FreeLing (Padró y Stanilovsky, 2012). Para analizar sintácticamente el corpus con dependencias universales (Nivre et al., 2016) se utilizó UDPipe (Straka, Hajic y Straková, 2016).

Para determinar la especificidad e identificar las palabras específicas del ámbito académico frente a un corpus de referencia, se empleó el test estadístico log-likelihood a partir de las frecuencias absolutas de HE y del corpus de referencia, en este caso, la sección de ficción del corpus de narrativa LEXESP (Sebastián-Gallés et al., 2000), de 5 millones de palabras. Como criterio de distribución, se seleccionaron aquellas palabras con ocurrencias en los 4 dominios y el 10% de palabras con una distribución más homogénea en términos de DP (Deviation of Proportions, Gries, 2008). La lista de palabras académicas resultante cuenta con 1.239 lemas que se corresponden con nombres, verbos, adverbios y adjetivos.

A partir de esta lista, se seleccionaron los nombres (n=602) que se emplearon como bases para extraer automáticamente una primera versión de colocaciones académicas, que están integradas en HARTA. Para este propósito, definimos las colocaciones, dentro del marco de la Lexicografía Explicativa y Combinatoria (Mel'čuk, 2012), como combinaciones léxicas con un significado composicional, que están formadas por una 'base', en este caso, un nombre, y un 'colocativo', y cuyos elementos tienden a coocurrir, como, por ejemplo, *alcanzar un objetivo*.

Para analizar la interdisciplinariedad de las bases y de las colocaciones académicas en español, partimos de las colocaciones ya integradas en HARTA y de un segundo grupo más numeroso de colocaciones, que fue extraído a partir de una ampliación del corpus HE, HARTA-Expertos-Plus (HEP). Este corpus contiene 21.068.482 palabras procedentes de 3.870 artículos de investigación: 19.043.390 palabras proceden del corpus académico-científico *Iberia* (Ahumada et al., 2011) y 2.025.092 palabras provienen del corpus HE. El corpus HEP se divide en los mismos cuatro dominios principales que HE, a su vez divididos en subdominios. Para los nuevos artículos, se aplicó el mismo proceso de tokenización, lematización y análisis sintáctico que se siguió para HE. Tras replicar los pasos seguidos para obtener la primera versión de colocaciones académicas, en primer lugar, se extrajeron colocaciones de 5 relaciones de dependencias sintácticas (N + N, V + Obj, Suj + V, N + Adj, N + Obl). En segundo lugar, se realizó una extracción automática, basada en medidas de asociación estadísticas (log-likelihood, Información Mutua, entre otras), y en criterios de frecuencia (≥5 ocurrencias) (Alonso-Ramos, García-Salido y Garcia, 2017). En tercer lugar, un grupo de anotadores refinó manualmente los candidatos extraídos automáticamente para obtener las colocaciones académicas, siguiendo criterios fraseológicos que se enmarcan dentro de la Teoría Sentido-Texto (Mel'čuk, 2012). A

pesar del refinamiento manual y de las medidas escogidas, tanto en la lista de nombres académicos como en las colocaciones seleccionadas se incluyen casos más asociados a la terminología, como, por ejemplo, la palabra *tejido* usada con el sentido de Biología 'cada uno de los agregados de células de la misma naturaleza' (DLE, s.m, def. 4), o la colocación *ingresar paciente,* que no son productivas para los varios dominios académicos.

3 Metodología

Con el fin de descartar los nombres especializados y los que son más frecuentes en la lengua general o en otros géneros, en primer lugar, aplicamos la medida IDF (Inverse Document Frequency, Jones 1972) a los 602 nombres de la lista *SpAKWL*. Esta medida se basa en el cálculo de la proporción de documentos que contiene un determinado término y se utiliza frecuentemente en el campo de la Extracción de la Información para identificar palabras clave, esto es, palabras específicas de un conjunto de textos que tienen un alto valor de IDF. Tras calcular el IDF de los nombres, ordenamos los valores de mayor a menor para visualizar en la posición más alta los candidatos más terminológicos. Calculamos la media de IDF (=0,659) y llevamos a cabo una revisión exhaustiva de los nombres situados por encima y por debajo de la misma para seleccionar el punto de corte, en este caso, la palabra *potencia* (IDF=0,900). En este estudio, el IDF nos ayudó a determinar precisamente las palabras que pueden ser descartadas (n=119) por ser menos interdisciplinares y más características de algunos artículos, que se corresponden con las ubicadas por encima del punto de corte (ver Anexo 1).

A partir del resultado obtenido, seguimos un proceso de filtrado dividido en diferentes fases para analizar con más profundidad los 119 nombres que se descartarían por el IDF. En la primera fase (F1), un grupo de anotadores clasificó los nombres con IDF ≥0,900 en dos grupos: los nombres que, tal y como indica esta medida, se descartan por ser especializados o porque pertenecen a la lengua general (G1) y los nombres que deben seguir un proceso de revisión posterior (G2). En la segunda fase (F2), en la que se incluyen los nombres clasificados en el G2, identificamos las colocaciones extraídas que contienen dichos nombres como bases para analizar su distribución y el número de colocativos con los que se combina. En este proceso, algunas bases se descartan, otras bases se reincluyen en la lista inicial y otro grupo de bases se selecciona para revisar en la siguiente fase. Por último, en la tercera fase (F3), contrastamos los equivalentes de los nombres en cuatro listas de inglés académico (*AVL, AWL, AKL*, y la lista de palabras académicas de *Collocaid*). En la Figura 1, exponemos el proceso seguido:

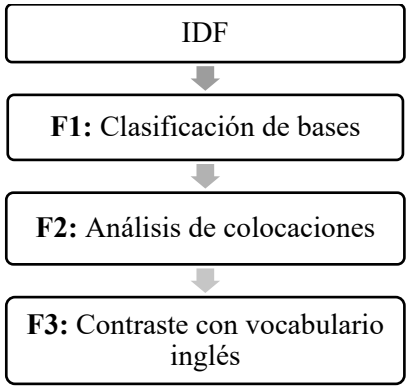


Figura 1. Proceso para filtrar los nombres de la *SpAKWL*

En las siguientes secciones, se explica en detalle el proceso que llevamos a cabo en las tres fases.

3.1. Clasificación de las bases a partir del resultado del IDF

Una vez identificados los 119 nombres que están por encima del punto de corte del IDF, un grupo de anotadores conformado por tres lingüistas los clasificó en dos grupos. En el G1 se incluyeron los nombres que, tras observar los ejemplos en contexto en el corpus HE y con la ayuda de los diccionarios para analizar los sentidos y las marcas de especialidad, resultaron ser terminológicos o más asociados a un número reducido de disciplinas. También se incluyeron aquellos nombres que presentan una frecuencia elevada en los corpus de lengua general, utilizando herramientas de corpus para consultar su frecuencia, como el corpus de *esTenTen* en

Sketch Engine (Kilgariff y Renau, 2013). En el G2 se incluyeron los nombres que, en cambio, requieren un análisis posterior por no mostrar indicios claros sobre su interdisciplinariedad. En función del descarte o mantenimiento, se les asignaron puntuaciones a los nombres. Por ejemplo, encontramos casos como el de la palabra *fracción,* que se descarta (=0), e *indicación,* que pasa a una siguiente fase (=análisis) (Tabla 1):

	Fase 1
fracción	0
indicación	análisis

Tabla 1. Ejemplo de puntuación asignada a nombres a descartar o analizar en la F1.

3.2.Análisis de colocaciones

En esta fase analizamos las bases clasificadas en el G2 (n=54), con el fin de llevar a cabo una valoración acerca del número de colocativos, que se asocia a la riqueza del vocabulario, así como de la distribución de las colocaciones que conforman, relacionada con la interdisciplinariedad. En este sentido, una base que se combina con varios colocativos y que forma colocaciones que están bien distribuidas en los textos académicos debería incluirse en la lista de nombres académicos.

Para analizar la distribución de las colocaciones, se representaron las frecuencias de las colocaciones en cada dominio (AH, CS, CF, BM) en forma de porcentajes, se calculó la desviación estándar (DE), y se indicó el número de subdominios en los que aparece cada colocación. Los valores de la DE oscilaron entre 0,00 y 0,50: cuanto más bajo es el valor, más homogénea es la distribución de la colocación en los textos de los cuatros dominios. En cuanto a los criterios de análisis de las colocaciones que contienen los nombres académicos, consideramos los tres parámetros: una colocación es interdisciplinar si presenta una DE entre 0,00 y 0,24, si se aproxima a un porcentaje de ≥20% en al menos tres dominios o bien en dos dominios, uno perteneciente a "ciencias duras" (CF y BM) y otro a "ciencias blandas" (CS y AH) y si aparece en ≥3 subdominios. Por lo tanto, si las colocaciones analizadas con un nombre del G2 están bien distribuidas, como, por ejemplo, la colocación *alcanzar difusión* (Tabla 2), el nombre, en este caso *difusión,* recibe una puntuación *=1* y se mantiene:

	DE	**Dom.**	**Sub.**	**=**
alcanzar difusión	0,27	AH 50%	5	Distri. alta
		CS 12,5%		
		CF 12,5%		
		BM 25%		

Tabla 2. Análisis de una colocación con distribución alta (homogénea).

Por el contrario, si las colocaciones que se forman a partir de un nombre presentan una DE entre 0,36-0,50, aproximadamente, una frecuencia de 0% en tres dominios o ≥90% en un dominio, y aparece solamente en uno o dos subdominios (Tabla 3), la base, en este caso, *abundancia*, con una puntuación *=0*, se descarta:

	DE	**Dom.**	**Sub.**	**=**
abundancia mayor	0,47	BM 97%	2	Distri. baja
		CF 3%		

Tabla 3. Análisis de una colocación con distribución baja (heterogénea).

Los nombres que pasan a la siguiente fase de revisión (F3) en la que se contrastan con las posibles equivalencias en las listas de inglés, y que reciben una puntuación *=análisis*, son aquellas bases que en este proceso presentan colocaciones que oscilan entre los límites de una distribución homogénea y heterogénea. Esto es, en este grupo se incluyen las bases cuyas colocaciones presentan una DE entre 0,25 y 0,35, aproximadamente, aparecen en dos dominios de un único grupo ("ciencias "duras" / "ciencias blandas"), pero con porcentajes equilibrados, o en tres subdominios de forma desequilibrada (Tabla 4):

	DE	**Dom.**	**Sub.**	**=**
indicación precisa	0,33	CS 18%	3	Distri. media
		AH 9%		
		BM 73%		

Tabla 4. Análisis de una colocación que requiere un análisis posterior.

Asimismo, pasan a la fase de revisión 3 un número reducido de bases que no presentan ningún colocativo productivo a nivel fraseológico de entre todos los candidatos extraídos, como, por ejemplo, el nombre *tipología*. Ahí decidimos su mantenimiento o descarte para formar parte de una nueva lista de nombres académicos.

En la siguiente Tabla (5) mostramos las puntuaciones de tres bases que, tras el análisis colocacional, reciben puntuaciones distintas:

	Fase 1	**Fase 2**
abundancia	1	0
indicación	1	análisis
difusión	1	1

Tabla 5. Ejemplos de puntuaciones asignadas a bases a descartar, analizar o incluir en la F2.

3.3. Contraste con listas de vocabulario de inglés académico

En esta última fase se analizaron las bases que presentaron dudas tras pasar por el análisis colocacional y aquellos nombres que no han podido analizarse en la fase 2, debido a que no presentaron ningún colocativo.

Se seleccionaron cuatro listas de palabras académicas en inglés para el análisis: *AKL, AVL, AWL* y la lista de palabras académicas de *Collocaid*. A partir de los nombres seleccionados en la fase anterior, se realizó una comparativa con las palabras académicas pertenecientes a las cuatro listas para encontrar su equivalente en inglés. En el proceso de búsqueda de los equivalentes, se consultaron dos diccionarios, el *Oxford English Dictionary* y el *Cambridge Dictionary*, tanto la versión bilingüe español-inglés como la monolingüe, así como el corpus paralelo *Linguee* para observar ejemplos en contexto. A la hora de buscar la equivalencia de cada nombre, se consideraron las diferentes traducciones posibles debido a la presencia de polisemia. Se considera que una palabra coincide con dos o más listas si en cada lista se presenta la traducción asociada a un mismo sentido, por ejemplo, la palabra *cultura* presenta su correspondencia en las cuatro listas con el nombre en inglés *culture.* Sin embargo, se dan otros casos en los que una palabra se traduce de distintas formas dependiendo del sentido que se adopte en cada una de ellas: por ejemplo, la palabra *señal* se traduce en la *AVL* con el sentido asociado a *signal,* mientras que en la lista de *Collocaid* y en la *AKL* solamente se encuentra el equivalente de otro sentido, *indication.* En estos casos, se considera el número de listas coincidentes para cada sentido: *señal* puede coincidir con una lista (la *AVL*) o dos listas (*Collocaid* y *AKL*), en función del sentido.

En relación con el criterio para filtrar la lista de nombres, se estableció un índice de ≥2, es decir, si un nombre aparece en al menos dos de las cuatro listas se mantiene. Por el contrario, las palabras que coinciden únicamente con una lista o que no coinciden con ninguna finalmente se descartan. Fijamos este índice debido a que la *AWL* aplica criterios más restringidos y no incluye palabras que también pertenecen a la lengua general, lo que provoca un porcentaje de correspondencia bajo porque en el vocabulario académico español se recogen palabras compartidas con la lengua general.

En la Tabla 6, podemos observar los ejemplos con las respectivas puntuaciones de nombres que se analizan en la fase 3, que incluyen tanto nombres que presentaron colocativos en la F2 (ej. *indicación* o *especificación*), como aquellos que no presentaron ningún colocativo (ej. *tipología* o *almacenamiento*):

	Fase1	**Fase2**	**Fase3**	**=**
almacenamiento	1	análisis (no coloc.)	0	1
tipología	1	análisis (no coloc.)	1	2
especificación	1	análisis	0	1
indicación	1	análisis	1	2

Tabla 6. Ejemplos de puntuaciones asignadas a nombres a descartar o incluir en la F3.

4 Resultados

La lista resultante se compone de 519 nombres académicos (Anexo 2), en contraste con los 602 nombres iniciales. La clasificación de los nombres en cada fase ha sido el resultado de las puntuaciones asignadas a cada uno de ellos. Una

puntuación final de 2 implicó la reinclusión del nombre a la lista inicial y una puntuación de 0 o 1 conllevó su descarte.

En la primera fase (F1), se descartaron 65 nombres, con una puntuación de 0, entre los cuales encontramos ejemplos como *emisión, fracción, tejido, geometría, prevención, infraestructura,* etc. Por otra parte, se mantuvieron 54 nombres que pasaron a una siguiente revisión, con una puntuación de 1, como *énfasis, concordancia, fiabilidad, puntuación, descenso, almacenamiento, procesamiento,* entre otros.

En cuanto a la fase de revisión de las colocaciones (F2), 6 bases se descartaron, como *especificación, asignación* o *resto*; 29 bases pasaron a la fase 3, como *trayectoria, gráfico, indicación* o *premisa*; y 19 bases se reincluyeron en la lista inicial por cumplir con el criterio de número y distribución de las colocaciones, como *sesgo, productividad, predominio* o *estándar*.

Por último, en la fase 3, de los nombres que se contrastaron con las listas de palabras académicas en inglés, se descartaron l2 nombres, entre los cuales encontramos ejemplos como *corrección, concordancia* y *fiabilidad*, ya que sus posibles equivalentes en inglés no aparecían en ninguna lista de inglés y *almacenamiento*, *formulación*, *asignación, acumulación*, *regulación*, *procesamiento, trayectoria, afirmación* y *barrera* porque aparecían únicamente en una lista. Por ejemplo, de la palabra *regulación* únicamente encontramos un posible equivalente en la *AVL* como *adjustment.*

Sin embargo, se reincluyeron 17 nombres: *gráfico, indicación, reproducción* y *experto,* que aparecían en dos listas; *heterogeneidad, premisa, variante, diferenciación, bibliografía, tipología, desempeño, instancia*, *supuesto* y *unión,* que aparecían en tres listas, y *paradigma, vínculo* y *rol,* cuyos equivalentes aparecieron en las cuatro listas. Por ejemplo, la palabra *instancia* aparece como *instance* en la *AVL*, en la *AKL* y en la lista de *Collocaid*.

A modo de resumen, en la Tabla 7, mostramos el número de nombres descartados, analizados o reincluidos en cada fase:

	F1	F2	F3	TOTAL
Descarte	65	6	12	83
Análisis	**54**	**29**	-	83
Reinclusión	-	19	17	36
TOTAL		**54**	**29**	119

Tabla 7. Nº de nombres clasificados en cada fase.

5 Discusión

La medida IDF ha permitido detectar las palabras clave de determinados documentos del corpus y, en consecuencia, aquellos nombres empleados más específicamente en algunas de las áreas científicas en las que está dividido el corpus HE. Cabe destacar que la decisión del punto de corte en la palabra *potencia* (IDF =0,900) ha implicado que un número reducido de nombres, como *software* o *regresión*, no se descartaran a pesar de presentar un valor alto de IDF (cerca de 0,800) y de ser especializados. Sin embargo, hemos optado por no establecer un punto de corte más bajo debido a que se habría descartado una gran parte de nombres académicos relevantes, como *disciplina, discurso, síntesis,* entre otros.

De los dos conjuntos de nombres obtenidos con el IDF, contrastamos la dispersión de algunas palabras con un alto valor de IDF con algunas palabras con valores más bajos en los artículos de cada subdominio. En efecto, observamos que la frecuencia y distribución no es proporcionada en el corpus en el primer caso (IDF alto), pero sí en el segundo (IDF bajo). En la Figura 1, podemos observar este comportamiento en una muestra de seis nombres con valores altos de IDF (IDF>1,06), es decir, menos distribuidos en los textos, y en la Figura 2, seis nombres con valores muy bajos (IDF<0.05). Para facilitar su lectura, presentamos la distribución de los nombres por subdominio en lugar de por artículos:

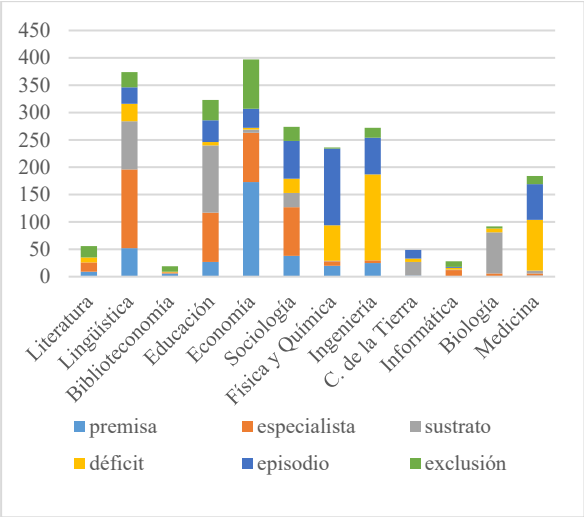


Figura 2. Seis nombres con IDF >1,05.

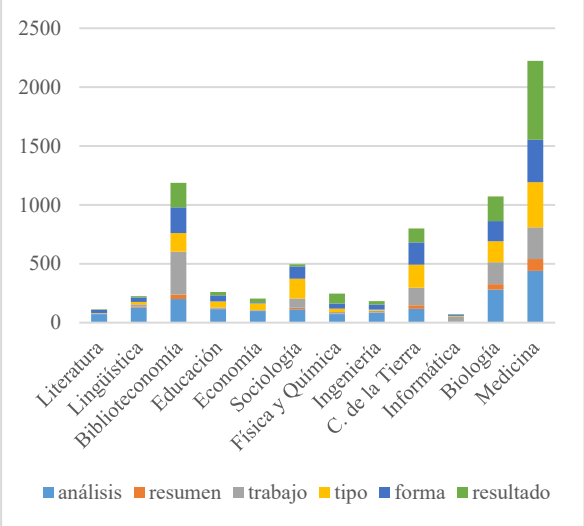


Figura 3. Seis nombres con IDF <0,05.

Como se puede apreciar, los nombres en la Figura 3 presentan una dispersión más homogénea que en la Figura 2. Por ejemplo, el nombre *episodio* presenta un gran predominio en los subdominios de Física y Química y Medicina, y no aparece en 3 de los 12 subdominios, mientras que el nombre *análisis* está bien distribuido, con ocurrencias proporcionales en los 12 subdominios.

En la primera fase de clasificación de los posibles nombres que se descartarían por el IDF, la mayoría se pudo clasificar en función de su especificidad en determinadas disciplinas, gracias al análisis de los contextos en los que aparecen dichas palabras en el corpus, así como a la ayuda de los diccionarios, como, por ejemplo, los nombres *geometría* o *motor*. Sin embargo, en línea con la problemática expuesta en la sección 1, se llevó a cabo una revisión más exhaustiva especialmente de aquellos nombres que se encuentran en un continuum entre la lengua general y la lengua académica, p.ej. *rol*, así como de los nombres polisémicos como *barrera,* que presenta un sentido más metafórico de 'obstáculo', y otro sentido de 'valla […] u otro obstáculo semejante con que cierra el paso' (DLE, f.s., def. 1), con el fin de identificar los sentidos más productivos en el discurso académico.

En relación con el análisis de las colocaciones (F2), un número muy reducido de nombres académicos presentaron únicamente un colocativo. En estos casos, si la colocación no presentó un nivel medio-alto de distribución, la base se eliminó. Por ejemplo, con el nombre *especificación*, únicamente se identificó el colocativo *cumplir*, y la colocación presentó una distribución heterogénea, con un 93% de ocurrencias en Ciencias Físicas (ejemplo 1):

(1) "Se diseñó la estructura de pavimento con agregados de La Calera, por *cumplir con* todas las *especificaciones*".

Por otra parte, la gran mayoría de nombres académicos presentaron ≥2 colocativos, por lo que fue necesario un análisis más detallado de la distribución de cada colocación para definir una media. Por ejemplo, con la base *productividad,* se identificaron los colocativos *aumento, alta* y *mayor,* que tienen una distribución medio-alta, como *productividad alta,* que presenta un 40% de ocurrencias en Ciencias Físicas, un 40% en Biología y Medicina y un 20% en Artes y Humanidades, una DE de 0,19, y aparece en 3 subdominios. Los casos que conllevaron más dudas se corresponden con aquellas bases que presentan 2-3 colocativos y una distribución media de colocaciones. Por ejemplo, con la base *indicación* se identificaron los colocativos *clara* y *precisa*: la colocación *indicación precisa* presenta una distribución homogénea, con una DE de 0,32, un porcentaje de aparición de un 9% en Artes y Humanidades, un 19% en Ciencias Sociales y un 72% en Biología y Medicina, y una aparición en 4 subdominios; sin embargo, la colocación *indicación clara* únicamente aparece

en los dominios de "ciencias blandas", con un 22% de ocurrencias en Artes y Humanidades y un 78% en Ciencias Sociales, una DE de 0,37 y una aparición en 3 subdominios. Debido a que únicamente se presentan dos colocaciones y la media de su distribución no proporciona indicios definitivos sobre su inclusión o descarte como base académica, en estos casos se contrasta el nombre con las listas de vocabulario académico en inglés. Cabe destacar que, en esta fase, también se identificaron 16 nombres que no presentaron ningún colocativo y, por lo tanto, no pudieron analizarse y pasaron a la fase 3. Como hemos mencionado en la sección 1, a pesar de que el objetivo principal del presente trabajo sea filtrar la lista de colocaciones académicas, los nombres sin colocativos se incluyen en este análisis con el propósito de obtener también una lista completa y más refinada de nombres académicos del español.

El análisis de las colocaciones también ha ofrecido indicios sobre el contraste de uso de las colocaciones en la lengua general y la lengua académica: la colocación *jugar un rol* no presenta una frecuencia alta ni una buena distribución en el discurso académico, pues su uso es más extendido en la lengua general, mientras que *desempeñar* y *ejercer un rol* presentan una distribución más homogénea entre los dominios y una frecuencia ligeramente mayor en el discurso académico.

A su vez, hemos podido observar casos en los que un nombre puede combinarse con colocativos y conformar colocaciones que están distribuidas de forma más homogénea que con otros colocativos: por ejemplo, con la base *puntuación,* identificamos la colocación *otorgar una puntuación,* que presenta una buena distribución, con un 14% de apariciones en AH, un 57% en CS, y un 39% en CF, y con una DE de 0,24; contrariamente, la colocación *puntuación mínima*, presenta un distribución heterogénea, con un 14% de ocurrencias en CS y un 86% en CF, una DE de 0,41 y con presencia únicamente en 2 subdominios. Estos casos indican que la base debe ser incluida en la lista, ya que es un nombre utilizado frecuentemente en distintos textos académicos, pero constituye colocaciones que se utilizan con más frecuencia en algunas disciplinas que en otras, probablemente debido a la variedad disciplinar.

En definitiva, hemos observado que los nombres que han tenido que pasar por distintas fases de análisis se corresponden especialmente con los nombres polisémicos, que poseen al menos dos sentidos distintos (ej. *barrera*) y los nombres que también se utilizan frecuentemente en la lengua general y, por lo tanto, no evidencian su especificidad en el discurso académico, como los nombres *unión* o *rol*.

6 Conclusiones

En este artículo se ha presentado una metodología para proponer una lista de nombres del discurso académico en español a partir de la lista *SpAKWL* (García-Salido 2021), aplicando criterios que identifiquen mejor la interdisciplinariedad. Aunque el objetivo principal es obtener una lista de colocaciones académicas que se integrará en HARTA, como objetivo secundario, hemos obtenido una lista de nombres académicos más prototípicos del discurso académico. Partiendo de la medida de IDF que es comúnmente utilizada para identificar de forma automática los nombres más asociados a la terminología, hemos aplicado diferentes análisis para valorar su efectividad y corroborar que los nombres identificados pueden descartarse.

Los resultados han demostrado que con esta metodología es posible identificar la interdisciplinariedad y establecer la lista de nombres académicos junto con una lista de colocaciones que puedan ser integradas en una herramienta que ayude a redactar textos académicos (HARTA). Específicamente, se ha obtenido un criterio para eliminar la terminología e identificar el vocabulario que puede ser utilizado independientemente de la disciplina y que ayuda a describir actividades y procesos académico-científicos y a estructurar la argumentación. No obstante, los resultados siguen remarcando la necesidad de desambiguar los sentidos de las palabras y la posibilidad de que, aunque las bases sean interdisciplinares en

el discurso académico, las colocaciones pueden utilizarse en unas disciplinas más que en otras.

El presente estudio forma parte de un proyecto de investigación más amplio, en el cual, a partir de este trabajo, planteamos integrar la nueva versión de la *SpAKWL* en HARTA de manera que, a partir de un texto, la herramienta pueda detectar las palabras académicas y sugerir las colocaciones correspondientes, en línea con lo que proponen herramientas como *LEAD* (Paquot 2012) o *Collocaid* (Frankenberg-Garcia et al. 2019) para la escritura académico-científica en inglés.

Agradecimientos

Este estudio ha sido posible gracias a la financiación del Ministerio de Ciencia e Innovación (PID2019-109683GB-C21); del Centro de Investigación de Galicia "CITIC", financiado por la Xunta de Galicia y la Unión Europea (FEDER GALICIA 2014-2020), con la ayuda ED431G 2019/01; y del Programa de Axudas á Etapa predoutoral da Xunta de Galicia, FSE Galicia 2014-2020.

Bibliografía

Ackermann, K. y Chen, Y.H. 2013. Developing the Academic Collocation List (ACL): A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12(4): 235–247.

Ahumada, I., Zamorano, J. P., García, E. D. R. y Lara, I. A. 2011. Design and development of Iberia: a corpus of scientific Spanish. *Corpora*, 6(2): 145-158.

Alonso-Ramos, M., García-Salido, M. y Garcia, M. 2017. Exploiting a Corpus to Compile a Lexical Resource for Academic Writing: Spanish Lexical Combinations. En I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček y V. Baisa (Eds.), *Proceedings of eLex 2017 conference*, páginas 571-586. Leiden, the Netherlands.

Cambridge Dictionary. Consultado el 27 de marzo de 2022 en: https://dictionary.cambridge.org/us/dictionary/.

Cobb, T., y Horst, M. 2004. Is there room for an academic word list in French?. En P. Bogaards y B. Laufer (Eds.), *Vocabulary in a Second Language. Selection, acquisition, and testing*, páginas 13-38, John Benjamins (Amsterdam/Philadelphia).

Coxhead, A. 2000. A new academic word list. *TESOL Quarterly*, 34(2): 213–238.

Drouin, P. 2007. Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, 7(2): 45-64.

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., y Sharma, N. 2019. Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, 31(1): 23-39.

García-Salido, M. 2021. Compiling an Academic Vocabulary List of Spanish. Disponible en: https://doi.org/10.13140/RG.2.2.27681.33123.

Garcia, M. y Gamallo, P. 2016. Yet another suite of multilingual NLP tools. En J. P. Leal J. L. SierraRodríguez et al. (Eds.), *Languages, Applications and Technologies. Communications in Computer and Information Science*, páginas 65–75, Springer (Cham).

Gardner, D., y Davies, M. 2013. A new academic vocabulary list. *Applied Linguistics*, 35(3): 305–327.

Gilquin, G., Granger, S., y Paquot, M. 2007. Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6(4): 319-335.

Gries, S. T. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics,* 13(4): 403–437.

Hatier, S., Augustyn, M., Tran, T. T. H., Yan, R., Tutin, A., y Jacques, M. P. 2016. French cross-disciplinary scientific lexicon: extraction and linguistic analysis. En *Proceedings of EURALEX*, páginas 355-366, Ivane Javakhishvili Tbilisi State University (Tbilsi).

Hyland, K. y Tse, P. 2007. Is there an "academic vocabulary"?. *TESOL quarterly*, 41(2): 235-253.

Hyland, K. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1): 4–21.

Jones, K. S. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1): 11-21.

Kilgarriff, A. y Renau, I. 2013. *esTenTen*, a vast web corpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences*, 95: 12-19.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J. y Suchomel, V. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1): 7–36.

Lei, L., y Liu, D. 2018. The academic English collocation list: A corpus-driven study. *International Journal of Corpus Linguistics*, 23(2): 216-243.

LINGUEE. Consultado el 28 de marzo de 2022 en: http://www.linguee.es.

Mel'čuk, I. 2012. Phraseology in the language, in the dictionary, and in the computer. *Yearbook of phraseology*, 3(1): 31-56.

Nivre, J., Marneffe, M.-C. D., Ginter, F., Goldberg, Y., Manning, C. D., Mcdonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. y Zeman, D. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. En *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016), páginas 1659–1666, European Language Resources Association (ELRA).

Oxford English Dictionary. Consultado el 27 de marzo de 2022 en: y https://www.oed.com/.

Padró, L. y Stanilovsky, E. 2012. Freeling 3.0: Towards wider multilinguality. En N. Calzolari et al., (Eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC2012), páginas 2473–2479, European Language Resources Association (ELRA).

Paquot, M. 2007. Towards a productively-oriented academic word list. En J. Walinski, K. Kredens, y S. Gozdz Roszkowski (Eds.), *Practical Applications in Language and Computers 2005*, páginas 127–140. Peter Lang (Frankfurt am main).

Paquot, M., y Bestgen, Y. 2009. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. *Language and Computers*, 68(1): 247 269.

Paquot, M. 2012. *The LEAD dictionary-cum-writing aid: An integrated dictionary and corpus tool*. En S. Granger y M. Paquot (Eds.), *Eletronic lexicography,* páginas 161-186, Oxford University Press (Oxford).

Real Academia Española: *Diccionario de la lengua española*, 23.ª ed., (versión 23.5 en línea). Consultado el 25 de marzo de 2022 en: https://dle.rae.es.

Sebastián-Gallés, N., Martí Antonín, M.A., Carreiras Valiña, M. F., y Cuetos Vega, F. 2000. *LEXESP: Léxico informatizado del español*. Barcelona: Edicions de la Universitat de Barcelona.

Straka, M., Hajic, J. y Straková, J. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. En *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016), páginas 1659–1666, European Language Resources Association (ELRA).

Tutin, A. 2007a. Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée*, 12(2), 5-14.

A Anexo 1: Nombres descartados a partir del IDF

premisa, trayectoria, especialista, déficit, sustrato, concordancia, trabajador, exclusión, ciudadano, episodio, fiabilidad, ejemplar, prioridad, geometría, infraestructura, señal, madurez, patología, creencia, amplitud, reproducción, paradigma, procedencia, almacenamiento, implantación, complicación, corrección, apertura, desplazamiento, motor, tipología, venta, puntuación, consenso,

ejecución, dispositivo, meta, asignación, autonomía, lector, continuidad, carencia, compuesto, bibliografía, supuesto, costo, normativa, emisión, correspondencia, variante, especificación, instalación, reto, experto, descenso, unión, formulación, expectativa, puesta, mediana, motivación, vínculo, inconveniente, departamento, productividad, desempeño, incertidumbre, plataforma, tejido, tensión, experimento, diferenciación, economía, barrera, satisfacción, requerimiento, dosis, sesgo, acumulación, entrevista, fundamento, regulación, expansión, explotación, transporte, abundancia, promoción, instancia, eliminación, separación, fracción, síntoma, heterogeneidad, efectividad, espectro, preferencia, difusión, presente, predominio, afirmación, transferencia, aceptación, gráfico, distinción, prevención, sugerencia, dispersión, fragmento, énfasis, varianza, canal, indicación, estadio, iniciativa, rol, procesamiento, transición, estándar, potencia.

B Anexo 2: Nombres descartados tras las tres fases

emisión, fracción, tejido, geometría, puesta, presente, prevención, ciudadano, compuesto, sustrato, trabajador, infraestructura, transferencia, carencia, explotación, estadio, transición, transporte, exclusión, varianza, ejemplar, venta, departamento, entrevista, síntoma, *dosis, patología, episodio, creencia, madurez, costo, abundancia, espectro, fragmento, iniciativa, economía, autonomía, potencia, prioridad, dispositivo, expectativa, incertidumbre, dispersión*, *preferencia, tensión*, *inconveniente, déficit, amplitud, desplazamiento, plataforma, requerimiento, expansión, separación, implantación, complicación*, *concordancia, fiabilidad, trayectoria, resto, afirmación, barrera, corrección, almacenamiento*, *formulación*, *acumulación*, *regulación*, *procesamiento, especificación, efectividad, fundamento, distinción, asignación.*

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 95-103 recibido 31-03-2022 revisado 20-05-2022 aceptado 30-05-2022

Compilación del corpus académico de noveles en euskera HARTAeus y su explotación para el estudio de la fraseología académica

Compilation of the academic corpus of novels in Basque HARTAeus and its exploitation for the study of academic phraseology

María Jesús Aranzabe,[1] Antton Gurrutxaga,[2] Igone Zabala[1]
[1] Centro HiTZ-Ixa, Universidad del País Vasco (UPV/EHU)
[2] Fundación Elhuyar
{maxux.aranzabe,igone.zabala}@ehu.eus, a.gurrutxaga@elhuyar.eus

Resumen: Se ha compilado un corpus académico de noveles para el euskera comparable con el corpus HARTA-noveles para el español. A partir del corpus se ha extraído una lista de vocabulario académico para el euskera, y sendas listas de colocaciones y fórmulas, a las que se les han asignado funciones discursivas. El objetivo último del proyecto HARTAes-vas, en el que se enmarca este trabajo, es diseñar una herramienta de ayuda a la escritura académica para las dos lenguas centrada en las combinaciones léxicas académicas, que integre diccionario y corpus.
Palabras clave: corpus académico, colocaciones, *lexical bundles*, funciones discursivas.

Abstract: An academic corpus of novices was compiled for Basque, comparable to the corpus HARTA-noveles for Spanish. A list of academic Basque vocabulary, collocations and formulas were extracted from the corpus, and then they were assigned discursive functions. The ultimate objective of the HARTAes-vas project, in which this work is framed, is to design a tool to help academic writing for Basque and Spanish focused on academic lexical combinations, integrating lexicographic information and corpora.
Keywords: academic corpus, collocations, lexical bundles, discursive functions.

1 Introducción

La introducción del euskera en ámbitos académicos, incluidos los de la educación superior, que se produjo a principios de la década de 1980, ha sido crucial para su revitalización (Zabala, 2019), ya que ha contribuido de forma muy significativa al aumento del número de hablantes y al desarrollo de los recursos expresivos necesarios para la comunicación especializada. Sin embargo, ¿podemos decir que el euskera ha "conquistado" los dominios académicos en el sentido de Laurén et al. (2002)? Dicho en otras palabras, ¿el euskera ha desarrollado los recursos expresivos necesarios para la comunicación académica en los diferentes ámbitos de especialidad? Laurén et al. (2002) defienden que a esta pregunta se puede responder de forma individual o de forma colectiva.

A nivel individual, los estudiantes universitarios adquieren los registros académicos necesarios para convertirse en miembros de la comunidad de expertos de su área gracias a numerosas tareas en las que el lenguaje resulta crucial (Biber, 2006). Algunos autores defienden que los textos académicos se elaboran siguiendo esquemas discursivos prefabricados que utilizan unidades fraseológicas semiautomáticas (Paquot, 2018), a las que nos referiremos de forma general como combinaciones léxicas académicas (CLA). No es de extrañar, por tanto, que las CLA del inglés hayan sido el objeto de estudio de numerosas investigaciones de lingüística de corpus con fines aplicados. Este auge es fácilmente explicable teniendo en cuenta el rol predominante del inglés como lengua académica internacional y el gran número de hablantes, principalmente, hablantes no-nativos, que necesitan recursos de ayuda para la

ISSN 1135-5948. DOI 10.26342/2022-69-8

redacción de artículos científicos y de trabajos académicos en general. Posteriormente, también se han ido extendiendo los proyectos de compilación de corpus académicos y de estudio de las CLA a otras lenguas como el francés, portugués de Brasil, sueco, noruego, danés y español (Alonso et al., 2017), ya que numerosos trabajos académicos se producen en las lenguas locales.

Si bien podemos pensar que el ser hablante nativo de una lengua es un factor que puede facilitar la producción de textos en dicha lengua, también está generalmente aceptado que no hay hablantes nativos de los registros académicos, y que éstos se adquieren gracias a la experiencia lingüística de lectura y producción de textos académicos. Debido a la internacionalización de la comunicación académica y del uso cada vez más extendido del inglés como lengua de instrucción y de elaboración de trabajos académicos en la educación superior, existe la preocupación de que los estudiantes universitarios, e incluso los expertos tengan cada vez más dificultades para adquirir los recursos expresivos necesarios para la comunicación académica en L1 diferentes del inglés (Swales, 2000; Görlach, 2002; Laurén et al., 2002; Johansson Kokkinakis et al., 2012; Gotti, 2012). Es por esto que se hace necesario elaborar recursos y herramientas de ayuda a la escritura académica también para otras lenguas. En el caso del euskera, la idea generalizada es que no ha habido suficiente tiempo para el desarrollo y estabilización de los registros académicos (Zabala et al., 2011; Zabala et al., 2021), y la preocupación por el impacto de la creciente internacionalización es aún mayor.

Las CLA son segmentos de palabras recurrentes que pueden o no ser semánticamente composicionales y que cubren funciones retóricas como añadir información, presentar ejemplos o expresar posibilidad. Incluyen colocaciones (*ondorioak atera* "extraer conclusiones"), locuciones (*oro har* "en general") y fórmulas, que coinciden en gran medida con las denominadas en la literatura *lexical bundles* (*azpimarratu beharra dago* "hay que remarcar"). La recurrencia es el resultado de su uso frecuente en discursos compartidos por la comunidad académica y, por lo tanto, las CLA constituyen un tipo de unidades privilegiadas para el estudio del nivel de desarrollo de los registros académicos del euskera.

Este trabajo se enmarca en el proyecto HARTAes-vas, proyecto coordinado entre la Universidade da Coruña y la Universidad del País Vasco (UPV/EHU) y financiado por el Ministerio de Ciencia e Investigación. El equipo de la Coruña cuenta con un corpus de expertos y otro de noveles en español, compilados en un proyecto anterior, ha explotado dichos corpus para la extracción y clasificación de CLA y ha desarrollado una herramienta de consulta de la fraseología académica en español (Herramienta de Ayuda a la Escritura Académica: HARTA)[1] (García-Salido et al., 2018). La colaboración con el grupo de la Coruña es fundamental para poder contar con elementos de comparación entre dos lenguas que se diferencian por su tipología y por su situación sociolingüística. En este trabajo describimos el corpus académico de noveles para el euskera compilado dentro del proyecto HARTAvas y su explotación para el estudio de la fraseología académica de cara a crear una herramienta de consulta coordinada con HARTAes, que ayude a la escritura académica en euskera, y que contribuya al desarrollo y estabilización de los registros académicos en dicha lengua. Hemos elaborado un corpus comparable con el corpus HARTA de noveles para el español (Villayandre, 2018; García-Salido et al., 2018) con el fin de poder contrastar los resultados obtenidos para el euskera, que es una lengua aglutinante en proceso de normalización, con los obtenidos para el español, lengua flexiva y bien desarrollada.

En el apartado 2 describimos la constitución del corpus de noveles HARTAeus. El apartado 3 lo dedicamos a la extracción, validación y análisis de las fórmulas y colocaciones académicas a partir del corpus. Finalmente, los apartados 4 y 5 recogen los resultados y conclusiones obtenidos hasta el momento.

2 Constitución del corpus HARTAeus

El corpus HARTAeus de noveles para el vasco está constituido por Trabajos Fin de Grado (TFG) y Trabajos Fin de Máster (TFM), por lo que se puede considerar una muestra de la escritura académica de los estudiantes universitarios. Al ser un corpus comparable con el corpus HARTA-noveles del español, su diseño ha seguido los criterios definidos en la

[1] http://www.dicesp.com:8083/search

creación de éste último (Villayandre, 2018). De este modo, los textos del corpus HARTAeus están divididos en cuatro secciones (Arte y Humanidades, Biología y Ciencias de la Salud, Ciencias Físicas y Ciencias Sociales), que se dividen a su vez en distintos dominios temáticos como, por ejemplo, Biología y Medicina en la sección de Biología y Ciencias de la Salud (ver Anexo 1 para la división del corpus en secciones y dominios).

El proceso de compilación ha comprendido seis fases: i) recolección de documentos para el corpus: los documentos en formato PDF proceden en su mayoría del repositorio ADDI (Archivo Digital para la Docencia e Investigación) de la Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU); ii) normalización: se ha realizado la conversión al formato DOCX de los documentos originales con el fin de realizar la limpieza y ordenación de las secciones y párrafos de los textos, y eliminar las marcas sobrantes; iii) codificación: se han introducido las etiquetas que marcan el inicio y final de las distintas secciones de los textos (título, resumen, presentación, introducción, cuerpo, metodología, resultados y discusión, conclusiones, agradecimientos, notas al pie de página y anexos); iv) se han incorporado de manera automática los textos en el entorno de trabajo Garaterm (Zabala et al, 2013) para su posterior procesamiento; v) almacenamiento: se han anotado los metadatos y las secciones de los textos del corpus de referencia con etiquetas XML. Asimismo, se ha adaptado el conversor de TEI existente en la plataforma Garaterm con el fin de mantener la estructura y las etiquetas de XML utilizadas para marcar los apartados de los documentos originales, y vi) procesamiento: el corpus ha sido tokenizado, lematizado y analizado morfológicamente por medio de Eustagger (Alegría et al., 2002), analizador morfológico y etiquetador de partes del discurso para el euskera. Por medio de este proceso se obtiene la información del lema y los rasgos morfosintácticos necesarios para poder extraer las combinaciones de palabras candidatas a colocaciones: N+N, N+V, N+Adj.

El resultado de este proceso ha sido la creación de un corpus académico monolingüe integrado por 398 textos (71 % TFG y 29 % TFM) y 3.285.098 palabras distribuidas en cuatro áreas de conocimiento (Tabla 1). La distribución en los distintos dominios temáticos puede verse en el Anexo 1.

Secciones del corpus	TFG nº de palabras (nº de documentos)	TFM nº de palabras (nº de documentos)	Total de palabras y documentos
Arte y Humanidades	450.859 (62)	203.831 (12)	654.690 (74)
Biología y Ciencias de la Salud	271.932 (65)	153.533 (26)	425.465 (91)
Ciencias Físicas	1.035.599 (121)	378.685 (36)	1.414.284 (157)
Ciencias Sociales	559.791 (46)	230.868 (30)	790.659 (76)
Totales	2.318.181 (294)	966.917 (104)	3.285.098 (398)

Tabla 1: Distribución de palabras y documentos por secciones y por tipología de textos (TGF y TFM).

Como se puede observar en la Tabla 1, el número de TFM es menor que el de TFG. Esto se debe a que el número de TFM que se elaboran en euskera es muy pequeño, a que bastantes trabajos están protegidos por cláusulas de confidencialidad y a que muchos de ellos no se publican en la plataforma ADDI.

3 Extracción y validación de CLA

Para la extracción y validación de las CLA hemos añadido tres módulos al extractor de terminología para el euskera Erauzterm (Alegría et al., 2004): un módulo para la identificación del vocabulario académico, un segundo módulo para la identificación de colocaciones académicas y un tercer módulo para la identificación de fórmulas académicas. Debido a las deficiencias del desarrollo de los registros académicos en euskera, encontramos CLA que superan los umbrales de frecuencia y dispersión establecidos pero que pueden ser consideradas como incorrectas o no óptimas. Como quiera que el último objetivo del proyecto es desarrollar una herramienta de ayuda a la escritura, en el proceso de validación hemos ido identificando las CLA incorrectas y elaborando una tipología de estas.

3.1 Extracción y validación del vocabulario académico

El módulo para la elaboración de la lista de vocabulario académico utiliza como contraste el corpus Dabilena,[2] obtenido de la web

[2] https://dabilena.elhuyar.eus/

(300.217.903 palabras): es el corpus mayor que tenemos para el euskera y ha sido elaborado por Elhuyar. Los candidatos se pueden filtrar según su categoría gramatical (N, V, Adj., Adv…) y según varias medidas de frecuencia y dispersión: nº de dominios y partes del texto, porcentaje de textos del corpus en los que aparecen, así como *log-likelihood*, frecuencia y umbrales de frecuencia esperada.

La tarea de identificación del vocabulario académico no es trivial, ya que se trata de identificar los lemas característicos del discurso académico, pero descartando los términos específicos de una determinada área de especialidad. Para que la lista obtenida para el euskera sea comparable con la obtenida para el español en el proyecto HARTA, se han probado algunas de las medidas descritas en García-Salido (2021). Se han realizado dos experimentos con 3 condiciones comunes: presencia de los candidatos en las 4 secciones del corpus y en el 20 % de los documentos, y valores de *log-likelihood* positivos. En el segundo experimento se ha añadido la condición de que la frecuencia no sea 3 veces superior a la frecuencia esperada en cada uno de las cuatro secciones del corpus. Los resultados obtenidos se resumen en las Tablas 2 y 3.

Experimento 1			
	candidatos	validados	precisión
N	443	338	76,30 %
V	167	165	98,80 %
ADJ	147	128	87,07 %
ADV	73	53	72,60 %
Total	830	684	82,41 %

Tabla 2: Validación de los candidatos para la lista de vocabulario académico: presencia en las 4 secciones del corpus y en el 20 % de los documentos + *log-likelihood* positiva.

Experimento 2			
	candidatos	validados	precisión
N	160	116	72,50 %
V	81	81	100 %
Adj.	70	62	88,57 %
Adv.	49	34	68,39 %
Total	360	293	81,39 %

Tabla 3: Validación de los candidatos para la lista de vocabulario académico: presencia en las 4 secciones del corpus y en el 20 % de los documentos + *log-likelihood* positiva + F< 3 $F_{esperada}$ en cada sección.

Como se puede ver en las Tablas 2 y 3, la condición añadida en el experimento 2, encaminada a descartar los términos específicos de las diferentes áreas de especialidad, no aumenta la precisión y, además, disminuye la cobertura en un 57 %.

3.2 Extracción y validación de colocaciones académicas

El módulo de extracción y validación de colocaciones académicas está conectado con el vocabulario académico, de tal manera que permite filtrar los candidatos a colocaciones con un solo lema o con los dos lemas incluidos en la lista de vocabulario académico. Las combinaciones candidatas a colocaciones académicas se extraen utilizando las medidas de asociación desarrolladas en Gurrutxaga et. al. (2011, 2018) y atienden a los siguientes patrones sintácticos: Sujeto-Verbo (*emaitzek erakutsi* "resultados mostrar"), Verbo-Objeto (*helburu lortu* "objetivo conseguir = conseguir objetivos"; *emaitzei erreparatu* "resultados-DATIVO atender = atender a los resultados", *lanetik atera* "trabajo-ABLATIVO = obtener a partir del trabajo"), Nombre-Modificador (*helburu nagusi* "objetivo principal", *funtsezko elementu* "fundamental elemento = elemento fundamental"), N-(posposición)-N (*lagin tamaina* "muestra tamaño = tamaño de muestra"; *laginaren tamaina* "muestra de la tamaño = tamaño de la muestra").

Se han excluido los nombres ligeros *mota* "tipo", *kopuru* "número", *falta* "falta", *multzo* "conjunto", *zati* "parte" y *maila* "nivel". También se han descartado los adjetivos *bakar* "único", *berdin* "igual", *ezberdin/desberdin* "diferente" y los modificadores prenominales *goiko* "superior", *beheko* "inferior", *honako*

"este", *horrelako* "similar", *hurrengo* "siguiente". La razón de este descarte es que, a pesar de que dan lugar a combinaciones recurrentes con nombres académicos, en la mayoría de los casos no se trata de colocaciones.

En la Tabla 4 se resumen algunos de los resultados obtenidos. El número total de tokens normalizado es de 9.471 tokens por millón de palabras.

	Types	Tokens
N-Modif.	495	13.221
N(pos.)N	142	4.112
Sujeto-V	11	192
V-Objeto	357	13.589
Total	1.005	31.114

Tabla 4: Colocaciones extraídas del corpus HARTAeus.

3.3 Extracción y validación de fórmulas académicas

Para la detección de fórmulas se han extraído n-gramas entre 2 y 5 elementos. Se han filtrado únicamente los que estaban presentes en las 4 secciones del corpus y cuya frecuencia era igual o superior a 10 apariciones por millón de palabras, criterio generalmente utilizado para la identificación de *lexical bundles* (Biber et al., 1999). A la hora de validar los candidatos, hemos asignado una o más funciones discursivas a cada n-grama validado, siguiendo la tipología usada en el proyecto HARTA para el español (García-Salido et al., 2019), ya que, con el fin de que los usuarios puedan encontrar las fórmulas fácilmente en la herramienta de consulta, es más detallada que la de Biber et al. (2004) y la de Hyland (2008). Además, una de las tareas principales del proyecto consiste en comparar las fórmulas extraídas de los corpus de noveles vasco y español.

En el proceso de validación y de asignación de función discursiva a los n-gramas, en algunos casos hemos eliminado algún elemento que no aportaba valor semántico a la función discursiva, como es la conjunción *eta* "y". Así por, ejemplo, si un candidato validado era *eta hala ere* "y aun así", lo hemos eliminado y hemos mantenido únicamente la fórmula de dos elementos *hala ere* "aun así". Con este procedimiento, hemos identificado algunas fórmulas monoléxicas plurimorfémicas, que en principio no esperábamos recoger. Por ejemplo, el n-grama *eta ondorioz* "y por consiguiente" lo hemos validado como la fórmula *ondorioz* "por consiguiente" y le hemos asignado la función "expresar consecuencia".

Una vez validados los n-gramas y asignadas las funciones discursivas, hemos identificado las variantes de una misma fórmula. Por ejemplo, *aipatu den moduan* "como se ha mencionado" y *aipatu dugun moduan* "como hemos mencionado" son dos variantes de la misma fórmula, y lo mismo sucede con las fórmulas *horrek esan nahi du* "eso quiere decir" y *horrek ez du esan nahi* "eso no quiere decir". En estos casos, las variantes las hemos considerado como un solo *type*. Se han validado y clasificado 644 fórmulas (*types*), 1.028 variantes y 125.398 tokens (38.171 tokens por millón de palabras). Como puede verse en la Tabla 5, a falta de estrategias complementarias para la extracción de fórmulas monoléxicas, las fórmulas de 2 palabras son las más numerosas.

	Número de palabras	Types
Fórmulas académicas	1 palabra	42
	2 palabras	490
	3 palabras	126
	4 palabras	12
	5 palabras	4
	Totales	644

Tabla 5: Nº de fórmulas académicas validadas una vez analizada la variación.

3.4 Identificación y clasificación de CLA incorrectas

Algunas colocaciones y fórmulas que llegan a los umbrales de frecuencia y dispersión establecidos, pueden considerarse como incorrectas o no óptimas. Estas CLA las hemos recogido y clasificado para poder así tenerlas en cuenta a la hora de diseñar la herramienta de consulta ya que, aunque no son muy numerosas, presentan un importante grado de recurrencia y compiten con formas más correctas o genuinas. En la Tabla 6, ofrecemos una clasificación preliminar y algunos ejemplos.

Ortografia no-estándar	
kontutan hartu behar da "hay que tener en cuenta"	*kontuan hartu behar da*
gutxi gora behera "poco más o menos"	*gutxi gorabehera*
pausuak eman "dar pasos" pausu "descanso"	*pausoak eman* pauso "paso"
Demostrativo de 1er grado como anáfora	
honek ez du esan nahi "esto no quiere decir" Demostrativo de 1er grado (catáfora)	*horrek ez du esan nahi* Demostrativo de 2º grado (anáfora)
Orden de palabras inadecuado	
Lan honen helburua […] *da* "El objetivo de este trabajo […] es"	*Lan honen helburua da* […]
Forma incorrecta de un conector	
Alde batetik […] *eta beste aldetik,* […] "Por un lado […] y por el otro lado […]"	*Alde batetik* […] *eta, bestetik,* […] "Por un lado […] y por el otro […]"
Asignación de una función discursiva incorrecta	
Hau da "esto es" "expresar causa"	*Hau da* "reformular"
Colocación incorrecta desde el punto de vista semántico-sintáctico	
datuek adierazi "datos expresar"	*datuek erakutsi* "datos mostrar"
datu adierazgarri "dato representativo"	*datu esangarri* "dato significativo"
Calcos	
besteen artean "entre otros"	*besteak beste* "entre otros"

Tabla 6: Clasificación y ejemplos de CLA incorrectas.

4 Resultados y discusión

No contamos con un corpus de expertos para el euskera comparable con el corpus HARTA de expertos, por lo que, para analizar los datos obtenidos hasta el momento, los contrastaremos con los ofrecidos en García-Salido (2021) y en Alonso y Zabala (2022).

La lista de vocabulario académico compilada hasta el momento para el euskera cuenta con 684 lemas. Esta lista es bastante más reducida que la obtenida tras contrastar los resultados de diferentes técnicas para el español (García-Salido, 2021): 833 lemas. La diferencia puede estar motivada por el descarte de los lemas con valores de *log-likelihood* negativos, ya que entre los lemas descartados puede haber algunos que son muy utilizados en textos generales pero que activan significados específicos en los discursos académicos. Nuestra idea es seguir completando la lista obtenida hasta el momento, probando otras técnicas y medidas.

Sin embargo, la principal aplicación de la lista de lemas académicos en el proyecto HARTAvas es la extracción de colocaciones académicas. Para ello, es fundamental la lista de nombres, ya que son los que constituyen las bases de las colocaciones, el número de N obtenidos para el euskera es menor pero comparable al del español: 338 eus / 358 es.

El número de colocaciones académicas obtenidas es también comparable al obtenido a partir del corpus HARTA de noveles para el español (Alonso y Zabala, 2022): 1.005 types eus / 1.197 es. Aunque hay que tener en cuenta que el corpus del euskera es mayor que el del español, que cuenta con 2 M de palabras. Aun siendo menor el número de colocaciones extraídas para el euskera, el número de tokens por M de palabras es notablemente superior: 9.471 tokens/M eus / 6.897 tokens/M es. Por lo tanto, como primera aproximación se puede decir que las colocaciones detectadas para el euskera son más recurrentes que las detectadas para el español en el corpus de noveles.

El número de fórmulas (types) extraídas para el euskera es notablemente superior al obtenido a partir del corpus de noveles en español: 644 types eus / 472 types es. También existe diferencia en la frecuencia de dichas fórmulas: 38.171 tokens/M eus / 20.474 tokens/M es. El mayor número de fórmulas detectadas en euskera podría estar relacionado con el menor grado de fijación de las fórmulas académicas en esta lengua, y con la variación entre fórmulas más genuinas y fórmulas calcadas del español: *besteen artean* (calco) / *besteak beste*; *orokorrean* (calco) / *oro har.* El número mayor de tokens, podría indicar una menor riqueza expresiva de los noveles vascos, que les haría recurrir más frecuentemente a las mismas secuencias dando lugar a un discurso más repetitivo. De cualquier modo, se requiere un análisis más minucioso de los criterios de validación que hemos utilizado para una y otra lengua, así como de los datos, con el fin de poder hacer una comparación más precisa de las CLA obtenidas en una y otra lengua de cara al diseño de la herramienta de ayuda a la escritura académica para las dos lenguas.

5 Conclusiones

A pesar de que el euskera es una lengua minorizada que se introdujo hace solo unas décadas en la enseñanza superior, hemos logrado compilar un corpus de trabajos académicos en euskera (TFG y TFM) comparable, e incluso algo más extenso, que el corpus de noveles para el español.

A partir del corpus hemos obtenido una lista de vocabulario académico en euskera (644 lemas), que aunque debe de considerarse preliminar, nos ha permitido identificar un gran número de colocaciones académicas (1.005 types).

Hemos extraído también n-gramas de 2, 3, 4 y 5 elementos, que hemos validado y a los que hemos asignado funciones discursivas partiendo de la tipología utilizada para HARTAes. Los n-gramas nos han permitido detectar fórmulas poliléxicas o *lexical bundles* como (*kontuan hartu beharrekoa da* "hay que tener en cuenta"), pero en el proceso de validación, hemos podido detectar también 42 fórmulas monoléxicas o *morphemic bundles* como *laburbilduz* "en resumen". Así hemos obtenido 644 fórmulas y 1.028 variantes, a las que les hemos asignado funciones discursivas.

Por último, hemos desarrollado una tipología de fórmulas incorrectas, de cara a elaborar su tratamiento lexicográfico en la herramienta de consulta.

Estamos implementando técnicas de semántica distribucional, con el fin de utilizar los corpus comparables del español y del euskera para la detección de fórmulas, sobre todo fórmulas monoléxicas, y equivalentes de colocaciones y fórmulas entre las dos lenguas.

Además, hemos comenzado la tarea de comparación más minuciosa de las listas de CLA elaboradas para las dos lenguas, con el fin de obtener una clasificación que tenga en cuenta las características tipológicas del español y del euskera. Dicha comparación nos servirá también para decidir el diseño de la herramienta de consulta para ambas lenguas.

Agradecimientos

Este trabajo es parte del proyecto HARTAvas (PID2019-109683GB-C22), financiado por el Ministerio de Ciencia e Innovación.

Bibliografía

Alegria, I., M.J. Aranzabe, A. Ezeiza A., N. Ezeiza, y R. Urizar R. 2002. Robustness and customisation in an analyser/lemmatiser for Basque. *En Third International Conference on Language Resources and Evaluation (LREC): Customizing Knowledge in NLP Applications-Strategies, Issues and Evaluation Workshop*, páginas 1-6, Las Palmas de Gran Canaria (Spain).

Alegría, I, A. Gurrutxaga, P. Lizaso, X. Saralegi, S. Ugartetxea, y R. Urizar. 2004. An Xml-Based Term Extraction Tool for Basque. En *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, páginas 1733-1736, Lisboa (Portugal).

Alonso-Ramos, M., M. García-Salido, y M. Garcia. 2017. Exploiting a Corpus to Compile a Lexical Resource for Academic Writing: Spanish Lexical Combinations. En Kosem, I., J. Kallas, C. Tiberius, S. Krek, M. Jakubíček, V. Baisa (Eds). *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, páginas 571-586, Leiden (the Netherlands).

Alonso-Ramos, M. y I. Zabala. 2022. HARTAes-vas: Combinaciones léxicas para una Herramienta de ayuda a la redacción de textos académicos en español y en vasco. En *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations, SEPLN*, September, A Coruña (Spain).

Biber, D. 2006. *University Language. A corpus-based study of spoken and written registers*. John Benjamins, Amsterdam.

Biber, D., E. Finegan, S. Johanson, S. Conrad, y G. Leech. 1999. *Longman Grammar of Spoken and Written Englis*h. Longman, London.

Biber, D., S. Conrad, y C. Viviana. 2004. If you look at…: lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3):371-405.

García-Salido, M., M. García, M., Villayandre, y M. Alonso-Ramos. 2018. A Lexical Tool for Academic Writing in Spanish based on Expert and Novice Corpora. En Calzolari N. et al. (Eds). *Proceedings of the Eleventh International Conference on Language*

Resources and Evaluation (LREC 2018), páginas 260-265, Miyazaki (Japan).

García-Salido, M., M. Garcia, y M. Alonso-Ramos. 2019. Identifying lexical bundles for an academic writing assistant in Spanish. En Corpas Pastor, G. y R. Mitkov (Eds). *Computational and Corpus-Based Phraseology*. Volume 11755 of *Lecture Notes in Artificial Intelligence*, páginas 144-158, Springer, Berlin.

García-Salido, M. 2021. Compiling an Academic Vocabulary List of Spanish. DOI: 10.13140/RG.2.2.27681.33123

Görlach, M. 2002. *Still More Englishes*. John Benjamins, Amsterdam.

Gotti, M. 2012. Variation in Academic Texts. En Gotti, M. (Ed). *Academic Identity Traits. A Corpus Based Investigation*, páginas 21-42, Peter Lang (Switzerland).

Gurrutxaga, A. e I. Alegria. 2011. Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. En *Proceedings of the Workshop on Multiword Expressions: from parsing and generation to the real world*, páginas 2–7, Portland, Oregon (USA).

Gurrutxaga, A., I. Alegria, y X. Artola. 2018. Caracterización computacional de la idiomaticidad: aplicación a la combinación nombre+verbo en euskera. En Ruiz Miyares, L. (Ed). *Estudios de Lexicología y Lexicografía. Homenaje a Eloína Miyares Bermúdez.* Santiago de Cuba (Cuba).

Hyland, K. 2008. As can be seen: lexical bundles and disciplinary variation. *English for Specific Porpuses*, 27(1): 4-21.

Johansson Kokkinakis, S., E. Sköldberg, B. Henriksen, K. Kinn, y J. Bondi Johannessen. 2012. Developing Academic Word Lists for Swedish, Norwegian and Danish a Joint Research Project. En Fjeld, R.V. y J.M. Torjusen (Eds). *Proceedings of the 15th EURALEX International Congress,* páginas 563–569, University of Oslo (Norway).

Laurén, Ch., J. Myking, y H. Picht. 2002. Language and domains: a proposal for a domain dynamics taxonomy. *LSP and Professional Communication*, 2(2):23-30.

Paquot, M. 2018. Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights from A Study of EFL Learnes's Use of Statistical Collocations. *Language Assessment Quarterly*, 15(1):29-43.

Swales, J. 2000. Language for Specific Purposes. *Annual review of Applied Linguistics,* 20:59-76.

Villayandre, M. 2018. "HARTA" de noveles: un corpus de español académico. *CHIMERA: Revista De Corpus De Lenguas Romances Y Estudios Lingüísticos*, 5(1): 131–140.

Zabala, I., I. San Martin, M. Lersundi, y A. Elordui. 2011. Graduate teaching of specialized registers in a language in the normalization process: Towards a comprehensive and interdisciplinary treatment of academic Basque. En Maruenda-Bataller, S. y B. Clavel-Arroita (Eds). *Multiple voices in academic and professional discourse*, páginas 208-218, Cambridge Scholars (Newcastle upon Tyne, UK).

Zabala, I., M. Lersundi, I. Leturia, I. Manterola, y G. Santander. 2013. GARATERM: euskararen erregistro akademikoen garapenaren ikerketarako lan-ingurunea. En Alberdi, X. y P. Salaburu (Eds). *Ugarteburu terminologia jardunaldiak (V). Terminologia naturala eta terminologia planifikatua euskararen normalizazioari begira*, páginas 98-114, Servicio Editorial de la UPV/EHU (Bilbao).

Zabala, I. 2019. The elaboration of Basque in Academic and Professional Domains. En Grenoble, L., P. Lane, y U. Røyneland (Editor-in-Chief), Igartua, I. y L. Oñederra (Basque Eds). *Linguistic Minorities in Europe Online*, De Gruyter Mouton.

Zabala, I., M.J. Aranzabe, y I. Aldezabal. 2021. Retos actuales del desarrollo y aprendizaje de los registros académicos orales y escritos del euskera. *Círculo de Lingüística Aplicada a la Comunicació*n, 88:31-50.

A Anexo 1: Corpus HARTAeus

Distribución de palabras y documentos (TFG y TFM) por secciones del corpus y dominios temáticos.

Secciones del corpus	Dominios temáticos	TFG nº de palabras (nº de documentos)	TFM nº de palabras (nº de documentos)	Total de palabras y documentos
Arte y Humanidades	Arte	42.956 (6)	0	42.956 (6)
	Lingüística	207.443 (27)	203.831 (12)	411.274 (39)
	Literatura	90.139 (12)	0	90.139 (12)
	Historia y Cultura	110.321 (17)	0	110.321 (17)
	Biblioteconomía y documentación	0	0	0
	Totales	450.859 (62)	203.831 (12)	654.690 (74
Biología y Ciencias de la Salud	Biología	182.906 (44)	153.533 (26)	336.439 (70)
	Medicina	89.026 (21)	0	89.026 (21)
	Totales	271.932 (65)	153.533 (26)	425.465 (91)
Ciencias Físicas	Ciencias de la Tierra	52.263 (7)	0	52.263 (7)
	Física	113.027 (16)	0	113.027 (16)
	Ingeniería	656.156 (69)	20.031 (1)	676.187 (70)
	Informática	75.160 (8)	332.982 (32)	408.142 (40)
	Química	138.993 (21)	25.672 (3)	164.665 (24)
	Totales	1.035.599 (121)	378.685 (36)	1.414.284 (157)
Ciencias Sociales	Economía y Empresa	158.433 (10)	0	158.433 (10)
	Educación	102.335 (16)	230.868 (30)	333.203 (46)
	Sociología	233.632 (16)	0	233.632 (16)
	Derecho	65.391 (4)	0	65.391 (4)
	Totales	559.791 (46)	230.868 (30)	790.659 (76)

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 105-116 recibido 31-03-2022 revisado 12-05-2022 aceptado 30-05-2022

Extraction and Semantic Representation of Domain-Specific Relations in Spanish Labour Law

Extracción y representación de relaciones específicas de dominio en la legislación laboral española

Artem Revenko,[1] **Patricia Martín-Chozas**[2]

[1]Semantic Web Company, Vienna, Austria
[2]Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
artem.revenko@semantic-web.com, patricia.martin@upm.es

Abstract: Despite the freedom of information and the development of various open data repositories, the access to legal information to general audience remains hindered due to the difficulty of understanding and interpreting it. In this paper we aim at employing modern language models to extract the most important information from legal documents and structure this information in a knowledge graph. This knowledge graph can later be used to retrieve information and answer legal question. To evaluate the performance of different models we formalize the task as event extraction and manually annotate 133 instances. We evaluate two models: GRIT and Text2Event. The latter model achieves a better score of ≈ 0.8 F_1 score for identifying legal classes and 0.5 F_1 score for identifying roles in legal relations. We demonstrate how the produced legal knowledge graph could be exploited with 2 example use cases. Finally, we annotate the whole Workers' Statute using the fine-tuned Text2Event model and publish the results in an open repository.
Keywords: Information Extraction. Knowledge Graphs. Semantic Web. Legal Domain.

Resumen: A pesar de la actual libertad de información y del desarrollo de diferentes repositorios de datos abiertos, el acceso a la información jurídica al público general sigue suponiendo un problema debido a la dificultad de comprensión e interpretación de dicha información. En este artículo, nuestro objetivo es emplear modelos de lenguaje punteros para extraer información relevante de documentos jurídicos; así como estructurar esta información en un grafo de conocimiento, con el objetivo de que este grafo pueda utilizarse más adelante para recuperar información y responder preguntas sobre el dominio jurídico. Para evaluar el rendimiento de los diferentes modelos, hemos formalizado este proceso como una tarea como extracción de eventos, y hemos anotado manualmente 133 instancias. Evaluamos dos modelos: GRIT y Text2Event. El último modelo consigue mejores resultados, de ≈ 0.8 F_1 para identificar clases jurídicas y de 0.5 F_1 para identificar roles en relaciones jurídicas. Asimismo, ejemplificamos cómo el grafo producido podría explotarse con diferentes casos de uso. Finalmente, hemos anotado todo el Estatuto de los Trabajadores con el modelo Text2Event y publicado los resultados en un repositorio abierto.
Palabras clave: Extracción de Información. Grafos de Conocimiento. Web Semántica. Dominio Jurídico.

1 Introduction

Due to its specific nature, the legal domain has always been a complex area for non legal users. The challenges include *finding* the correct document for a purpose and *interpreting* the document. With the recent rise of the data sharing and open data technologies in the last decade, legal knowledge is more accessible than ever. Well-known legal practitioners have already exposed their legal data in open and machine readable formats, developing platforms such as the European Data Portal[1], a platform funded by the Eu-

[1]`https://data.europa.eu/`

ISSN 1135-5948. DOI 10.26342/2022-69-9

ropean Union and managed by the Publications Office that gathers legal data from different subdomains such as justice, legal system and government. At a national level, one of the most important sources of legal data in Spain is the Official State Gazette[2], which is constantly being updated and accessed by lawyers in Spain. It contains documents in the labour law domain, such as state collective agreements and the Spanish Workers Statute. This availability of legal data efficiently addresses the task for finding the correct documents and accessing them. Yet, the interpretation of legal data and, therefore, exploitation of the legal results by general public remains an important challenge (Robaldo et al., 2019). We are driven by the idea of tackling this challenge and enabling more human-friendly interfaces for accessing this kind of information. We choose the labour law sub-domain for our experimentation as this domain is relevant for everyday use by general audience and the tasks of enabling natural language search, information retrieval, question answering over the labour law are highly demanded. To solve these kind of tasks and ease the access to legal information from general audience, we aim at structuring legal data into a knowledge graph.

Modern (multilingual) language models have celebrated many successes on various NLP tasks (named entity recognition, relation extraction, paraphrase detection, question answering, etc.). We employ these models to tackle our challenge as well. We noticed that most models are trained with general corpora and, therefore, fail to identify the peculiarities of domain specific texts. The lack of domain specific annotated corpora and domain specific language resources published in machine readable formats hinders the fine-tuning of the models.

Consequently, the purpose of this paper is twofold: on the one hand, we test different language models on a domain specific corpus to extract relations amongst terms and, on the other hand, we provide annotated data in the labour law domain and structure the results in Semantic Web formats so they can be reused in the future by upcoming researchers.

This work (code, input and output data) is openly available and published in a GitHub repository[3].

[2] `https://www.boe.es/`

[3] `https://github.com/pmchozas/term_relex/`

1.1 A look at the Semantic Web

More than 20 years ago, the *World Wide Web Consortium (W3C)* promoted the publication of data in structured, machine-readable and interlinked formats, in which the meaning of data can be interpreted by machines to achieve more complex and effective data understanding. This initiative is known as the *Semantic Web* or the *Web of Data* (Berners-Lee, Hendler, and Lassila, 2001).

The most common format for publishing data on the Semantic Web is the *Resource Description Framework (RDF)*. RDF is at the core of the Linked Open Data paradigm for publishing information, based on the *Linked Data Principles* (Berners-Lee, 2006). According to these principles, resources need to be identified by a *Uniform Resource Identifier (URI)* (a unique identifier that follows the HTTP standard web protocols), and that resources need to contain pointers to other resources.

The inner structure of Linked Data is determined by the *ontology* (also known as *model* or *vocabulary*) that defines how to represent the concepts of a certain domain (Chandrasekaran, Josephson, and Benjamins, 1999). These ontologies are composed of classes, relations, rules and restrictions.

In this paper, we make use of RDF to structure the extracted knowledge following the Linked Data Principles, publishing the results in a machine readable and linked dataset, also called *knowledge graph*.

1.2 Motivation

We have already approached the task of creating legal knowledge graph in one of the pilots (Spanish labour law pilot) of the [Lynx][4] project, an Innovation Action funded by the European Union's Horizon 2020, that was active from 2017 until 2021. Although the project already ended, this pilot served as an inspiration to work deeply on the extraction of knowledge over labour law documentation, since several issues were spotted during its development:

1. The labour law texts are highly domain specific. This fact hinders the reuse of already pre-trained language models that are usually trained with texts from the general domain.

[4] `https://lynx-project.eu/`

2. Annotated corpora in this domain are scarce; even after annotating a part of the Statute, the size of the resulting annotated corpus is not sufficient for a language model to obtain good results.

Therefore, within this paper we tackle those issues trying to untangle labour law data, easing the understanding of duties and rights related to the labour domain to anyone willing to know them. For instance, if we suppose that a worker is in trouble with a company for taking a leave from work, we may want to answer questions such as:

Q1) *In which situations can a worker claim for a right?*

Q2) *When must a worker come back to work after a leave from work?*

The rest of the paper is divided as follows: Section 2 explains the statement of the task; Section 3 describes the corpus and the annotation process; Section 4 identifies the language models applied in the experiment; Section 5 reports on the results and evaluation; Section 6 covers the conversion of the results to Semantic Web formats; Section 7 shows examples on how to exploit the resulting knowledge graph; Section 8 gathers related work on event extraction and legal ontologies and, finally, Section 9 collects conclusions and future work.

2 Task Statement

As described in previous work by the authors (Martín-Chozas and Revenko, 2021), we analysed the nature of conceptual relations in legal texts and noticed that labour law texts are full of Hohfeld *deontic relations*, part of the *Hohfeldian fundamental relations* (Hohfeld, 1913), that are divided into two sets of relations:

- *Deontic relations*, that are those that modify ordinary actions: Right, Duty, No-Right and Priviledge.
- *Potestative relations*, that are those that modify deontic relations: Power, Liability, Disability and Immunity.

In this work, we focus on the extraction of deontic relations (see Figure 1), since they are the basis of the fundamental relations and the most common within the Spanish labour law.

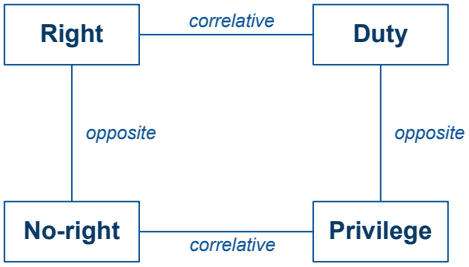


Figure 1: Hohfeld's Deontic Relations.

Hohfeld classes provide information about the particular type of legal relation. However, it is difficult to use this information alone. In fact, more complex use cases, such as question answering over legal texts or the merge of different legal documents into a single knowledge graph, require extracting additional information about the participants of these relations.

We identify the following roles of the participants: subjects of relations, objects of relations and complements of relations. These roles correspond to the classes identified by the Provision Model (see Section 6):

- The *subject* is the agent of the action, who performs the action.
- The *object* is the patient of the action, who receives the action.
- The *complement* is the item which is handled in the relation.

Consequently, the model should be capable of (1) classifying a string from a legal text into one of Hohfeld relation classes (if there is one); (2) identifying the roles of participants of the extracted relation. We formalise the task as *sentence-level event extraction*[5]. Event extraction is an essential task for natural language understanding, aiming to transform the text into structured event records (Doddington et al., 2004; Ahn, 2006). These event records can further be transformed into a knowledge graph, see Section 6.

[5]In preliminary experiments we also considered an alternative formalisation as a relation extraction task (Hendrickx et al., 2019). However, in that case, we need to extract the entities in a separate step and then use relation extraction to identify relation between those entities. Moreover, the roles that we want to extract are better described as roles within a sentence rather than being in a relation with a particular entity. In sight of these difficulties, we refrained from using relation extraction as the task formulation.

We illustrate the event extraction task in Figure 2 with an example from Spanish labour law. From the sentence, we extract an event record of type "Right" corresponding to the Hohfeld deontic class "Right", together with the roles of the different participants.

Generally, event extraction allows the definition of different sets of roles for each event. For our use case we do not exploit this flexibility of task formulation as we define the same set of roles for all event types.

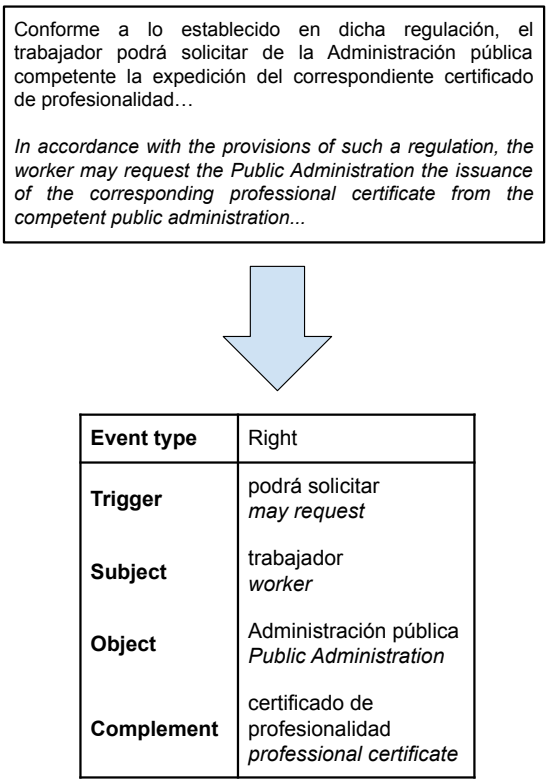


Figure 2: Annotated sample of an event event from a sentence from Spanish labour law, with approximate translations.

3 Training Data

As mentioned before, this experiment is based on the Spanish Workers' Statute, that is published in the Official State Gazette. The text is the main legislative labour law document in Spain, therefore, it clearly corresponds to our aim. The Spanish Workers' Statute is a representative example of a domain-specific legal corpus, therefore, any obtained results could be extrapolated to other legal sub-domains.

The Statute is divided into three main sections named as "titles". The first title covers individual labour relations; the second title covers the rights of collective representation and workers' assemblies inside companies, and the third title covers collective bargaining and collective agreements. In total, the three sections gather 92 articles, containing approximately 50.000 tokens. With the current state of analysis we estimate the density of relations in the Spanish labour law to be 3.65 relations per article.

Regarding the manual annotation of this document, it started by identifying its most relevant terms. To speed up this process, we made use of an open source terminology extraction tool that extracts the most frequent terms in the statute. For more information about the evaluation of the tool's performance, we refer the reader to its research paper (Oliver and Vàzquez, 2015).From those most frequent terms, we identified those that could hold the roles of subject and object of a Hohfeld relation. This is, *legal agents*, such as worker or employer, and *legal entities*, such as company or worker union.

Having the document automatically annotated with these entities, we focused on discovering Hohfeld relations amongst them and extracted the corresponding text excerpts. Optionally, these excerpts could also include a complement, usually an object that takes part in the relation. Not only positive samples were annotated, but also negative samples: text excerpts with legal entities that do not present any relation at all amongst them. Corpus and annotated data statistics are shown in Table 1. Figure 2 shows an example of a positive annotation. Regarding the negative ones, we have identified 2 types: 1) Annotations with entities but no relations and 2) Annotations with neither entities nor relations. Examples of these types are shown below:

1. Mediante los convenios colectivos, y en su ámbito correspondiente, los trabajadores(e1) y empresarios(e2) regulan las condiciones de trabajo y de productividad. (*By means of the collective agreements, in their corresponding field, workers(e1) and employers(e2) regulate working and productivity conditions.*)

2. Igualdad de remuneración por razón de sexo. (*Equaly pay based on sex*)

Type of Element	Total number
Sentences in the corpus	1568
Tokens in the corpus	54849
Annotated samples	133
Positive samples	97
Negative samples	36
Legal agents	127
Legal entities	86
Subjects	90
Objects	69
Complements	100

Table 1: Statistics of the corpus and the annotated data.

4 Models

For this work we focused on joint multi-task deep learning classification models as they achieve state of the art results on common event extraction datasets, see also Section 8. As these common datasets are in English, for the final choice of the model it was important for us that the model code is publicly available, so that we can reuse the model and that the base model is either multilingual or can be changed to multilingual. Finally, we proceed with two models: GRIT (Du, Rush, and Cardie, 2021) and Text2Event (Lu et al., 2021). For both models

GRIT The model GRIT is a generative role-filler transformer model, i.e. it is capable of identifying the (predefined) roles of entities in text. In order to apply GRIT to event extraction task we found it necessary to declare the trigger as a separate role and extend the model to also classify the class (event type) of the input text. The extended model is a joint generative event extraction model. The base model of GRIT is BERT (Devlin et al., 2018), we used the model with BETO (Cañete et al., 2020) that is a BERT model trained on a big Spanish corpus.

Text2Event The event extraction model is Text2Event. This model relies on the description or names of roles, more precisely the names of roles are input to the language model together with the legal text. Therefore, we translated the names of the roles and also the names of the Hohfeld classes to Spanish. As recommended by the authors of the original paper, we have pre-trained the model on ACE dataset, and only then fine-tuned the model on our dataset[6]. The base model of Text2Event is T5 (Raffel et al., 2020) pre-trained on English corpora, we used Text2Event with multilingual T5 (mT5) (Xue et al., 2020).

Roles The final set of roles for both models consists of trigger, subject, object, and complement roles as described in Section 2.

5 Results and Evaluation

For the evaluation of the performance of our model we will use well established metrics such as precision (P), recall (R) and F_1 score. Let the *gold standard* be the correct manually annotated data. Let the *true positives (TP)* be all the correctly predicted relations; *false positives (FP)* – incorrectly predicted relations; *false negatives (FN)* – those cases when a relation is not predicted, though it does exist in the gold standard; *true negatives (TN)* – the relation is not predicted and it does not exist in the gold standard. Then $P = \frac{TP}{TP+FP}$, $R = \frac{TP}{TP+FN}$ and $F_1 = 2 * \frac{P*R}{P+R}$. These measures are well established and widely used for evaluation of different classification models, see also Section 8. It should be noted that we use strict scores, i.e. only exactly correct matches are counted as true predictions.

For computing the results we used the 126 samples from the training set in the following split: 116 samples used for training and 10 samples used for development set. The test set consists of 20 samples, all the results are reported for the test set. This setup is applied to both models. The training was performed with default parameters as set by the authors of the original models, except for the extension of GRIT and the change of the base models as described in Section 4.

The comparison of the models is in Table 2. The two columns present the F_1 scores for the task of classifying all roles including "trigger" role and classifying the Hohfeld class, respectively. The Text2Event model outperforms the GRIT model by a significant margin. Manual checks of the results confirm this finding. A possible explanation of the difference in performance could by the ability of the Text2Event model to include the

[6]Though the roles and the language in the ACE dataset are quite different from our use case, we use ACE in the pre-training to enable the model to learn the constrained generative language as suggested by the authors of the original Text2Event paper.

model name	F_1 roles	F_1 Hohfeld class
GRIT	0.26	0.65
Text2Event	0.47	0.82

Table 2: Comparison of F_1 scores achieved by GRIT and Text2Event models. "F_1 roles" is the F_1 score of extracting all the different roles, including the trigger. "F_1 Hohfeld class" is the F_1 of the legal relation classification.

data	F_1	p	r
all roles	0.47	0.45	0.50
↳ trigger	0.82	0.75	0.90
↳ subject	0.57	0.50	0.67
↳ object	0.00	0.00	0.00
↳ complement	0.45	0.42	0.50
Hohfeld class	0.82	0.75	0.90

Table 3: Detailed scores of Text2Event model.

names of roles and classes, i.e. "derecho", "sujeto", etc., as input to the model. This ability allows for pre-training on the large ACE dataset and efficient knowledge transfer for the few-shot learning with our labour law dataset even despite the different types of events, different domain and different language (English) of the pre-training dataset.

More detailed results for the better performing Text2Event model are presented in Table 3. The model can classify and extract triggers with good confidence, however, the model experiences difficulties with other roles, in particular with identifying objects. Manual investigation of the final results reveals additional problems in identifying the *complements*. Moreover, model rarely predicts other classes than "Right" and "Duty". Nevertheless, the Hohfeld classes, subjects, objects, and triggers are predicted with F_1 scores in the range of 0.5-0.8, which we consider to be reasonably good.

6 Triplification

The second part of our work is focused on the publication of the results obtained following Semantic Web formats. Normally, the results of this type of experiments are usually published in unstructured formats, such as txt, or semi-structured, such as csv.

In this case, we consider that it is highly important to publish these data in structured, open and machine-readable formats, to support their reuse and update. This is possible thanks to the data models of the Semantic Web. In this specific case, we have combined linguistic data representation models with legal information representation models, that are described in Section 6.1. Consequently, we have transformed our results following those vocabularies, as explained in Section 6.2, and the resulting dataset is presented in Section 6.3.

6.1 Ontology selection

As mentioned in the Motivation (Section 1.2), the results of this experiment are to be transformed into a labour law knowledge graph, with the aim of generating a rich resource of concepts and relations that can be applied to other NLP tasks in the future. Therefore, to represent these relations, we chose the Provision Model mentioned in Section 8 since, although LegalRuleML also allows the representation of deontic operators, the Provision Model contains classes corresponding to the potestative relations in case we would like to include them in the future.

On the other hand, to represent the linguistic information apply SKOS[7] and labelling properties from RDF Schema[8].

In this case, since the envisioned output is a rich terminological resource with many different kind of data, we need both vocabularies to represent the complexity of the information contained. Additionally, we combine them with other ontologies such as the Schema data models[9], to add extra information to the relation, as explained in the following section.

6.2 Schema

The heuristics behind the semantic representation of this dataset are as follows:

- Every time we find a Hohfeld relation, we create a new Hohfeld class with the Provision Model, depending on the nature of the relation. Therefore, this would be `prv:Right`, `prv:Duty`, `prv:Prohibition` or `prv:Permission`.
- Following the nomenclature of the Provision Model, every class needs to have a *Bearer*, a *Counterpart* and, optionally, an *Object*. These elements correspond

[7] `https://www.w3.org/TR/skos-reference/`
[8] `https://www.w3.org/TR/rdf-schema/`
[9] `https://schema.org/`

to the *Subject*, *Object* and *Complement* from our annotations, respectively. These items are represented with the class `skos:Concept`, and they are linked to the Hohfeld class with different properties: `prv:hasRightBearer`, `prv:hasRightCounterpart` and `prv:hasRightObject`.

- The `skos:Concepts` are thought to be URIs, to assign unambiguous identifiers to each Hohfeld element. To represent their labels, we use the `rdfs:label` property.
- Additionally, we also include the relation trigger in this data model, that is represented with the class schema:Action, and linked to the Hohfeld class with the properties `prv:hasRightAction`, `prv:hasDutyAction`, `prv:hasProhibitionAction` or `prv:hasPermissionAction`.

6.3 Dataset in RDF

First, we have split the Statute into individual sentences, that were used one by one as input to our fine-tuned Text2Event model. Then we recorded the results. For each sentence that is classified into one of Hohfeld classes we created a subgraph as described in Section 6.2. The statistics of the resulting dataset are presented in Table 4.

Type of Element	Total number
Hohfeld classes	791
Right	578
Duty	213
Subjects	659
Objects	31
Complements	312

Table 4: Statistics of the resulting dataset.

7 Exploitation

In this section, we translate the questions in natural language formulated at the end of Section 1.2 into SPARQL queries[10], to exemplify how to navigate through the generated graph.

First, in Listing 1, we collect all the prefixes that are needed to formulate the rest of the queries. These prefixes identify the vocabularies that have been used to structure the data: RDF and RDF Schema, Provision Model, Schema and SKOS.

Now, Listing 2 formalises Question 1, as stated in Section 1.2: *In which situations can a worker claim for a right?*. Therefore, we look for something (?s) that is a right (prv:Right), which has a right bearer (prv:hasRightBearer), whose ID we do not know (?bearer), and that has right action (prv:hasRightAction), whose ID we do not know either (?action). However, we do know their labels (rdfs:label), which are *trabajador* in Spanish (@es) for the bearer, and *podrá reclamar* in Spanish (@es) for the action. The result of this query are three URIs which are the identifiers of the right instances that satisfy these criteria: http://www.testuri.com/test_hohfeld#1032; http://www.testuri.com/test_hohfeld#762; http://www.testuri.com/test_hohfeld#764.

On the other hand, Listing 3 formalises the Question 2: *When must a worker come back to work after a leave from work?*. In this case, we look for something (?s) that is a duty (prv:Duty), with a duty bearer (prv:hasDutyBearer) and a duty action, which is the trigger (prv:hasDutyAction). Here, the label of the duty bearer is *trabajador*, and the label of the duty action, which is of the type schema:Action, is *deberá reincorporarse*. The result of this query is a URI which is the identifier of the right instance that satisfies this criteria: http://www.testuri.com/test_hohfeld#1051

Additionally, since we already have obtained the ID of a given instance, we could make a simple query to retrieve the textual excerpt (with the property skos:note) from which the relation has been extracted, as exposed in Listing 4. The result of this query is the following excerpt: *En los supuestos de suspensión por ejercicio de cargo público representativo o funciones sindicales de ámbito provincial o superior, el trabajador deberá reincorporarse en el plazo máximo de treinta días naturales a partir de la cesación en el cargo o función.*

8 Related Work

Event Extraction In this work we focus on sentence-level event extraction and consider document-level extraction for future work. Event extraction task has recently received widespread attention (Liu, Min, and Huang, 2021). Most work in event extraction

[10] `https://www.w3.org/TR/rdf-sparql-query/`

Model name	Trigger F_1	Roles F_1
OneIE	72.8	54.8
Text2Event	71.8	54.4

Table 5: State of the art results on ACE dataset.

has focused on the ACE sentence level event task (Walker et al., 2006), which requires the detection of an event trigger and extraction of its arguments from within a single sentence. This dataset consists of 599 documents and includes 8 event types and 6000 individual events. Further important dataset is MUC-4 (muc, 1992) with 1700 documents, 400 tokens per document on average. Note that these datasets are significantly larger than the one we consider in this paper. Most existing event extraction dataset are available in English language, no event extraction dataset in Spanish is known to us.

The most prominent approaches to solving the task include

1. Decomposition into subtasks such as entity recognition and argument classification (Ma et al., 2020; Zhang et al., 2020);
2. Semantic grounding, i.e. mapping entities to external knwoledge sources (Zhang, Wang, and Roth, 2020; Huang et al., 2018);
3. Question-answering based approaches (Zhou et al., 2021; Liu et al., 2020);
4. Joint multi-task classification models (Lin et al., 2020; Paolini et al., 2021; Du, Rush, and Cardie, 2021; Lu et al., 2021).

Recent state of the art results are achieved by the joint multi-task deep learning models. GRIT (Du, Rush, and Cardie, 2021) achieves joint (for classifying all roles) F_1 score of 54.5 on MUC-4 dataset. The results on ACE dataset are collected in Table 5.

Legal Ontologies Regarding the representation of legal information in Semantic Web formats, we find many approaches depending on the type of data and on the purpose of the ontology. Likewise, the type of data depends on the legal subarea to which a resource belongs (labour law, civil law, etc.) and on the type of legal document (provisions, rules, licenses...). The purpose of the ontology is also varied. Some are merely intended to structure the information while others are designed to reason over data and infer knowledge.

Amongst the most used ontologies to structure legal documents we find the ELI ontology[11], the European Legislation Identifier. This vocabulary is widely used by the publishers of legal data in the European Union to represent metadata of legislative documents as Linked Data. To complement ELI, the Publications Office of the European Union also applies the Common Data Model (CDM) vocabulary and the Functional Requirements for Bibliographic Records (FRBR) to represent legal resources and their relationships.

Apart from those ontologies intended to represent legal documents, we also find well-known vocabularies to represent common general terms in the legal domain such as Akoma Ntoso, which was created as an XML standard and afterwards evolved to an ontology (Palmirani and Vitali, 2011), and LegalRuleML (Athan et al., 2015), that is able to represent the particularities of the legal normative rules with a rich, articulated, and meaningful markup language. Similarly, we find the Provision Model (Biagioli, 1996), to annotate rules and rule ammendments in normative provisions, that was subsequently extended in (Francesconi, 2016), to cover Hohfeld's relations (which are described in Section 2).

In this section, we have mentioned those ontologies that are directly related to our work. For more information about legal ontologies, we refer the reader to more comprehensive surveys such as (Valente, 2005) and (de Oliveira Rodrigues et al., 2019).

9 Conclusions and Future Work

In this paper we experimented with pretrained multilingual language models for extracting knowledge from a domain-specific labour law corpus in Spanish. We formalised the task as *sentence-level event extraction* and applied two models: GRIT and Text2Event. To train and evaluate the model, we annotated the Workers' Statute with 133 individual sentences containing Hohfeld roles and relations. The latter model (Text2Event) outperforms the former by a significant margin of around 0.2 of F_1 scores both on identifying the roles and classifying

[11] `https://op.europa.eu/es/web/eu-vocabularies/eli`

into Hohfeld classes. Text2Event obtains a satisfying results above 0.8 F_1 score for classifying Hohfeld classes. The model also efficiently extracts the triggers ($F_1 \approx 0.75$), but looses the quality for other roles ($F_1 \approx 0.5$ for all roles including trigger).

Furthermore, we split the Workers' Statute into individual sentences and apply the fine-tuned Text2Event model on this input. The results of this automatic information extraction tasks have been also triplified, following existing Semantic Web vocabularies, such as SKOS for concepts and the Provision Model for Hohfeld relations. The resulting labour law knowledge graph is publicly available for to be reused by the community.

Future Work In the short term, we plan to develop a post-processing script based on NLP rules to clean the results with the aim of improving precision. This idea is based on the observation that the current models sometimes interchange relation agents with the relation complement, which we think that could be avoided with a role labeling task over the results. Furthermore, we focus on extracting sentence-level roles and relations as most relations are expressed in a single sentence. However, we note that the number of relations spanning over multiple sentences is not negligible. Hence, in the future work we also plan to experiment with extracting relations beyond sentence level.

Regarding the representation in RDF, we plan to add more linguistic information to the graph, linking it to existing legal resources published in Semantic Web formats, such as EuroVoc[12]. We may also want to link our labour law graph with more general resources such as Wikidata[13], to extend the graph with information from a wider scope. To represent this additional linguistic data we plan to use Ontolex[14] to complement SKOS. This combination is widely applied to represent language resources in the Semantic Web: while SKOS is used for thesauri and concept schemes, Ontolex is intended to represent lexical information such as dictionaries.

Figure 3 shows an example of the semantic representation of a right relation amongst the subject "trabajador" (*worker*), the object "administración pública" (*public administration*) and the complement "*certificado de profesionalidad*" (professional certification), being these labels represented with the Ontolex vocabulary. In the example, terms are also linked with matches in EuroVoc and Wikidata.

[12] https://eur-lex.europa.eu/browse/eurovoc.html

[13] https://www.wikidata.org/

[14] https://www.w3.org/2016/05/ontolex/

Acknowledgements

This work has been supported by the European Union's Horizon 2020 research and innovation programme through the Prêt-à-LLOD[15] project, with grant agreement No. 825182, and by COST (European Cooperation in Science and Technology) through NexusLinguarum, the "European network for Web-centred linguistic data science" COST Action (CA18209)[16].

[15] https://pret-a-llod.eu/

[16] https://nexuslinguarum.eu/

References

1992. *Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Ahn, D. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia, July. Association for Computational Linguistics.

Athan, T., G. Governatori, M. Palmirani, A. Paschke, and A. Wyner. 2015. Legalruleml: Design principles and foundations. In *Reasoning Web International Summer School*, pages 151–188. Springer.

Berners-Lee, T. 2006. Design issues.

Berners-Lee, T., J. Hendler, and O. Lassila. 2001. The semantic web. *Scientific american*, 284(5):34–43.

Biagioli, C. 1996. Law making environment: model based system for the formulation, research and diagnosis of legislation. *Artificial Intelligence and Law*.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Chandrasekaran, B., J. R. Josephson, and V. R. Benjamins. 1999. What are ontologies, and why do we need them? *IEEE Intelligent Systems and their applications*.

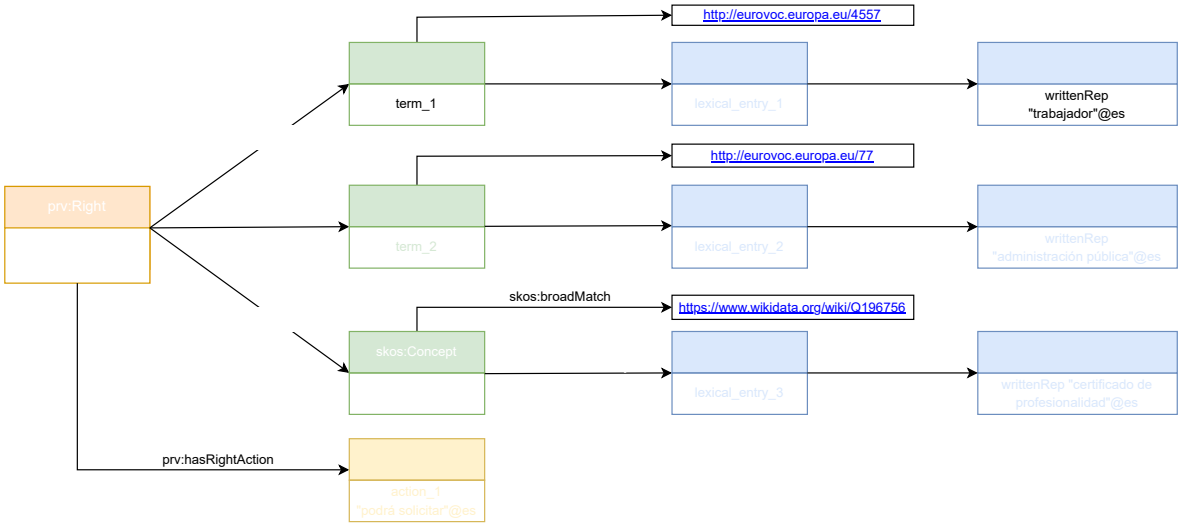


Figure 3: Example of a right modeled with SKOS and Ontolex, and linked with external knowledge bases.

de Oliveira Rodrigues, C. M., F. L. G. de Freitas, E. F. S. Barreiros, R. R. de Azevedo, and A. T. de Almeida Filho. 2019. Legal ontologies over time: A systematic mapping study. *Expert Systems with Applications*, 130:12–30.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, page arXiv:1810.04805, October.

Doddington, G., A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Du, X., A. Rush, and C. Cardie. 2021. GRIT: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 634–644, Online, April. Association for Computational Linguistics.

Francesconi, E. 2016. Semantic model for legal resources: Annotation and reasoning over normative provisions. *Semantic Web*, 7(3):255–265.

Hendrickx, I., S. N. Kim, Z. Kozareva, P. Nakov, D. Ó. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz. 2019. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations Between Pairs of Nominals. *arXiv e-prints*, page arXiv:1911.10422, November.

Hohfeld, W. N. 1913. Some fundamental legal conceptions as applied in judicial reasoning. *Yale Lj*, 23:16.

Huang, L., H. Ji, K. Cho, I. Dagan, S. Riedel, and C. Voss. 2018. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170, Melbourne, Australia, July. Association for Computational Linguistics.

Lin, Y., H. Ji, F. Huang, and L. Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online, July. Association for Computational Linguistics.

Liu, J., Y. Chen, K. Liu, W. Bi, and X. Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online, November. Association for Computational Linguistics.

Liu, J., L. Min, and X. Huang. 2021. An overview of event extraction and its applications. *arXiv e-prints*, page arXiv:2111.03212, November.

Lu, Y., H. Lin, J. Xu, X. Han, J. Tang, A. Li, L. Sun, M. Liao, and S. Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online, August. Association for Computational Linguistics.

Ma, J., S. Wang, R. Anubhai, M. Ballesteros, and Y. Al-Onaizan. 2020. Resource-enhanced neural model for event argument extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3554–3559, Online, November. Association for Computational Linguistics.

Martín-Chozas, P. and A. Revenko. 2021. Thesaurus enhanced extraction of hohfeld's relations from spanish labour law. In *Proceedings of the 2nd International Workshop on Deep Learning meets Ontologies and Natural Language Processing (DeepOntoNLP 2021) co-located with 18th Extended Semantic Web Conference 2021*, volume 2918, pages 30–38. CEUR-WS.org.

Oliver, A. and M. Vàzquez. 2015. Tbxtools: a free, fast and flexible tool for automatic terminology extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*.

Palmirani, M. and F. Vitali. 2011. Akomantoso for legal documents. In *Legislative XML for the semantic Web*. Springer, pages 75–100.

Paolini, G., B. Athiwaratkun, J. Krone, J. Ma, A. Achille, R. ANUBHAI, C. N. dos Santos, B. Xiang, and S. Soatto. 2021. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Robaldo, L., S. Villata, A. Wyner, and M. Grabmair. 2019. Introduction for artificial intelligence and law: special issue "natural language processing for legal texts".

Valente, A. 2005. Types and roles of legal ontologies. In *Law and the semantic web*. Springer, pages 65–76.

Walker, C., S. Strassel, J. Medero, and K. Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Xue, L., N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv e-prints*, page arXiv:2010.11934, October.

Zhang, H., H. Wang, and D. Roth. 2020. Unsupervised Label-aware Event Trigger and Argument Classification. *arXiv e-prints*, page arXiv:2012.15243, December.

Zhang, Z., X. Kong, Z. Liu, X. Ma, and E. Hovy. 2020. A two-step approach for implicit event argument detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7479–7485, Online, July. Association for Computational Linguistics.

Zhou, Y., Y. Chen, J. Zhao, Y. Wu, J. Xu, and J. Li. 2021. What the role is vs. what plays the role: Semi-supervised event argument extraction via dual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14638–14646, May.

```
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX prv:<http://www.ittig.cnr.it/ontologies/def/ProvisionModel#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema:<https://schema.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
```

Listing 1: Prefixes used in the SPARQL queries over the resulting dataset.

```
SELECT ?s
WHERE {
    ?s rdf:type prv:Right;
        prv:hasRightBearer ?bearer ;
        prv:hasRightAction ?action .
    ?bearer  rdfs:label "trabajador"@es .
    ?action rdfs:label "podrá reclamar"@es.
}
```

Listing 2: Example of Question 1 translation into SPARQL.

```
SELECT ?s
WHERE {
    ?s rdf:type prv:Duty;
        prv:hasDutyBearer ?bearer ;
        prv:hasDutyAction ?action .
    ?bearer  rdfs:label "trabajador"@es .
    ?action rdf:type schema:Action ;
    rdfs:label "deberá reincorporarse"@es . }
```

Listing 3: Example of Question 2 translation into SPARQL.

```
SELECT *
WHERE {
    <http://www.testuri.com/test_hohfeld#1051> skos:note ?note .
    }
```

Listing 4: SPARQL query to retrieve the textual excerpt of a given duty.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 117-127 recibido 31-03-2022 revisado 21-05-2022 aceptado 23-05-2022

A Semantic-Proximity Term-Weighting Scheme for Aspect Category Detection

Ponderación de Términos basada en Proximidad Semántica para la Detección de Categorías de Aspecto

Monserrat Vázquez-Hernández, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, México
{mvazquez, villasen, mmontesg}@inaoep.mx

Abstract: Aspect category detection is a subtask of aspect-level sentiment analysis, which aims at identifying the aspect categories present in an opinion. It is a difficult task because the category must be inferred from the terms of the opinion, and also because each opinion may include judgments for more than one aspect category. In recent years, the use of attention mechanisms has improved performance in different tasks, allowing the identification and prioritization of terms that mostly contribute to the classification. However, in multi-label problems, such as aspect category detection, different terms must be selected based on each category, which is a drawback for these models. Motivated by the same idea of identifying and highlighting the importance of terms, this paper proposes a weighing scheme that emphasizes terms in an opinion based on their *semantic proximity* to each aspect category. The proposed scheme has been evaluated on different SemEval datasets, demonstrating its effectiveness in this multi-label scenario. Moreover, it can be applied in scenarios with limited training data and can be combined with different classification models, including deep neural networks.
Keywords: semantic proximity, weighting of terms, aspect category detection.

Resumen: La detección de categorías de aspecto es una subtarea dentro del análisis de sentimientos a nivel de aspecto. Esta subtarea aborda la identificación de aquellas categorías de aspecto presentes en una opinión. Se trata de una tarea desafiante pues la categoría debe inferirse de los términos de la opinión, aunado a esto, una opinión puede incluir evaluaciones de más de una categoría de aspecto. En los últimos años, el uso de mecanismos de atención ha permitido mejorar los resultados en distintas tareas, éstos permiten identificar y priorizar los términos clave que contribuyen a la clasificación. Sin embargo, en problemas multi-etiqueta, como la detección de categorías de aspecto, se deben seleccionar diferentes términos dependiendo de cada categoría lo cual es un inconveniente para estos modelos. Motivados por esta misma idea de identificar y destacar la importancia de términos clave, en este trabajo se propone un esquema que permite enfatizar los términos de una opinión en función de su *proximidad semántica* a cada categoría de aspecto. El esquema propuesto se evaluó en distintos conjuntos de datos de SemEval demostrando su efectividad en este escenario multi-etiqueta. Además, es posible aplicarlo a pesar de contar con pocos datos de entrenamiento, y puede combinarse con distintos modelos de clasificación, incluyendo redes neuronales profundas.
Palabras clave: proximidad semántica, ponderación de términos, detección de categorías de aspecto.

ISSN 1135-5948. DOI 10.26342/2022-69-10

1 Introduction

Sentiment Analysis aims to identify emotions, attitudes or opinions in a subjective text about a product, service or topic of interest (Liu and Zhang, 2012). Different consumers may have different opinions about the same product or service. Through their opinions, consumers express their approval or rejection on particular aspects that they wish to highlight, which poses the challenge of grouping the opinions into different predefined aspect categories in order to identify relevant groups (López Ramos and Arco García, 2019). This challenge is tackled by the sub-task of Aspect Category Detection.

Aspect Category Detection (ACD) attempts to identify the general concepts to which each of the different aspects named in an opinion belong (Pontiki et al., 2016). For example, given the opinion: "the spaghetti was tasteless but the staff was nice", the aspects named are "spaghetti" and "staff", and the corresponding categories are "food" and "service". ACD is a multi-label problem since more than one aspect can be evaluated in an opinion and each one corresponds to a specific category.

Identifying the terms associated with the different categories is a difficult task because the category must be inferred from the context. One possibility for this is to observe different modifiers in an opinion. Through the nature of each category, it is possible to infer the associated terms. For example, in a restaurant, the term "tasteless" is used to describe the "food" but not the "staff". In this context, a common approach used to address this subtask is the use of lexicons (Mowlaei, Saniee Abadeh, and Keshavarz, 2020). Methods based on this approach perform category detection using sets of words to identify the corresponding categories. However, the construction of these lexicons is difficult, expensive, domain and language dependent.

Recently, deep learning approaches using attention mechanisms have been applied to address this task. These mechanisms examine the context of a sentence and identify and prioritize the most relevant terms for its classification. In same way, they are like lexicon-based approaches in that they emphasize the most relevant terms associated with a category, except that these approaches do so automatically without relying on external resources (Chaudhari et al., 2021). Unfortunately, these approaches have some drawbacks for this application. On the one hand, this is a multi-label problem, so the terms related to all different aspect categories mentioned in a single text must be jointly identified (Movahedi et al., 2019). On the other hand, like any deep learning approach, they require large training sets to achieve good results (Chaudhari et al., 2021), and for this subtask datasets are usually limited and also highly imbalanced.

Similar to previous works, in this paper we propose a new term weighting scheme whose aim is also to identify and prioritize terms associated with each aspect category. Based on a given set of category-oriented lexicons, which may have been manually or automatically defined, the proposed scheme weights each term in an opinion according to its semantic proximity to the different aspect categories. To do this, it considers pre-trained word embeddings; in a first step it computes a representative vector for each category, and then, in a second step, it measures the similarity of each term vector with respect to each category vector. Accordingly, the terms that contribute the most to identify each category are highlighted before feeding the classification algorithm, acting as a kind of non-supervised pre-attention mechanism. In this manner, the solution proposed can deal with multi-label problems, is less sensitive to data scarcity and distribution, and can be combined with different classification models, including neural networks.

The evaluation of the proposed approach was carried out on datasets from SemEval (Pontiki et al., 2016), considering English and Spanish languages, two different application domains, as well as several works for comparison purposes.

Summarizing, the two main contributions of this paper are:

- A new term weighting scheme specially suited to the aspect category detection task, which acts as a kind of non-supervised attention mechanism.
- A detailed study on the effectiveness and adaptability of the proposed weighting scheme, considering different languages, domains and classification models.

The remainder of the paper is organized as

follows. Section 2 presents a brief overview of previous work on aspect category detection. Section 3 describes the proposed weighting scheme. Sections 4 reports the experiments, results, and their analysis. Finally, Section 5 points out our conclusions and future work.

2 Related work

Aspect category detection is a subtask of aspect-based sentiment analysis, which attempts to assign a subset of categories from a set of predefined aspect categories to a given opinion (López Ramos and Arco García, 2019). This subtask was introduced and defined at the SemEval workshop: "An aspect category expresses, in a general way, the characteristics evaluated of an entity. Aspect categories are usually not defined by terms present in opinions instead they are inferred through terms used to evaluate different aspects. Category detection is a challenging problem due to the existence of overlapping categories" (Pontiki et al., 2016).

Previous research works in this topic can be organized according to their classification strategy in: lexicon-based, unsupervised, supervised and hybrid (Liu and Zhang, 2012). Despite the existence of successful lexicon-based and unsupervised methods (Ghadery et al., 2018), the majority of the proposed approaches follow a supervised learning approach, considering hand-crafted representations and using classification algorithms such as SVM, K-NN, Logistic Regression or ensembles of them (Xenos et al., 2016), (Hercig et al., 2016), (Hetal and others, 2021).

Over the last few years, deep learning models have brought significant advances to the aspect category detection task. For example, in (Toh and Su, 2016) it is described the construction of a set of binary classifiers, one for each category, considering a variety of lexical and syntactic features, along with extra features learned from a Convolutional Neural Network (CNN). This approach was the best performer at SemEval 2016, and based on the analysis of its results, their authors concluded that the CNN output probabilities were the most relevant features, and that the combination of two different machine learning methods is a feasible approach for the task. In (Xue et al., 2017) the multi-label aspect classification was also handled by multiple one-vs-all binary classifiers, implemented through a neural network with BiLSTM and CNN layers. In addition, this work models the task as a multi-task learning problem, jointly solving the detection of aspect categories and the extraction of aspect terms. Its results showed an important performance improvement, confirming the synergy between both tasks. In (He et al., 2017) it is presented a deep neural network approach based on an attention mechanism, which was later modified and improved in (Movahedi et al., 2019). In this last work, instead of training several one-vs-all models, the authors proposed a single model, namely Topic-Attention Network, which detects aspect categories of a given review sentence by attending to different parts of the sentence based on different topics. Their results confirmed that a single attention may not be able to provide a good representation for reviews containing multiple aspects, and, therefore, pointed out the relevance of learning to weigh the terms based on the different categories. More recently, and in this direction, in (Zhang et al., 2021) it is presented a multilayer self-attention model to deal with aspect category detection. Particularly, it is a BERT-based multi-self-attention model, which uses multiple attentions to obtain relevant information of the multiple target categories. Despite obtaining competitive results, its authors pointed out the difficulties of the attention model to correctly handle short texts, since they provide very limited contextual information.

Following the previous ideas, our approach seeks to emphasize the opinion terms by weighting them in accordance to their semantic proximity to each of the aspect categories, thus generating as many weights for each term as the number of categories. In this way, the terms that could contribute the most to identifying any given category are highlighted prior to feeding the corresponding binary classifier, acting as a kind of non-supervised attention mechanism.

3 Proposed method

Figure 1 shows the general diagram of the proposed method. It follows a *one-vs-all* approach, which means that it uses as many classifiers as aspect categories in the training set. The main components of our method are: *i*) the weighting of terms, *ii*) the construction of the opinions' representations, and *iii*) the classification process. The following sub-

sections detail the first two components since they represent the core contribution of our work. For the classification process we consider traditional as well as deep leaning models, which are described in Section 4.

3.1 New term weighting scheme

The purpose of the proposed term weighting scheme (named as SP for *semantic proximity*) is to emphasize the contribution of each opinion term for the detection of each aspect category. Accordingly, we calculate as many weights for each term as the number of aspect categories.

Given a set of aspect categories, $\mathbf{C} = \{C_1, C_2, ..., C_n\}$, where each category is represent by a pre-defined lexicon or set of terms, $C_i = \{t_{i1}, ..., t_{in}\}$, and using pre-trained word embeddings from GloVe (Pennington, Socher, and Manning, 2014)[1] to represent each term, with $emb(t)$ indicating the embedding vector of term t, we compute the semantic proximity weight of term t_i for category C_j as follows:

1. Define the representative vector of the category C_i, referred as $emb(C_i)$. This vector is computed as the average of the term vectors of the category lexicon:

$$emb(C_i) = \frac{1}{|C_i|} \sum_{\forall t \in C_i} emb(t) \quad (1)$$

2. Measure the semantic proximity of term t_i with respect to category C_j. This proximity is computed by the cosine similarity between the term and category embeddings:

$$SP_{C_j}(t_i) = \cos(emb(t_i), emb(C_j)) \quad (2)$$

According to this new term weighting scheme, the opinion's words that are strongly related to the lexicon of the aspect category that is under analysis will have a greater weight than those from less related words. Figure 2 shows two opinions along with the semantic proximity weights of their words relative to the ambience-general and food-quality categories, respectively.

[1]The experiments were carried out using pre-trained embeddings on Wikipedia 300d and Twitter 200d. Twitter GloVe embeddings were chosen due to their orientation towards language and text size in social networks, Wikipedia GloVe were used to extend the coverage of the vocabulary. Nonetheless, alternative options could be considered.

3.2 Opinion representations

As we previously mentioned, the proposed weighting scheme, which acts as a kind of unsupervised pre-attention mechanism, can be used in combination with different classification models, including traditional classifiers such as the SVM, as well as deep neural networks like a CNN. In the first case, the SP weights are integrated under the Bag of Words representation, while in the second case these weights are used to alter the embeddings that feed the networks. Both cases of opinion representation are described below.

Representation for a traditional classifier. In this case, opinions are represented using a Bag of Words model. Accordingly, each opinion or document is represented by a vector $d = \langle w_1, w_2, ..., w_m \rangle$, where m is the size of the training vocabulary and w_i indicates the weight of term t_i in the opinion. We propose to define these weights as a combination of the term frequency and the term semantic proximity to the aspect category of interest[2] C_j as follows:

$$w_i = tf(t_i) \times SP_{C_j}(t_i) \quad (3)$$

Representation for a deep neural network. In this case, opinions are represented by the array of their word embeddings. Thus, an opinion or document having k terms will be represented by an array of the form $d = [emb(t_1), emb(t_2), ..., emb(t_k)]$. We propose to alter each of these embeddings by multiplying them by a scalar that indicates the relevance of each term, that is, by the semantic proximity SP of each term to the category of interest C_j. Based on this, the new embedding of a term t_i, denoted as $emb'(t_i)$, is computed as indicated in Formula 4, and the new representation of the opinion d is as indicated in Formula 5.

$$emb'(t_i) = SP_{C_j}(t_i) \times emb(t_i) \quad (4)$$

$$d = [emb'(t_1), emb'(t_2), ..., emb'(t_k)] \quad (5)$$

[2]Please note that we follow a one-vs-all classification approach, and, thus, we have a different classifier for each aspect category.

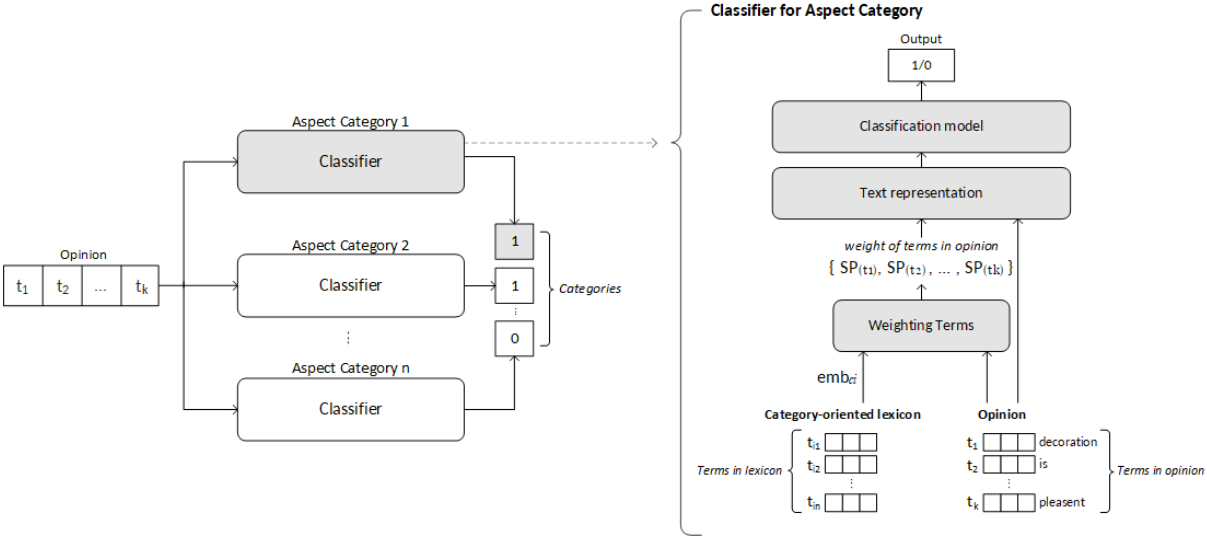


Figure 1: The proposed method. The left part depicts its general view, which is based on a one-vs-all classification approach; the right part details the main components of each aspect category classifier.

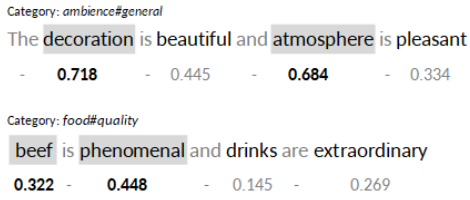


Figure 2: Two opinions with their respective SP weights. Words such as "decoration" and "atmosphere" are highly related to the ambience-general category, whereas "beef" and "phenomenal" to the food-quality category.

4 Experiments

4.1 Datasets

For the experiments we considered two datasets from SemEval 2016 (Pontiki et al., 2016). Particularly, we used the collections for the restaurant domain in both English and Spanish. Table 1 describes the distribution of these collections. As pre-processing operations, we eliminated punctuation marks and special characters; when using the SVM classifier we also removed stopwords.

4.2 Category-Oriented Lexicons

The category-oriented lexicons that our method takes as input can be manually or automatically defined. For the experiments reported, we extracted these lexicons from the training sets using the SS3 method recently proposed in (Burdisso, Errecalde, and Montes-y Gómez, 2019). This method associates each vocabulary term with a confidence value for each of the categories. This value is a number in the interval [0,1] and represents the degree of confidence with which a term is believed to exclusively belong to given category. For example, the term "tequila" will have a confidence value close to 1 for the "drinks" category, whereas the term "of" will have a value equal or close to 0 because it is similarly distributed across all categories.

	instances			
	spanish		**english**	
Categories	*traing*	*test*	*traing*	*test*
ambience#general	293	126	255	66
drinks#quality	31	10	47	22
drinks#style_options	29	11	32	12
drinks#prices	14	10	20	4
food#quality	845	291	849	313
food#style_options	192	69	137	55
food#prices	127	41	90	23
restaurant#general	540	222	422	142
restaurant#miscellaneous	14	13	98	33
restaurant#prices	115	39	80	21
service#general	504	222	449	155
location#general	15	18	28	13

Table 1: Aspect categories in restaurant datasets.

As defined in the SemEval workshop, aspect categories are formed by an entity and attribute pair (e.g., for the category food#quality, the entity is "food" and the attribute is "quality"). In particular, for the restaurant data set there are 12 predefined aspect categories which are listed in Table

1. From these categories, 6 different entities E = {ambience, drinks, food, restaurant, service, location} and 5 different attributes A = {general, miscellaneous, prices, quality, style_options} are derived. We extracted lexicons for each one of the entities and for each one of the attributes, and then we made the corresponding unions to define the lexicons for the 12 aspect categories.

As stated above, the SS3 method (Burdisso, Errecalde, and Montes-y Gómez, 2019) determines a confidence value for all terms with respect to all categories. In order to only include in the lexicons the terms most strongly associated with each category, we propose to filter them using the following criteria: we consider confidences are normally distributed, and thus we keep the terms whose confidence values are equal or above β standard deviations from the mean, considering $\beta = \{1, 2, 3\}$. Table 2 shows five terms for four different lexicons, two in Spanish and two in English.

ambience#general		**food#quality**	
term	*value*	*term*	*value*
ambiente	1.000	calidad	1.000
mesas	1.000	caros	1.000
tranquilo	0.976	platos	0.977
terraza	0.921	excesivos	0.809
bonito	0.846	proporcion	0.749

a) Category-oriented lexicons in Spanish

location#general		**drinks#style_options**	
term	*value*	*term*	*value*
located	1.000	champagne	1.000
chart	0.870	martinis	1.000
block	0.870	well	0.922
sidewalk	0.870	generously	0.922
conveniently	0.870	guaranteed	0.802

b) Category-oriented lexicons in English

Table 2: Example of category-oriented lexicons. For each term its confidence value according to SS3 is indicated.

4.3 Experimental settings

For the experiments, we employed the method proposed in combination with two classification algorithms, a SVM and a CNN. For each classification model, we trained 12 binary classifiers, one for each aspect category. For the sake of simplicity, for all classifiers we used the same settings; the used hyperparameter values are as follows:

- **SVM**: C = 2.5, kernel = linear, and Degree = 2.5. For this particular case, the terms considered for the BOW representation were those present in training set opinions from the category under analysis, without considering empty words[3].
- **CNN**: We used a combination of kernels of sizes 1,2,3 in the convolutional layer to create different feature maps (Gehrmann et al., 2018). The rest of settings for them are: activation=relu, pool-size = max_length - kernel-size + 1, strides=1. The general settings of CNN architecture are: epochs = 9, filters = 256, dropout = 0.5, activation function = sigmoide, loss function = binary cross-entropy, and optimizer = adam.

Due to the fact that the datasets show a high level of imbalance, particularly because the task was approached under the one-vs-all approach, we carried out additional experiments applying oversampling over the minority categories, which indeed correspond to the aspect category under analysis for each binary classifier. In particular, we applied an oversampling technique that consists in randomly replicate instances of the minority class until reaching the size of the majority class.

4.4 Baseline results

As baseline results we used those obtained by the same two classifiers but using the traditional representations. That is, for the SVM we used a BOW with tf-idf weigths, without including our SP weights. In the case of the CNN, we fed it with the pre-trained Glove embeddings without having altered them with our SP weights. We also consider the baseline results reported for each SemEval 2016 task (Pontiki et al., 2016).

Additionally, we compared our results against those from state-of-the-art constrained[4] methods. In particular, we considered the top 3 results for the Spanish and the English datasets. The works considered are:

- *GTI* (Alvarez-López et al., 2016). It uses a support vector machine, but also

[3]Empty words defined in the Python NLTK library are considered.

[4]Works that use external resources such as datasets or dictionaries are classified as *U: Unconstrained* and those that do not use any type of additional resource are classified as *C: Constrained*.

a manually debugged list of words obtained from the training set to remove inter-category noise.

- *TGB* (Çetin et al., 2016). It uses a two-layer approach. In first layer considers a one-vs-all classification approach, where probabilities for entities and attributes are computed. Then, in a second layer, these probabilities are combined to understand which is the best combination of entity and attribute in order to determine the target aspect categories.
- *UWB* (Hercig et al., 2016). it implements a classifier per category using maximum entropy approach. A large number of features are considered for the construction of classifiers.
- *BUTkn* (Macháček, 2016). It uses a set of word n-grams manually compiled for each aspect category and then classifiers the opinions by looking for the occurrence of these n-grams.
- *XRCE* (Brun, Perez, and Roux, 2016). It adapts a component that extracts semantic information about entities and attributes. A dependency graph is created in which the relationships of a term with respect to categories are represented. The classification is performed by looking for word matches in the opinions and their relationships with categories.

4.5 Experimental Results

Table 3 presents a summary of the best results obtained with our method as well as with the baseline configurations. The second and third columns refer to the configuration of the classifier (kind and whether or not oversampling was used), while the fourth and third columns indicate the configuration used to calculate the SP term weights. When comparing the results obtained with and without using the SP weights, the usefulness of the proposed method is clearly appreciated. For the Spanish collection, our best result is achieved with the SVM classifier using oversampling, and with $\beta = 2$ and using the Twitter Glove embeddings of 200 dimensions for the computation of the SP weights. For English, the best results were achieved with the CNN architecture with oversampling, and with $\beta = 2$ and using Wikipedia Glove embeddings of 300 dimensions for the computation of the SP weights.

Regarding the comparison against state-of-the-art methods, Table 4 shows our method's results as well as the best constrained results from SemEval 2016 in both Spanish and English. These comparisons evidence the relevance of the proposed method, since, in addition to its simplicity and generality, it shows competitive results in both scenarios; in particular it outperforms the best results previously reported in the Spanish dataset.

4.5.1 Discussion

In contrast to most previous works, the proposed method does not necessarily need to use external resources for its implementation. For example, the lexicons used by the proposed weighting scheme can be automatically extracted from the training set, as we did in the experiments. For those categories with a considerably reduced number of instances, the number of terms extracted for their lexicons is also reduced. However, our experiments showed that the number of terms per lexicon is not directly related to the results obtained per category. Tables 5 and 6 contrast the number of instances, number of terms in lexicons, and results achieved for each of the aspect categories of the two datasets.

Despite the small number of terms in the lexicons, the proposed weighing scheme demonstrates its usefulness by paying different levels of attention to those terms strongly related to the categories, which helps to determine whether or not a category is present in an opinion. Analyzing the terms that constitute all category-oriented lexicons (refer to Figure 3), we observe that they are clearly representative of the different categories, despite they were automatically extracted. In consequence, an advantage of this approach is that it can be easily adapted to other domains or languages, as was observed in the experiments.

To demonstrate the generality of our method, we performed another experiment using the English laptop dataset. The achieved results are shown in Table 7. In this case, only the best constrained work is taken as a reference for comparison. Our best result was achieved using the SVM classifer, applying oversampling for class balancing, and considering $\beta = 2$ with GloVe Wikipedia em-

Spanish	Classifier	Oversampling	Embeddings	Threshold	micro-F1
Our method	**SVM**	**Yes**	**Twitter 200d**	$\beta = 2$	**71.11**
Baseline	SVM	Yes	-	-	64.30
Our method	CNN	Yes	Twitter 200d	$\beta = 1$	69.99
Baseline	CNN	Yes	Wikipedia 300d	-	60.50
Baseline SemEval	-	-	-	-	54.68

English	Classifier	Oversampling	Embeddings	Threshold	micro-F1
Our method	SVM	Yes	Twitter 200d	$\beta = 2$	66.50
Baseline	SVM	No	-	-	64.30
Our method	**CNN**	**Yes**	**Wikipedia 300d**	$\beta = 2$	**68.50**
Baseline	CNN	Yes	Wikipedia 300d	-	66.80
Baseline SemEval	-	-	-	-	58.92

Table 3: Micro-F1 results of the aspect category detection task, for the restaurant domain in Spanish and English. For each method the result of its best configuration is included; the best overall result is highlighted in bold.

	micro-f1
SP+SVM	**71.111**
GTI	70.027
TGB	63.551
UWB	61.968

a) Spanish

	micro-f1
BUTKn.	71.494
XRCE	68.701
SP+CNN	**68.500**
UWB	67.817

b) English

Table 4: Comparison against SOTA results by constrained methods in the aspect category detection tasks from SemEval 2016. The results of our method correspond to the SP+SVM and SP+CNN configurations for Spanish and English, respectively.

Category	**Training Intances**	**#Terms**	**Test Instances**	**%Errors**
ambience#general	293	40	126	30.73
drinks#quality	31	17	10	90.00
drinks#style_options	29	9	11	45.45
drinks#prices	14	8	10	55.55
food#quality	845	17	291	20.66
food#style_options	192	62	69	58.06
food#prices	127	8	41	20.00
restaurant#general	540	23	222	29.22
restaurant#miscellaneous	14	28	13	100.00
restaurant#prices	115	23	39	52.63
service#general	504	31	222	16.55
location#general	15	47	18	83.33

Table 5: Number of opinions to classify and lexicon terms for each aspect category, as well as the percentage of incorrectly classified instances by our best result in the Spanish dataset.

Category	**Training Instances**	**#Terms**	**Test Instances**	**%Errors**
ambience#general	255	17	66	78.79
drinks#quality	47	10	22	45.45
drinks#style_options	32	9	12	58.33
drinks#prices	20	7	4	75.00
food#quality	849	6	313	19.17
food#style_options	137	6	55	50.91
food#prices	90	6	23	47.29
restaurant#general	422	4	142	42.14
restaurant#miscellaneous	98	4	33	78.79
restaurant#prices	80	10	21	47.62
service#general	449	4	155	21.29
location#general	28	1	13	53.85

Table 6: Number of opinions to classify and lexicon terms for each aspect category, as well as the percentage of incorrectly classified instances by our best result in the English dataset.

beddings of 300 dimensions for the computation of the SP weights. As can be noticed, our result is significantly higher than the reported SemEval's baseline result, and it also outperforms the best reported constrained result. It is important to highlight that, in spite of the large number of categories in this dataset, 67 aspect categories derived from 22 entities and 9 attributes, for the experiments we did not carried out any additional hyperparameter adjustment.

	micro-f1
SP+SVM	**62.10**
UWB	60.45
Baseline SemEval	52.68

Table 7: Micro F1 results in the aspect category detection using the English laptop dataset.

4.5.2 Error analysis

Doing a detailed analysis of the errors (refer to Tables 5 and 6), it can be noticed that

Figure 3: Terms of three category-oriented lexicons of the restaurant domain.

there is no a clear relationship between the characteristics of the datasets and error rates. For the Spanish dataset, the correlation between the number of training instances and the percentage of errors is $r = -0.690$, while for the English dataset $r = -0.664$. Surprisingly, these values indicate that there was a slight tendency to misclassify instances from categories with many examples. This is partly because these more frequent categories are more diverse and also because they usually appeared together with others.

On the other hand, when correlating the size of the categories' lexicons with the classification errors, for Spanish we obtained a $r = 0.214$ and for English $r = 0.355$, suggesting little influence of this variable on the results. However, analyzing these lexicons in greater depth, we found that although they seemed related and relevant to the different categories, they showed a high overlap. In the Spanish dataset, for the smallest categories more than 80% of their terms are also included in others. One interesting example is the entity category "drinks", for which any term was unique; even more, 40% of its terms are also in the lexicons of four or more categories. For English, a similar behavior was observed, but, in addition, we noticed that for the category "location#general", with a single term but exclusive to this category, we obtained a better result than for other categories with larger lexicons. This suggests us the influence of the category-oriented lexicons in whole process, as well as the need to define them more precisely.

5 Conclusions

The main contribution of this work is the proposal of a new term weighting scheme specially suited to the aspect category detection task. It is based on the evaluation of the semantic proximity of each term in an opinion with respect to the categories' description, acting as a kind of non-supervised attention mechanism.

The proposed weighting scheme relies on the availability of a set of category-oriented lexicons, nonetheless, they can be automatically extracted from the training dataset. This latter characteristic makes the method easily adaptable to different domains and languages.

From the results obtained, it is possible to conclude that the proposed term weighting scheme has a positive impact on the identification of the categories of aspects expressed in an opinion. Moreover, it has the advantage of being able to be combined with different classification models, including traditional machine learning classifiers as well as deep neural networks.

On the other hand, although the method is less sensitive to small and imbalanced datasets than other supervised approaches, it is affected by these conditions. As it was observed, the method achieved better results in Spanish than in English, being the latter the collection with less instances and high imbalance rates.

As working directions, we plan to evaluate the method using different kinds of category lexicons, both manually and automatically generated. Besides that, we seek to evaluate our method in collections having different volumes of information, as well as using different contextual word embeddings such as those from BERT. Furthermore, due to the generality of the proposed method, we plan to apply it in other text classification tasks such as polarity classification, author profiling and fraud detection.

Acknowledgments

The present work was supported by CONACyT/México (scholarship 756974 and grant CB-2015-01-257383). In addition, the authors thank CONACYT for the computational resources provided by the Deep Learning Platform for Language Technologies.

References

Alvarez-López, T., J. Juncal-Martínez, M. Fernández-Gavilanes, E. Costa-Montenegro, and F. J. González-Castano. 2016. Gti at semeval-2016 task 5: Svm and crf for aspect detection and unsupervised aspect-based sentiment analysis.

In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 306–311.

Brun, C., J. Perez, and C. Roux. 2016. XRCE at SemEval-2016 task 5: Feedbacked ensemble modeling on syntactico-semantic knowledge for aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 277–281, San Diego, California, June. Association for Computational Linguistics.

Burdisso, S. G., M. Errecalde, and M. Montes-y Gómez. 2019. A text classification framework for simple and effective early depression detection over social media streams. *Expert Systems with Applications*, 133:182–197.

Çetin, F. S., E. Yıldırım, C. Özbey, and G. Eryiğit. 2016. Tgb at semeval-2016 task 5: multi-lingual constraint system for aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 337–341.

Chaudhari, S., V. Mithal, G. Polatkan, and R. Ramanath. 2021. An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–32.

Gehrmann, S., F. Dernoncourt, Y. Li, E. T. Carlson, J. T. Wu, J. Welt, J. Foote Jr, E. T. Moseley, D. W. Grant, P. D. Tyler, et al. 2018. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS one*, 13(2):e0192360.

Ghadery, E., S. Movahedi, H. Faili, and A. Shakery. 2018. An unsupervised approach for aspect category detection using soft cosine similarity measure. *arXiv preprint arXiv:1812.03361*.

He, R., W. S. Lee, H. T. Ng, and D. Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, Vancouver, Canada, July. Association for Computational Linguistics.

Hercig, T., T. Brychcín, L. Svoboda, and M. Konkol. 2016. Uwb at semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 342–349.

Hetal, V. et al. 2021. Ensemble models for aspect category related absa subtasks. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(13):2348–2364.

Liu, B. and L. Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*. Springer, pages 415–463.

López Ramos, D. and L. Arco García. 2019. Aprendizaje profundo para la extracción de aspectos en opiniones textuales. *Revista Cubana de Ciencias Informáticas*, 13(2):105–145.

Macháček, J. 2016. BUTknot at SemEval-2016 task 5: Supervised machine learning with term substitution approach in aspect category detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 301–305, San Diego, California, June. Association for Computational Linguistics.

Movahedi, S., E. Ghadery, H. Faili, and A. Shakery. 2019. Aspect category detection via topic-attention network. *CoRR*, abs/1901.01183.

Mowlaei, M. E., M. Saniee Abadeh, and H. Keshavarz. 2020. Aspect-based sentiment analysis using adaptive aspect-based lexicons. *Expert Systems with Applications*, 148:113234.

Pennington, J., R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pontiki, M., D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *10th International Workshop on Semantic Evaluation (SemEval 2016)*.

Toh, Z. and J. Su. 2016. NLANGP at SemEval-2016 task 5: Improving aspect

based sentiment analysis using neural network features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 282–288, San Diego, California, June. Association for Computational Linguistics.

Xenos, D., P. Theodorakakos, J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos. 2016. Aueb-absa at semeval-2016 task 5: Ensembles of classifiers and embeddings for aspect based sentiment analysis. In **SEMEVAL*.

Xue, W., W. Zhou, T. Li, and Q. Wang. 2017. Mtna: a neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 151–156.

Zhang, X., X. Song, A. Feng, and Z. Gao. 2021. Multi-self-attention for aspect category detection and biomedical multilabel text classification with bert. *Mathematical Problems in Engineering*, 2021.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 129-140 recibido 21-03-2022 revisado 17-05-2022 aceptado 30-05-2022

Detección de Indicios de Autolesiones No Suicidas en Informes Médicos de Psiquiatría Mediante el Análisis del Lenguaje

Detecting Signs of Non-suicidal Self-Injury in Psychiatric Medical Reports Using Language Analysis

Juan Martinez-Romo[1],[2] **Blanca Reneses**[1],[2],[3] **Ignacio Martínez-Capella**
Lourdes Araujo[1],[2] **J. Sevilla-Llewellyn-Jones**[1] **Germán Seara-Aguilar**
[1]NLP & IR Group
Universidad Nacional de Educación a Distancia (UNED)
[2]Instituto Mixto UNED-ISCIII IMIENS
juaner,lurdes@lsi.uned.es

[1]IdISSC
Hospital Clínico San Carlos
[2]CIBERSAM
[3]Universidad Complutense
blanca.reneses@salud.madrid.org
julia.sevilla@salud.madrid.org

Unidad de Innovación
IdISSC
Hospital Clínico San Carlos
Madrid
imcapella@salud.madrid.org
gseara@shealth.eu

Resumen: La autolesión no suicida, a menudo denominada autolesión, es el acto de dañarse deliberadamente el propio cuerpo, como cortarse o quemarse. Normalmente, no pretende ser un intento de suicidio. En este trabajo se presenta un sistema de detección de indicios de autolesiones no suicidas, basado en el análisis del lenguaje, sobre un conjunto anotado de informes médicos obtenidos del servicio de psiquiatría de un Hospital público madrileño. Tanto la explicabilidad como la precisión a la hora de predecir los casos positivos, son los dos principales objetivos de este trabajo. Para lograr este fin se han desarrollado dos sistemas supervisados de diferente naturaleza. Por un lado se ha llevado a cabo un proceso de extracción de diferentes rasgos centrados en el propio mundo de las autolesiones mediante técnicas de procesamiento del lenguaje natural para alimentar posteriormente un clasificador tradicional. Por otro lado, se ha implementado un sistema de aprendizaje profundo basado en varias capas de redes neuronales convolucionales, debido a su gran desempeño en tareas de clasificación de textos. El resultado es el funcionamiento de dos sistemas supervisados con un gran rendimiento, en donde destacamos el sistema basado en un clasificador tradicional debido a su mejor predicción de clases positivas y la mayor facilidad de cara a explicar sus resultados a los profesionales sanitarios.
Palabras clave: Detección de autolesiones no suicidas, análisis del lenguaje, aprendizaje automático, redes neuronales.

Abstract: Non-suicidal self-injury, often referred to as self-injury, is the act of deliberately harming one's own body, such as cutting or burning oneself. It is not usually intended as a suicide attempt. This paper presents a system for detecting signs of non-suicidal self-injury, based on language analysis, on an annotated set of medical reports obtained from the psychiatric service of a public hospital in Madrid. Both explainability and accuracy in predicting positive cases are the two main objectives of this work. In order to achieve this goal, two supervised systems of different natures have been developed. On the one hand, a process of extraction of different features focused on the world of self-injury itself has been carried out using natural language processing techniques to subsequently feed a traditional classifier. On the other hand, a deep learning system based on several layers of convolutional neural networks, due to its high performance in text classification tasks. The result are two supervised systems with high performance, where we highlight the system based on a traditional classifier due to its better prediction of positive classes and the greater ease to explain its results to health professionals.
Keywords: Non-suicidal Self-injury detection, language analysis, machine learning, neural networks.

ISSN 1135-5948. DOI 10.26342/2022-69-11

1 Introducción

Los trastornos de salud mental, como las autolesiones, son problemas cuya incidencia en la población aumenta de manera alarmante en los últimos años. Estas afecciones pueden pasar desapercibidas durante muchos años, lo que hace que las personas que las padecen no reciban la asistencia médica adecuada. Los problemas de salud mental sin tratar pueden acarrear graves consecuencias, como el deterioro personal o incluso el suicidio. Las autolesiones, también conocidas como autolesiones deliberadas o autoagresiones, son un tipo de problema de salud mental menos conocido que afecta principalmente a los jóvenes (Young et al., 2007). La autolesión se refiere al acto de causarse daño corporal a sí mismo sin intención suicida, como cortarse, quemarse o tirarse del pelo, y se ha relacionado con problemas de salud mental subyacentes, como la depresión y la ansiedad (Greaves, 2018b). Entre las diferentes acciones que las personas afectadas llevan a cabo dentro del concepto general de autolesión, existen grandes diferencias tanto en los motivos para llevarlas a cabo, como en el género (Rodham, Hawton, y Evans, 2004). La naturaleza a menudo impulsiva de estos actos (especialmente el auto-corte) significa que la prevención debe centrarse en fomentar métodos alternativos de gestión de la ansiedad, la resolución de problemas y la búsqueda de ayuda antes de que se desarrollen pensamientos de autolesión. Dada la gravedad de los síntomas y los riesgos, es importante dedicar esfuerzos a detectar mejor los problemas de salud mental en la sociedad para que puedan recibir la ayuda que necesitan. También se han encontrado diferencias en la forma de comunicarse y en el lenguaje empleado por las personas que sufren problemas de salud mental (Pennebaker, Mehl, y Niederhoffer, 2003). A pesar de que los informes médicos están generalmente escritos por médicos, en ocasiones tratan de plasmar las ideas subyacentes del paciente e incluso escriben literalmente frases o expresiones utilizadas y que puedan denotar de forma clara estados de ánimo o pensamientos. De esta forma, el análisis del lenguaje empleado en estos informes médicos mediante técnicas de Procesamiento del Lenguaje Natural (PLN) pueden ayudar a la detección temprana de pacientes con otros trastornos previos. Sin embargo, cabe señalar que la mayoría de los estudios sobre detección temprana de peligros para la seguridad y la salud se han centrado en el texto en inglés. Por otra parte, hay que señalar que apenas existen conjuntos de datos (datasets o corpora) para entrenar modelos de identificación en las tareas mencionadas, y los existentes se limitan al inglés y son de tamaño reducido, lo cual es un claro indicador del camino que aún queda por recorrer para que los profesionales sanitarios puedan disponer de herramientas maduras de análisis de textos. En este trabajo se presenta un sistema de detección de autolesiones en informes médicos procedentes del servicio de psiquiatría del Hospital Clínico San Carlos de Madrid. Este sistema de detección de autolesiones entrenado y evaluado sobre un corpus anotado de informes médicos, tiene el objetivo de aplicarse a cualquier informe del servicio de psiquiatría para permitir la detección temprana de este trastorno en pacientes que hayan sido tratados por dicho servicio. De esta forma, pacientes con otro tipo de trastornos de carácter más leve, podrían ser diagnosticados y recibir tratamiento antes de adquirir estos hábitos tan perjudiciales. Y es que la detección temprana es clave en el tratamiento de los problemas de salud mental, ya que una intervención rápida mejora las probabilidades de un buen pronóstico.

El resto del artículo se organiza de la siguiente forma: en la Sección 2 se analiza el estado del arte y trabajos relacionados. en la Sección 3 se describe el corpus y las técnicas utilizada para anotarlo. En la Sección 4 se detallan las características del sistema de detección de autolesiones. La Sección 5 se centra en la experimentación y el análisis de resultados. Finalmente, en la Sección 6 se extraen las principales conclusiones y se exponen las líneas de trabajo futuro.

2 Estado del Arte

El estudio de las autolesiones y más concretamente de su investigación a través del análisis de textos no es demasiado extenso. Existen trabajos (Baetens et al., 2011) en los que se investigaron la prevalencia de las autolesiones no suicidas (NSSI) y las autolesiones suicidas (SSI) en una muestra de adolescentes de entre 12 y 18 años, así como las diferencias psicosociales entre los adolescentes que practican NSSI y los que practican SSI. También hay trabajos (Nicolai, Wielgus, y Mezulis, 2016) que apoyan la teoría de la cascada emocio-

nal, en la que la rumiación distingue entre las personas que se autolesionan y las que no lo hacen, y hacen especial hincapié en la relación entre el afecto negativo y las NSSI.

Hay trabajos (Burke, Ammerman, y Jacobucci, 2019) que se han centrado en abordar las limitaciones de los sistemas de detección de riesgo y el tiempo de cómputo utilizando herramientas analíticas avanzadas, como el procesamiento del lenguaje natural (PLN) y el aprendizaje automático. Existen estudios centrados en la ideación de suicidio que usan enfoques de PLN y que han utilizado en gran medida modelos basados en la historia clínica electrónica (HCE) (Haerian, Salmasian, y Friedman, 2012; Kessler et al., 2017) y modelos de predicción basados en PLN y rasgos lingüísticos (Fernandes et al., 2018; McCoy et al., 2016; Poulin et al., 2014).

En 2017, un trabajo (Walsh, Ribeiro, y Franklin, 2017) utilizó aprendizaje automático para predecir el riesgo de suicidio en pacientes de autolesiones a lo largo del tiempo analizando informes médicos de una gran base de datos médica. También se han usado tests adaptativos informatizados (CAT) para entrenar un árbol de decisión con el objetivo de predecir el riesgo de suicidio (Delgado-Gomez et al., 2016). Otros trabajos (Metzger et al., 2017) han empleado algoritmos de clasificación como random forest and naïve Bayes sobre informes médicos para predecir suicidios, demostrando que los métodos de aprendizaje automático pueden mejorar la calidad de los indicadores epidemiológicos en comparación con la actual vigilancia nacional de los intentos de suicidio de un país como Francia. Los árboles de decisión también se han usado en otros trabajos (Mann et al., 2008) para estudiar la correlación entre pacientes de psiquiatría analizando su conducta suicida pasada frente a la ideación de suicidio que mostraban en un momento dado.

La clasificación de textos clínicos mediante redes neuronales ha resultado una herramienta de gran utilidad en problemas como la identificación de fenotipos en informes médicos para pacientes con un conjunto determinado de signos y síntomas clínicos (Obeid et al., 2019). En los últimos años se han producido avances significativos en los enfoques de aprendizaje profundo, como las redes neuronales convolucionales (CNN), y su aplicación ha sido un éxito en problemas como el procesamiento y la clasificación de textos o el reconocimiento del habla (LeCun, Bengio, y Hinton, 2015).

Recientemente ha surgido un estudio (Obeid et al., 2020) que aprovecha la información de las notas clínicas utilizando redes neuronales profundas (DNNs) para identificar los pacientes tratados por autolesión intencional y predecir futuros eventos de autolesión. Los autores utilizaron dos modelos basados en una CNN y en una LSTM con resultados prometedores. También se han usado técnicas de clasificación como el Gradient Boosting para la detección de autolesiones e ideación suicida en las notas de triaje de los servicios de urgencias (Rozova et al., 2022).

En los últimos años han aparecido bastantes trabajos en el ámbito de las redes sociales y la salud mental. Aunque el formato del texto de las redes sociales y en lenguaje escrito en primera persona hacen abordar este problema desde un enfoque diferente, queremos destacar algunos trabajos por su relevancia y su cercanía al problema de las autolesiones.

Desde 2017 y de forma anual se celebra la tarea competitiva eRisk (Losada, Crestani, y Parapar, 2019; Losada, Crestani, y Parapar, 2020; Parapar et al., 2021), dentro del congreso CLEF (Cross Language Evaluation Forum). eRisk trata de avanzar en la predicción temprana en redes sociales de problemas relacionados con la salud mental. Depresión, anorexia, ludopatía y autolesiones desde el año 2019 han sido los trastornos elegidos por los organizadores. Dentro de esta competición, un sistema con resultados prometedores fue el equipo iLab (Martınez-Castano et al., 2020) en el que los investigadores propusieron un sistema de clasificación basado en BERT y transformers. En contraposición al uso pesado de las redes neuronales y los transformers, también resultan interesantes las participaciones del grupo NLP-UNED (Ageitos, Martínez-Romo, y Araujo, 2020; Campillo-Ageitos et al., 2021), que emplearon técnicas de PLN y análisis de sentimientos para alimentar un clasificador rápido y eficiente. Finalmente, un trabajo que obtuvo buenos resultados con un sistema innovador fue el del grupo UNSL(Loyola et al., 2021), que empleó políticas de alerta, un sistema basado en reglas y un modelo de aprendizaje por refuerzo.

3 Corpus

El corpus de evaluación procede de un conjunto de informes médicos anonimizados pro-

cedentes del servicio de psiquiatría del Hospital Clínico San Carlos de Madrid en España.

La preparación de los informes para su análisis ha sido desarrollada por la Unidad de Innovación del Hospital Clínico San Carlos, a partir de la descarga autorizada de informes informatizados del Servicio de Psiquiatría correspondientes a un periodo de cuatro años. Dicha preparación ha consistido en tres fases: limpieza de los informes, compleción y anonimización. Previamente, esta cesión de datos fue evaluada y aprobada por el Comité de Ética de la Investigación (20/586-E).

A partir de este conjunto de informes anonimizados, se llevó a cabo un proceso de anotación por parte de expertos dando lugar a un corpus de 1252 informes anotados. Los diagnósticos de estos informes son diversos, pero entre ellos no se incluye sufrir autolesiones. Por ello ha sido necesaria una anotación manual supervisada por los médicos expertos en base al contenido textual de los informes. Tras la anotación manual en busca de indicios de autolesiones, 1138 han sido anotados como negativos y 114 como positivos. Durante el proceso de anotación, se buscaban indicios claros de que el profesional sanitario que hubiera atendido al paciente indicara que las autolesiones se habían producido. La mera ideación o pensamiento de esta situación fue tratada como un caso negativo. En el caso de situaciones en las que las autolesiones tenían un fin autolítico también fueron etiquetadas como casos negativos al buscar otro fin diferente al ansiolítico. Estos últimos casos deberían tratarse en un estudio diferente como parte de los pacientes con riesgo de suicidio. Los informes tienen una media de 1310 palabras y 7566 caracteres por cada informe, teniendo el informe de mayor tamaño 1639 palabras y 32767 caracteres y el de menor tamaño 84 palabras y 510 caracteres. El corpus se ha dividido en dos conjuntos de entrenamiento (80 %) y test (20 %), resultando dos conjuntos de 1001 y 251 informes respectivamente. La división se ha llevado a cabo de forma estratificada para respetar la proporción de clases en los conjuntos de entrenamiento y test.

4 Sistema de Detección de Autolesiones

Para la tarea de detección de autolesiones se han desarrollado dos sistemas supervisados, uno de ellos basado en la extracción de rasgos y la aplicación de algoritmos clásicos de clasificación y el otro basado en redes neuronales con la aplicación de un modelo BERT para el tokenizado. Los dos sistemas desarrollados solo analizan el texto anonimizado del informe sin tener en cuenta el diagnóstico que aparece en otro campo y que sólo ha sido tenido en cuenta en el proceso de anotación manual para ayudar a los expertos en caso de duda.

4.1 Sistema de Aprendizaje Automático

El sistema supervisado está compuesto de tres módulos diferentes que se encargan de las tareas de pre-procesamiento, extracción de rasgos y aplicación de algoritmos de clasificación.

4.1.1 Pre-procesamiento de los Informes Médicos

En cuanto al pre-procesamiento se han aplicado las técnicas habituales, como son la conversión a minúsculas del texto, la eliminación de caracteres especiales, la normalización de determinados conectores, el tokenizado del texto y el borrado de palabras vacías. También se ha llevado a cabo un proceso de stemming mediante el segundo algoritmo de Porter para extraer la raíz de las palabras.

4.1.2 Extracción de Rasgos y Algoritmos de Clasificación

La extracción de rasgos se puede agrupar en cinco conjuntos de características:

- **CountVectorizer**: En primer lugar se ha utilizado la herramienta CountVectorizer de la biblioteca scikit-learn en Python para obtener vectores de palabras a partir de los informes médicos. Esta herramienta se utiliza para transformar un texto dado en un vector sobre la base de la frecuencia de cada palabra que aparece en todo el texto. La función crea una matriz en la que cada palabra única está representada por una columna de la matriz, y cada muestra de texto del documento es una fila en dicha matriz. El valor de cada celda no es más que la frecuencia de la palabra en esa muestra de texto en particular.
- **Vocabulario de Autolesiones**: Se ha compilado un conjunto de 53 palabras relacionadas con el contexto de las autolesiones. En este conjunto hay palabras

como morder, cortar, pellizcar, etc. Este conjunto de palabras, se emplea como entrada del CountVectorizer para no usar todo el vocabulario completo sino solo estas 53 palabras, algo que proporciona mayor rapidez y precisión.

- **Diccionario NSSI**: Greaves (Greaves, 2018a) desarrolló un trabajo en el que llevó a cabo la clasificación de un conjunto de conceptos relacionados con las autolesiones. El resultado de este trabajo es un diccionario de palabras relacionadas con la autolesión llamado Diccionario Non-Suicidal Self-Injury (NSSI), donde las palabras se dividen en cinco categorías: 1) Métodos de NSSI; 2) Términos de NSSI; 3) Instrumentos utilizados; 4) Razones de NSSI; y (5) Términos específicos de cortes. De esta forma, se han creado cuatro rasgos de NSSI, uno para cada categoría. Estas características cuentan la frecuencia de las palabras de su categoría en el texto.
- **Distancia de Términos de Autolesiones**: Existen numerosos trabajos que han probado la relevancia de las primeras palabras de un documento en relación al texto completo. En este grupo de rasgos se ha tratado de medir por un lado la distancia entre el inicio del documento y la primera palabra del vocabulario de autolesiones presente en el texto y por otro lado la distancia media entre palabras del vocabulario de autolesiones. Estas medidas se han realizado en función del número de palabras y del número de caracteres, dando lugar a cuatro rasgos.
- **Negación**: Se ha llevado a cabo un proceso de detección de la negación mediante una arquitectura (Fabregat, Araujo Serna, y Martínez Romo, 2019; Fabregat et al., 2019) basada en aprendizaje profundo. La detección de la negación se ha aplicado a los grupos de rasgos definidos anteriormente para eliminar la presencia de los términos de autolesiones que han sido negados y de esta forma restar su incidencia. Es decir, si en el texto aparece una afirmación como "No se aprecian cortes", la detección de negación evita que el término "cortes" se contabilice en ninguno de los rasgos calculados en este trabajo.

Una vez extraídos los rasgos descritos anteriormente y con la ayuda del corpus anotado, se ha llevado a cabo la aplicación de los algoritmos más efectivos según el estado del arte en este tipo tareas.

4.2 Sistema basado en Aprendizaje Profundo

El segundo sistema usa redes neuronales con una arquitectura en la que se disponen tres capas de redes neuronales convolucionales y para la que se ha adaptado la tecnología de BERT (Devlin et al., 2018) para el proceso de tokenizado, que está basado en la representación de codificadores binarios a partir de Transformers. En este caso, hemos adaptado esta tecnología para la clasificación de textos.

Aparte de la preparación del texto, para el tokenizado de los textos médicos, hemos usado dos modelos pre-entrenados: Un modelo base de BERT[1] que está disponible en seis idiomas, incluido el español, y fue creado para tareas de clasificación de textos. Y el modelo RoBERTa-base-bne[2], que es un modelo de lenguaje enmascarado basado en transformers para el español. Está basado en el modelo base de RoBERTa y ha sido pre-entrenado utilizando el mayor corpus en español conocido hasta la fecha, con un total de 570GB de texto limpio y procesado expresamente para este trabajo. El texto procede de una compilación de páginas web realizada por la Biblioteca Nacional Española desde 2009 hasta 2019.

De esta forma, se ha adoptado una arquitectura que consta de tres capas de redes neuronales convolucionales concatenadas. La arquitectura del sistema de aprendizaje profundo usada para este trabajo puede apreciarse en la Figura 1, en la que se muestra a nivel general la arquitectura de la red neuronal, con tres capas de redes neuronales convolucionales (CNN) y dos capas de redes neuronales densamente conectadas, la última empleada como capa de clasificación.

5 Resultados

Para la evaluación de los dos sistemas desarrollados vamos a usar las medidas tradicionales de clasificación precisión, cobertura y

[1] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

[2] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne

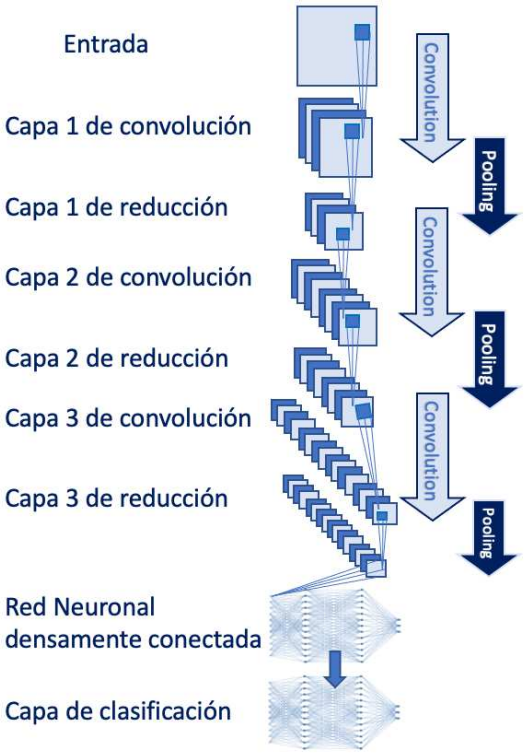


Figura 1: Arquitectura de la red neuronal.

medida-F. Además, como la detección de casos de autolesiones es una tarea cuyas implicaciones requieren una gran precisión, uno de los principales objetivos de este trabajo es la búsqueda de un buen desempeño a la hora de predecir casos positivos. También el hecho de ser una tarea relacionada con la salud, requiere de un grado satisfactorio de explicabilidad de cara a los profesionales que en última instancia deben de tomar las decisiones.

5.1 Baselines

En primer lugar, hemos desarrollado cuatro baselines para medir la calidad de los sistemas supervisados.

- **Most frequent Class (MFC):** El sistema anota todos los casos con la clase más frecuente, que en este caso es la clase negativa.
- **Less frequent Class (LFC):** El sistema anota todos los casos con la clase menos frecuente, que en este caso es la clase positiva.
- **Random prediction:** El sistema asigna una predicción aleatoria a cada instancia.
- **Random Ratio prediction:** El sistema asigna mediante una función de probabilidad una predicción aleatoria a cada instancia, manteniendo el mismo ratio (positivas/negativas) de anotaciones que el conjunto de test.

La Tabla 1 muestra los resultados tras la aplicación de los baselines. Como era de esperar, al tratarse de un corpus desbalanceado, el baseline que mejor rendimiento obtiene es aquel que predice como negativos todos los casos al ser la clase mayoritaria.

5.2 Combinación de Rasgos

En la Tabla 2 se pueden apreciar los resultados obtenidos tras diferentes combinaciones de los rasgos descritos en la sección 4.1.2. Para estos resultados se ha aplicado un algoritmo de regresión logística. De forma evidente en cuanto a la hipótesis de partida, los peores resultados se obtienen con los vectores de palabras formados por el vocabulario completo del corpus (32K palabras) y los mejores se consiguen con la combinación de todos los rasgos computados. En cuanto a la parte más interesante de esta combinación, destaca la diferencia entre la mejora obtenida por la clase negativa y la positiva al introducir los rasgos. La clase negativa solo aumenta tres puntos su medida-F, mientras que la clase positiva aumenta 21 puntos al introducir todos los rasgos. Esta diferencia demuestra la eficiencia de los rasgos introducidos en cuanto al propósito general de mejorar sobre todo la predicción de los casos positivos. En cuanto a los rasgos, analizados de manera individual, destaca la aportación de los vectores de palabras obtenidos a partir del vocabulario compilado manualmente de 53 palabras. La diferencia entre usar el vocabulario completo o solo las 53 palabras, se refleja en un aumento de 14 puntos en la medida-F de la clase positiva, aumentando tanto la precisión como la cobertura. Los rasgos que menos aportación parecen tener en el cómputo global son las distancias entre términos de autolesiones y la negación, quizás debido a que no se producen demasiadas en los informes médicos o su relevancia es menor de la esperada en cuanto al contexto global del informe. En cuanto a la clase positiva, la precisión y cobertura aumentan de forma desigual, teniendo los rasgos computados un impacto mayor en la precisión que en la cobertura. Este hecho era de esperar dado que al menos el rasgo que usa los vectores de palabras con un vocabulario

BASELINES					
Baseline	**F1 Todas Clases**			**Clase Positiva**	
	F1-NO	**F1-SI**	***F1-weighted Avg***	**P-SI**	**R-SI**
MFC	**0.95**	0.00	**0.86**	0.00	0.00
LFC	0.00	**0.17**	0.02	**0.09**	**1.00**
Random Prediction	0.65	0.14	0.60	0.08	0.43
Random Ratio Prediction	0.91	0.05	0.83	0.05	0.04

Tabla 1: F1-SI: F1-measure de los casos positivos, F1-NO: F1-measure de los casos negativos, F1-weighted Avg: F1-measure media de todo el conjunto de test, P-SI: Precisión de los casos positivos, R-SI: Recall de los casos positivos

reducido implica profundizar en esa dirección precisamente.

5.3 Análisis de diferentes algoritmos de Clasificación

En la sección anterior se empleó un algoritmo de regresión logística para la tarea de clasificación. En la Tabla 3 se muestran los resultados al aplicar los diferentes algoritmos de clasificación que mejor rendimiento han obtenido en diferentes trabajos del estado del arte consultados. Para esta comparativa se ha usado la combinación de rasgos que mejor rendimiento obtuvo en la sección anterior y cuyos resultados se pueden observar en la Tabla 2. Como se puede ver, hay tres algoritmos (Logistic Regression, Gradient Boosting y SVM) que obtienen los mejores resultados en cuanto a la medida-F global. Sin embargo, como uno de los objetivos de este trabajo consiste en mejorar la detección de la clase positiva, se observa que el algoritmo "Gradient Boosting" obtiene el mejor rendimiento en la predicción de casos positivos. Esto unido a que era uno de los tres algoritmos que de forma global obtenían mejores resultados lo convierten en la mejor opción para nuestro sistema. Profundizando en los resultados de "Gradient Boosting", aparte de obtener los mejores resultados en las clases positivas, negativas, y de forma global, obtiene mejor cobertura que ningún otro algoritmo. Esta parece ser su mejor aportación, ya que su precisión en la clase positiva es superada por otros algoritmos.

5.4 Sistema basado en Aprendizaje Profundo

En la Tabla 4 se puede observar el rendimiento de los sistemas basados en redes neuronales. Se ha optado por variar dos hiperparámetros como son el número de épocas y el dropout. En todos los experimentos se han usado embeddings de 200 dimensiones. De los resultados obtenidos en cuanto a las diferentes combinaciones no se pueden obtener demasiadas conclusiones. Quizás se puede observar que un dropout bajo mejora el rendimiento global aunque no es concluyente. De forma general podría decirse que cinco épocas han funcionado mejor que diez, al igual que ocurre para la clase positiva con la que ligeramente se observan mejores resultados. En cuanto a la precisión de los casos positivos, con una combinación se obtienen mucho mejores resultados que con el resto, sin embargo su cobertura y medida-F se ven negativamente afectadas. La única conclusión evidente a nivel de diferentes combinaciones se produce en la cobertura de la clase positiva. En este caso un menor número de épocas y un bajo dropout implican un significativo mejor rendimiento que los casos opuestos con una diferencia de 50 puntos.

5.5 Análisis Global de Resultados

De forma general y tal como se muestra en la Tabla 5, los sistemas desarrollados superan ampliamente a los baselines propuestos al inicio del trabajo. En cuanto a la comparativa entre el mejor sistema supervisado y el mejor sistema basado en redes neuronales, globalmente el sistema de aprendizaje profundo obtiene mejores resultados si atendemos a la medida-F. Sin embargo, la pequeña diferencia a favor de las redes neuronales en relación a la medida-F global y de la clase negativa, se ve ampliamente superada en cuanto a la clase positiva tanto en la medida-F como en la precisión y la cobertura. Destaca notablemente la diferencia en la precisión, de forma muy significativa la medida-F y de forma relevante la cobertura, siendo esta última la medida donde la diferencia de rendimiento es algo

COMBINACIÓN DE RASGOS					
Features	**F1-NO**	**F1-SI**	***F1-W Avg***	**P-SI**	**R-SI**
CV	0.94	0.47	0.90	0.61	0.33
CV + NSSI	0.96	0.49	0.92	0.64	0.39
CV + VAL	0.96	0.61	0.93	0.84	0.50
CV + VAL + NSSI	**0.97**	0.63	0.94	0.86	0.52
CV + VAL + NSSI + DIS + NEG	**0.97**	**0.65**	**0.95**	**0.87**	**0.54**

Tabla 2: CV: CountVectorizer, VAL: Vocabulario de Autolesiones, NSSI: Rasgos de NSSI, DIS: Rasgos de distancia de terminos de autolesion, NEG: Negación

ALGORITMOS DE CLASIFICACIÓN					
Algoritmo	**F1 Todas Clases**			**Clase Positiva**	
	F1-NO	**F1-SI**	***F1-weighted Avg***	**P-SI**	**R-SI**
Logistic Regression	**0.97**	0.65	**0.94**	0.86	0.52
Random Forest	0.95	0.53	0.91	0.50	0.57
Gradient Boosting	**0.97**	**0.68**	**0.94**	0.71	**0.65**
K Neighbours	0.96	0.36	0.91	**1.00**	0.22
SVM	**0.97**	0.59	**0.94**	0.91	0.43
Adaboost	0.96	0.59	0.93	0.62	0.57

Tabla 3: F1-SI: F1-measure de los casos positivos, F1-NO: F1-measure de los casos negativos, F1-weighted Avg: F1-measure media de todo el conjunto de test, P-SI: Precisión de los casos positivos, R-SI: Recall de los casos positivos

menor. De esta forma, y teniendo en cuenta las implicaciones en cuanto a mejor explicabilidad del sistema supervisado basado en el algoritmo "Gradient Boosting", consideramos que la opción más óptima para la tarea concreta en la que se centra este trabajo es dicho sistema.

5.6 Análisis de la Incidencia de las Categorías de Rasgos

Dado que el sistema supervisado ofrece un mejor rendimiento en la detección de casos positivos y además su grado de explicabilidad es mayor, hemos decidido profundizar en los rasgos extraídos y su incidencia en los resultados. En la figura 2 se muestra un gráfico de barras que representa la frecuencia de aparición de los rasgos que componen las diferentes categorías en función de la clase a la que pertenecen. Dicha frecuencia se ha normalizado en función del número de documentos de cada clase y tamaño del vocabulario de cada categoría, dado que el corpus está muy desbalanceado. En esta figura destacan positivamente categorías como el vocabulario de autolesiones, los términos de NSSI, los conceptos de autolesiones por cortes de NSSI y la negación. En cuanto al vocabulario de autolesiones, se intuía esta diferencia debido a los resultados obtenidos. En cuanto a la negación, también se observa una gran disparidad. Finalmente en cuanto a los conceptos de NSSI, los que mejor parecen representar a la clase positiva son los "Términos" y el "Cutting". Sin embargo, los conceptos de "Razones", "Métodos" e "Instrumentos" ofrecen una menor divergencia. Una posible explicación del distinto funcionamiento de estos conceptos de NSSI reside en el hecho de que los informes médicos tratan de reflejar los hechos más relevantes representados seguramente de una forma genérica y sin profundizar en determinados aspectos. De esta forma, los conceptos de "Razones", "Métodos" e "Instrumentos" implican un mayor detalle en la descripción del suceso del que se suele encontrar en un informe.

En la figura 3 se observa un gráfico de barras en el que aparecen las raíces de los términos más frecuentes del vocabulario de autolesiones ordenados por su frecuencia normalizada de aparición y en función de la clase a la que pertenecen. Como se puede observar, raíces como "autolesión", "cort" y "rasg" presentan una gran diferencia a favor de las clases positivas, mientras que otras raíces co-

Red Neuronal					
	F1 Todas Clases			Clase Positiva	
Modelo	**F1-NO**	**F1-SI**	***F1-weighted Avg***	**P-SI**	**R-SI**
Bert Multilingüe					
CVN + BERT (5ep,0.05do)	0.97	0.56	**0.95**	0.50	**0.64**
CVN + BERT (5ep,0.1do)	0.96	0.51	0.94	0.43	**0.64**
CVN + BERT (5ep,0.2do)	0.96	0.47	0.94	0.40	0.57
CVN + BERT (5ep,0.3do)	**0.98**	**0.57**	**0.95**	0.57	0.57
CVN + BERT (5ep,0.4do)	0.96	0.44	0.93	0.36	0.57
CVN + BERT (10ep,0.05do)	0.97	0.55	**0.95**	0.53	0.57
CVN + BERT (10ep,0.1do)	**0.98**	0.33	0.94	**0.75**	0.21
CVN + BERT (10ep,0.2do)	0.97	0.22	0.93	0.50	0.14
CVN + BERT (10ep,0.3do)	0.97	0.24	0.93	0.67	0.14
CVN + BERT (10ep,0.4do)	0.97	0.20	0.93	0.33	0.14
RoBERTa					
CVN + RoBERTa (5ep,0.05do)	0.96	0.39	0.92	0.43	0.35
CVN + RoBERTa (5ep,0.1do)	0.96	0.17	0.91	0.33	0.12
CVN + RoBERTa (5ep,0.2do)	0.96	0.37	0.93	0.50	0.29
CVN + RoBERTa (5ep,0.3do)	0.96	0.48	0.93	0.50	0.29
CVN + RoBERTa (5ep,0.4do)	0.96	0.54	0.94	0.50	0.59
CVN + RoBERTa (10ep,0.05do)	0.96	0.48	0.93	0.50	0.47
CVN + RoBERTa (10ep,0.1do)	0.96	0.41	0.93	0.50	0.35
CVN + RoBERTa (10ep,0.2do)	0.95	0.28	0.90	0.26	0.29
CVN + RoBERTa (10ep,0.3do)	0.97	0.55	**0.95**	0.67	0.47
CVN + RoBERTa (10ep,0.4do)	0.96	0.41	0.93	0.50	0.35

Tabla 4: F1-SI: F1-measure de los casos positivos, F1-NO: F1-measure de los casos negativos, F1-weighted Avg: F1-measure media de todo el conjunto de test, P-SI: Precisión de los casos positivos, R-SI: Recall de los casos positivos. Los sistemas varían en función del número de épocas (5-10 ep) y el dropout (0.05-0.4 do).

COMPARATIVA DE RESULTADOS					
Sistema	**F1 Todas Clases**			**Clase Positiva**	
	F1-NO	**F1-SI**	***F1-weighted Avg***	**P-SI**	**R-SI**
Baseline MFC	0.95	0.00	0.86	0.00	0.00
Baseline LFC	0.00	0.17	0.02	0.09	**1.00**
Gradient Boosting	0.97	**0.68**	0.94	**0.71**	**0.65**
CVN + BERT (5ep,0.3do)	**0.98**	0.57	**0.95**	0.57	0.57

Tabla 5: F1-SI: F1-measure de los casos positivos, F1-NO: F1-measure de los casos negativos, F1-weighted Avg: F1-measure media de todo el conjunto de test, P-SI: Precisión de los casos positivos, R-SI: Recall de los casos positivos

mo "tir", e "inger" muestran una aparición más equilibrada dado los conceptos más neutrales que representan en cuanto a las autolesiones. Destaca la raíz "sangr", teniendo más peso en la clase negativa debido seguramente a que las lesiones relacionadas con la sangre no son las más frecuentes en el problema estudiado.

6 Conclusiones y trabajo futuro

En este trabajo se presenta un sistema de detección de indicios de autolesiones no suicidas, basado en el análisis del lenguaje, sobre un conjunto anotado de informes médicos obtenidos del servicio de psiquiatría de un Hospital público madrileño. Dada la naturaleza tan crítica de la tarea y las implicaciones a

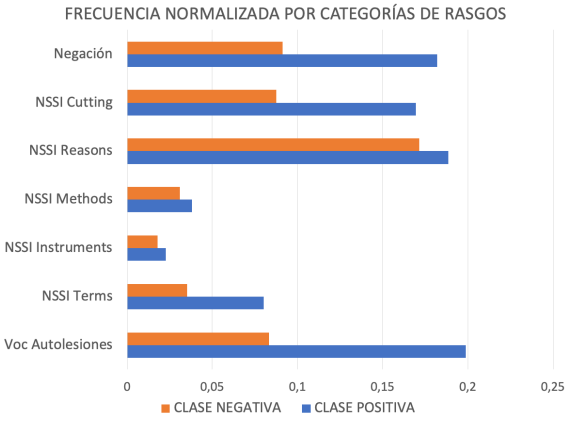


Figura 2: Frecuencia normalizada de las diferentes categorías de rasgos en función de la clase.

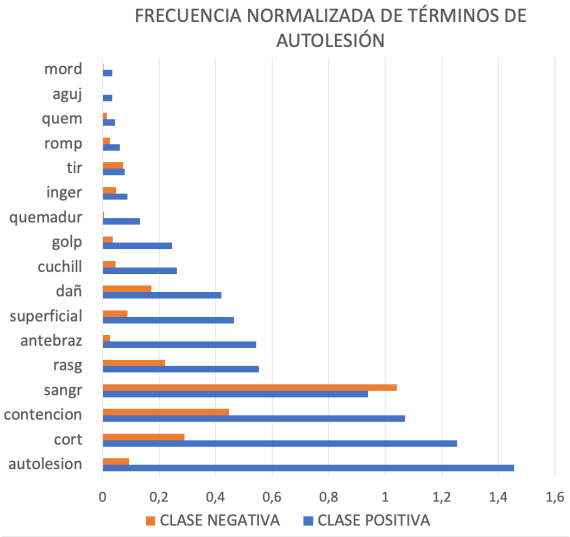


Figura 3: Frecuencia normalizada de términos de autolesiones en función de la clase.

la hora de predecir incorrectamente un caso, que realmente tiene detrás a un ser humano real, es necesario tratar este tipo de trabajos desde un punto de vista diferente al mero resultado obtenido por un conjunto de métricas de evaluación. Al inicio del trabajo se fijaron tres objetivos prioritarios: por un lado el sistema debería obtener un alto rendimiento para impedir en la medida de lo posible las predicciones erróneas, por otro lado la clasificación correcta de los casos positivos debería ser prioritaria, y finalmente se debería buscar la mayor explicabilidad del sistema para que el profesional sanitario pudiera tener la máxima información de cara a tomar una decisión final. Teniendo en cuenta estos requisitos, el trabajo ha cumplido con los objetivos. Por un lado el rendimiento global obtenido alcanza unos valores de medida-F de 0.95, alcanzando un 0.68 de medida-F para los casos positivos, lo cual es una prueba de su buen funcionamiento. Con la extracción de un conjunto de rasgos muy focalizados en alcanzar un mayor rendimiento en cuanto a la detección de los casos positivos, se ha conseguido el segundo objetivo equilibrando y mejorando tanto la precisión como la cobertura de forma significativa. Y finalmente, gracias al esfuerzo realizado para equiparar un sistema supervisado basado en algoritmos tradicionales de clasificación a un sistema basado en redes neuronales, se ha hecho posible el poder elegir el primero de los sistemas ya que con un rendimiento global similar tiene dos ventajas como son el mejor rendimiento en cuanto a la clasificación de casos positivos y una mejor explicabilidad debido a que los rasgos obtenidos forman parte de la decisión tomada finalmente por el sistema. En este último caso, el sistema basado en redes neuronales, a pesar de tener un ligero mejor rendimiento global, obtiene peores resultados en la clase positiva y además sus decisiones a día de hoy son difíciles de explicar de cara a un psicólogo o psiquiatra.

En cuanto al trabajo futuro, consideramos varias líneas de actuación. Por un lado el corpus está desbalanceado y además no dispone de un gran número de casos positivos. De esta forma trabajaremos para obtener un corpus de mayor tamaño con la esperanza de que un mayor número de casos positivos nos ayude a mejorar aún más la detección de este tipo de casos. Por otro lado, debido al gran potencial de tecnologías como las redes neuronales, trabajaremos para optimizar el sistema e incluir transformers con los objetivos de mejorar su rendimiento en cuanto a los casos positivos e iniciar un trabajo de estudio para mejorar la explicabilidad de este tipo de sistemas.

Agradecimientos

This work has been partially supported by the Spanish Ministry of Science and Innovation within the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32 and the project RAICES (IMIENS 2022).

Bibliografía

Ageitos, E. C., J. Martínez-Romo, y L. Araujo. 2020. Nlp-uned at erisk 2020: Self-harm early risk detection with sentiment

analysis and linguistic features. En *CLEF (Working Notes)*.

Baetens, I., L. Claes, J. Muehlenkamp, H. Grietens, y P. Onghena. 2011. Non-Suicidal and Suicidal Self-Injurious Behavior among Flemish Adolescents: A Web-Survey. *Archives of Suicide Research*, 15(1):56–67.

Burke, T. A., B. A. Ammerman, y R. Jacobucci. 2019. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *Journal of affective disorders*, 245:869–884.

Campillo-Ageitos, E., H. Fabregat, L. Araujo, y J. Martinez-Romo. 2021. Nlp-uned at erisk 2021: self-harm early risk detection with tf-idf and linguistic features. *Working Notes of CLEF*, páginas 21–24.

Delgado-Gomez, D., E. Baca-Garcia, D. Aguado, P. Courtet, y J. Lopez-Castroman. 2016. Computerized adaptive test vs. decision trees: development of a support decision system to identify suicidal behavior. *Journal of affective disorders*, 206:204–209.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fabregat, H., L. Araujo Serna, y J. Martínez Romo. 2019. Deep learning approach for negation trigger and scope recognition.

Fabregat, H., A. Duque, J. Martínez-Romo, y L. Araujo. 2019. Extending a deep learning approach for negation cues detection in spanish. En *IberLEF@ SEPLN*, páginas 369–377.

Fernandes, A. C., R. Dutta, S. Velupillai, J. Sanyal, R. Stewart, y D. Chandran. 2018. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Scientific reports*, 8(1):1–10.

Greaves, M. M. 2018a. *A Corpus Linguistic Analysis of Public Reddit and Tumblr Blog Posts on Non-Suicidal Self-Injury, An abstract*. Ph.D. tesis, College of Education, Oregon State University.

Greaves, M. M. 2018b. A corpus linguistic analysis of public reddit and tumblr blog posts on non-suicidal self-injury.

Haerian, K., H. Salmasian, y C. Friedman. 2012. Methods for identifying suicide or suicidal ideation in ehrs. En *AMIA annual symposium proceedings*, volumen 2012, página 1244. American Medical Informatics Association.

Kessler, R. C., M. B. Stein, M. V. Petukhova, P. Bliese, R. M. Bossarte, E. J. Bromet, C. S. Fullerton, S. E. Gilman, C. Ivany, L. Lewandowski-Romps, y others. 2017. Predicting suicides after outpatient mental health visits in the army study to assess risk and resilience in servicemembers (army starrs). *Molecular psychiatry*, 22(4):544–551.

LeCun, Y., Y. Bengio, y G. Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Losada, D. E., F. Crestani, y J. Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. En *International Conference of the Cross-Language Evaluation Forum for European Languages*, páginas 340–357. Springer.

Losada, D. E., F. Crestani, y J. Parapar. 2020. erisk 2020: Self-harm and depression challenges. En *European Conference on Information Retrieval*, páginas 557–563. Springer.

Loyola, J. M., S. Burdisso, H. Thompson, L. Cagnina, y M. Errecalde. 2021. Unsl at erisk 2021: A comparison of three early alert policies for early risk detection. En *Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum, Bucarest, Romania*.

Mann, J. J., S. P. Ellis, C. M. Waternaux, X. Liu, M. A. Oquendo, K. M. Malone, B. S. Brodsky, G. L. Haas, y D. Currier. 2008. Classification trees distinguish suicide attempters in major psychiatric disorders: a model of clinical decision making. *The Journal of clinical psychiatry*, 69(1):2693.

Martınez-Castano, R., A. Htait, L. Azzopardi, y Y. Moshfeghi. 2020. Early risk detection of self-harm and depression severity using bert-based transformers. *Working Notes of CLEF*, página 16.

McCoy, T. H., V. M. Castro, A. M. Roberson, L. A. Snapper, y R. H. Perlis. 2016. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. *JAMA psychiatry*, 73(10):1064–1071.

Metzger, M.-H., N. Tvardik, Q. Gicquel, C. Bouvry, E. Poulet, y V. Potinet-Pagliaroli. 2017. Use of emergency department electronic medical records for automated epidemiological surveillance of suicide attempts: a french pilot study. *International journal of methods in psychiatric research*, 26(2):e1522.

Nicolai, K. A., M. D. Wielgus, y A. Mezulis. 2016. Identifying Risk for Self-Harm: Rumination and Negative Affectivity in the Prospective Prediction of Nonsuicidal Self-Injury. *Suicide and Life-Threatening Behavior*, 46(2):223–233.

Obeid, J. S., J. Dahne, S. Christensen, S. Howard, T. Crawford, L. J. Frey, T. Stecker, y B. E. Bunnell. 2020. Identifying and predicting intentional self-harm in electronic health record clinical notes: deep learning approach. *JMIR medical informatics*, 8(7):e17784.

Obeid, J. S., E. R. Weeda, A. J. Matuskowitz, K. Gagnon, T. Crawford, C. M. Carr, y L. J. Frey. 2019. Automated detection of altered mental status in emergency department clinical notes: a deep learning approach. *BMC medical informatics and decision making*, 19(1):1–9.

Parapar, J., P. Martín-Rodilla, D. E. Losada, y F. Crestani. 2021. Overview of erisk 2021: Early risk prediction on the internet. En *International Conference of the Cross-Language Evaluation Forum for European Languages*, páginas 324–344. Springer.

Pennebaker, J. W., M. R. Mehl, y K. G. Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Poulin, C., B. Shiner, P. Thompson, L. Vepstas, Y. Young-Xu, B. Goertzel, B. Watts, L. Flashman, y T. McAllister. 2014. Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one*, 9(1):e85733.

Rodham, K., K. Hawton, y E. Evans. 2004. Reasons for deliberate self-harm: Comparison of self-poisoners and self-cutters in a community sample of adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 43(1):80–87.

Rozova, V., K. Witt, J. Robinson, Y. Li, y K. Verspoor. 2022. Detection of self-harm and suicidal ideation in emergency department triage notes. *Journal of the American Medical Informatics Association*, 29(3):472–480.

Walsh, C. G., J. D. Ribeiro, y J. C. Franklin. 2017. Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, 5(3):457–469.

Young, R., M. Van Beinum, H. Sweeting, y P. West. 2007. Young people who self-harm. *The British Journal of Psychiatry*, 191(1):44–49.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 141-152 recibido 31-03-2022 revisado 11-05-2022 aceptado 19-05-2022

Semantic Relations Predict the Bracketing of Three-Component Multiword Terms

Las Relaciones Semánticas Predicen la Desambiguación Estructural de las Unidades Terminológicas Poliléxicas con Tres Formantes

Juan Rojas-Garcia
University of Granada, Granada, Spain
juanrojas@ugr.es

Abstract: For English multiword terms (MWTs) of three or more constituents (e.g., *sea level rise*), a semantic analysis, based on linguistic and domain knowledge, is necessary to resolve the dependency between components. This structural disambiguation, often known as bracketing, involves the grouping of the dependent components so that the MWT is reduced to its basic form of modifier+head, as in [*sea level*] [*rise*]. Knowledge of these dependencies facilitates the comprehension of an MWT and its accurate translation into other languages. Moreover, the resolution of MWT bracketing provides a higher overall accuracy in machine translation systems and sentence parsers. This paper thus presents a pilot study that explored whether the bracketing of a ternary compound, when used as an argument in a sentence, can be predicted from the semantic information encoded in that sentence. It is shown that, with a random forest model, the semantic relation of the MWT to another argument in the same sentence, the lexical domain of the predicate, and the semantic role of the MWT were able to predict the bracketing of the 190 ternary compounds used as arguments in a sample of 188 semantically annotated sentences from a Coastal Engineering corpus (100% F_1-score). Furthermore, only the semantic relation of an MWT to another argument in the same sentence proved enormous capability to predict ternary compound bracketing with a binary decision-tree model (94.12% F_1-score).
Keywords: Semantic Relation, Multiword-Term Bracketing, Random Forest, Decision Tree.

Resumen: En unidades terminológicas poliléxicas (UTP) con tres o más formantes en lengua inglesa (p.ej., *sea level rise*), establecer la dependencia entre dichos formantes requiere de un análisis lingüístico y de conocimiento especializado del área concreta en que se emplean las UTP. Esta desambiguación estructural, o *bracketing*, implica el agrupamiento de los formantes para reducir la UTP a su estructura básica de modificador+núcleo, como en [*sea level*] [*rise*]. Conocer el *bracketing* de una UTP no solo facilita su comprensión y traducción a otras lenguas, sino que también mejora el desempeño de los sistemas de traducción automática y de los analizadores sintácticos. Por tanto, en este artículo presentamos un estudio piloto que explora si el *bracketing* de una UTP con tres formantes, al emplearse como argumento en una oración, puede predecirse a partir de la información semántica codificada en dicha oración. Se muestra que, con un modelo random forest, la relación semántica de la UTP con otro argumento en la misma oración, el dominio léxico del verbo y el rol semántico de la UTP son capaces de predecir el *bracketing* de las 190 UTP ternarias que se usan como argumento en una muestra de 188 oraciones, anotadas semánticamente y extraídas de un corpus sobre ingeniería de costas (con un valor de F_1 del 100 %). Además, únicamente la relación semántica que mantiene una UTP ternaria con otro argumento en la misma oración posee una enorme capacidad para predecir su *bracketing* mediante un árbol de decisión binario (con un valor de F_1 del 94,12 %).
Palabras clave: Relación Semántica, Desambiguación Estructural de Unidades Terminológicas Poliléxicas, Random Forest, Árbol de Decisión.

ISSN 1135-5948. DOI 10.26342/2022-69-12

1 Introduction

A set of 1,694 sentences from a Coastal Engineering corpus, in which a named river (e.g., Salinas River) was an argument of the predicate of the sentences, were semantically analyzed and annotated with the semantic relation between the arguments, the lexical domain of the predicates, and the semantic role of the arguments.

This paper presents the statistical analysis of those semantic annotations with a view to finding evidence that the structural disambiguation, or bracketing, of a three-component multiword term (e.g., [*sand supply*] [*decrease*]) can be predicted from the semantic information encoded in the sentence where the ternary compound is used as an argument. For this experiment, we assumed that the context, which constrains the factors that drive understanding (Leech G., 1981), also helps to resolve the structural disambiguation of a ternary compound. This assumption comes from the daily experience of a translator who must deal with ternary compounds in a specialized text. Although the compounds are somewhat familiar, it is useful to craft definitions for them to facilitate their translation into another language based on their context of use.

The rest of this paper is organized as follows. Section 2 presents a fundamental background of bracketing of multiword terms. Section 3 provides a literature review of predictive models for bracketing, mostly from the perspective of variables and resources used for the task of compound bracketing prediction. Section 4 explains the materials used in this study. Section 5 covers our semantic approach to predicting ternary compound bracketing based on two supervised models, namely decision tree, and random forest. Also described are the sample of ternary compounds, the training and testing phases for the predictive models, and the results, which provide linguistic insights as to how semantic relations, predicate lexical domains, and semantic roles are intertwined with the bracketing of ternary compounds. Section 6 discusses the results and compares them to those outlined in the literature review. Finally, Section 7 presents the conclusions derived from this work along with plans for future research.

2 Bracketing of Multiword Terms

When multiword expressions are used in specialized domains, they are known as multiword terms (MWTs). MWTs often have more than two components. For instance, in Coastal Engineering, *beach size sand supply* refers to the supply of sand, usually provided by rivers, whose grain size is appropriate to mitigate beach erosion. The most frequent MWTs in specialized texts are endocentric because they specify a broader concept or hypernym. For example, *beach size sand supply* is a type of *sand supply* since the grain size of the sand is specified. It is thus the dimension activated to form the hyponym.

For MWTs of three or more constituents, a semantic analysis, based on linguistic and domain knowledge, is necessary to resolve the dependency between components. This structural disambiguation, often known as *bracketing* or *parsing*, involves the grouping of the dependent components so that the MWT is reduced to its basic form of modifier+head, as in [*beach size*] [*sand supply*]. Knowledge of these dependencies facilitates the comprehension of an MWT and, consequently, its accurate translation into other languages.

Therefore, before including MWTs in terminological knowledge bases, it is often necessary to structurally disambiguate them to make their relational structure explicit and thus favor knowledge acquisition (León-Araúz P. et al., 2021). Furthermore, the resolution of MWT bracketing provides a higher overall accuracy in machine translation systems (Green N., 2011), sentence parsers (Vadas D. and Curran J.R., 2008), and in systems aimed at determining the implicit semantic relation holding between modifier and head in MWTs of three or more components (Kim S.N. and Baldwin T., 2013).

3 Review of Bracketing Prediction Methods

Previous work on compound parsing/bracketing exploits either unsupervised methods (e.g., based on bigram corpus frequency) or supervised ones (i.e., based on training data, containing manually parsed/bracketed compounds, which are used to train an algorithm for predicting compound bracketing).

The two basic unsupervised approaches are the adjacency model (Marcus M., 1980; Pustejovsky J. et al., 1993), and the dependency model (Lauer M., 1994). For a ternary compound such as *sea level rise* (i.e., increase in sea level), the adjacency model concludes whether *level* is more closely associated with *sea* (leading to a left-branched structure) or to *rise* (leading to a right-branched structure). In contrast, the dependency model resolves whether *sea* is more strongly associated with *level* (leading to a left-branched structure) or with *rise* (leading to a right-branched structure). In this case, the correct bracketing of *sea level rise* is left-branched. The way of measuring the association strength between two of the words (or constituents) in the compound is based on association measures estimated from corpus data,

such as bigram frequency, point-wise mutual information, or chi-squared, among others.

Resnik P.S.'s (1993) method for ternary compounds, based on the adjacency model and the association measure called *selectional association*, estimated from the parsed *Wall Street Journal* corpus (30 million words), achieved an overall accuracy of 72.6% (with a sample of 157 ternary compounds from the Penn Treebank corpus, 64.1% left-branched, and 35.9% right-branched). In contrast, Lauer M. (1995) adopted the dependency model for his method, based on the ratio of left- to right-bracketing probability for a ternary compound, estimated from *Grolier's Encyclopedia* (8 million words). The author calculated probabilities of conceptual categories in the taxonomy underlying *Roget's Thesaurus* (Roget P.M., 1852)[1], rather than for individual words, to avoid data sparsity problems. His method reached an overall accuracy of 80.7% (with a sample of 244 ternary compounds from *Grolier's Encyclopedia*, 66.8% left-branched, and 33.2% right-branched).

Nakov P. and Hearst M. (2005) developed an unsupervised, knowledge-rich method for parsing ternary compounds. Their approach included:

(1) Ten types of surface variable, such as dashes (e.g., *beach-sand transport* points to a left-bracketed compound), possessive markers (e.g., *city's water supply* indicates a right-bracketed compound), and acronyms (e.g., *pH quality control (QC)* reveals a right-bracketed compound).

(2) Three types of paraphrase variable, namely prepositional phrases (e.g., *distance from the river mouth* means that *river mouth distance* is left-bracketed), copula paraphrases (e.g., *water product that/which is a mixture* proves that *mixture water product* is right-bracketed), and verbal paraphrases (e.g., *impacts associated with river pollution* implies that *river pollution impact* is left-bracketed).

The authors concluded that the adjacency and dependency models showed comparable performance when using the chi-squared association measure and the number of web search engine page hits for approximating corpus frequencies, as suggested by Lapata M. and Keller F. (2004). Although their method achieved an overall accuracy of 95.35%, this result was probably biased toward the majority left-bracketing class because of the bracketing-imbalanced sample of 430 ternary compounds from a corpus of biomedical domain abstracts retrieved from MEDLINE (84% left-branched, and only 16% right-branched).

Girju R. et al. (2005) implemented a supervised model for bracketing ternary compounds with the machine-learning technique decision tree. They employed a total of 15 semantic variables based on WordNet senses, five variables for each compound constituent, namely the top three WordNet semantic categories for each constituent, derivationally-related forms, and whether the constituent was a nominalization. The algorithm reached an overall accuracy of 83.10%, with a sample of 728 ternary compounds from the *Wall Street Journal* component of the Penn Treebank corpus (Marcus M. et al., 1993), 67.4% left-branched, and 32.6% right-branched.

Kim S.N. and Baldwin T. (2013) devised a method that consisted of automatically determining the semantic relations between the pairs of words in a ternary compound, and then predicting bracketing from the constituent pair whose semantic relation coincided with that of the ternary compound. When this method was combined with that of Nakov P. and Hearst M. (2005), it achieved an overall accuracy of 74.1% with a sample of 1,571 ternary compounds from the *Wall Street Journal* corpus. However, no information was provided regarding the percentage of left- and right-bracketing within the sample.

The supervised method by Bergsma S. et al. (2010) used both n-gram variables (the logarithm of the frequency of all constituent subsets appearing in the Google V2 corpus), and Boolean lexical variables that indicated the presence or absence of a particular string at a given position in the compound (the constituents and their position, the entire ternary compound, as well as a capitalization pattern of the constituent sequence). As a machine-learning technique, the authors applied the support-vector machine algorithm, which reached an overall accuracy of 91.6%, with a sample of 2,150 ternary compounds from the *Wall Street Journal* corpus (70.5% left-branched, and 29.5% right-branched).

Vadas D. and Curran J.R. (2007) developed a supervised method for parsing ternary compounds based on the machine-learning technique of logistic regression. They used 88,568 variables, an extremely large number, which can be summarized as follows:

(1) Bigram frequencies were collected from two sources, namely hit counts from the web search engine Google, and frequencies in the Google Web 1T corpus (Brants T. and Franz A., 1993).

(2) The pairs of compound constituents, and the surface variables by Nakov P. and Hearst M. (2005), were compared according to both the adjacency and dependency models by means of the chi-squared, and bigram probability association measures.

[1] http://www.gutenberg.org/ebooks/10681.

(3) Lexical features for all unigrams and bigrams in a ternary compound, along with their position within the compound.

(4) Contextual variables, consisting of bag-of-word features for both the words in the sentence where the compound is used, and for a two-word window on each side of the compound.

(5) For every n-gram and context window feature, their part-of-speech tags and named entity tags were added.

(6) For each sense of each constituent in the ternary compound, a semantic feature for its synset, as well as the synset of each of its hypernyms up to the root, were extracted from WordNet, and incorporated into the supervised model as additional variables.

This method achieved an F_1-score of 93.01%, with a sample of 5,582 ternary compounds from the Penn Treebank corpus (58.99% left-branched, and 41.01% right-branched).

The supervised system by Pitler E. et al. (2010) was able to bracket compounds of three or more constituents (including the conjunction *and*). Applying the support-vector machine algorithm, the system first calculated the probability that a word sequence, within a compound, was a constituent, given the entire compound as context. Then, using these probabilities, the system predicted the bracketing of a compound with the CYK parser (i.e., Cocke-Younger-Kasami algorithm). As variables for the system, the authors employed:

(1) The position of the proposed bracketing within the compound.

(2) The association measure point-wise mutual information (PMI) between all word pairs in the compound, derived from the Google V2 corpus (Lin D. et al., 2010).

(3) Boolean lexical variables to indicate the presence of a particular word at each position in the compound.

(4) Boolean variables to inform about the shape of the compound, namely the presence of capitalized letters and hyphenated words provided information concerning the possibility that the compound included a named entity.

The system reached an overall accuracy of 95.4%, with a sample of 64,844 compounds of three or more constituents from the Penn Treebank corpus, but bracketing-related information in the form of percentages was not provided.

Lazaridou A. et al. (2013) tackled the parsing of a ternary compound using a *semantic plausibility measure* derived from a distributional semantic model trained on a corpus of 2.8 billion tokens, where the vector of a ternary compound was obtained from the combination of the vectors of each of its constituents. This supervised method relied on the support-vector machine algorithm with 14 variables, summarized as follows: (1) 12 variables for representing the semantic plausibility of either the left- or right-bracketing; (2) two variables for the PMI values of the word pairs in the compound, according to the adjacency model. The method achieved an overall accuracy of 85.6%, with a sample of 2,227 ternary compounds from the Penn Treebank corpus (34.4% left-branched, and 65.6% right-branched).

Faruqui M. and Dyer C. (2015) also addressed the ternary compound bracketing with word vectors. However, their semantic model was non-distributional because the vectors did not encode any word co-occurrence information. Instead, the vector dimensions were Boolean variables that represented linguistic knowledge derived from resources such as WordNet (Fellbaum C.A., 1998), FrameNet (Ruppenhofer J. et al., 2010), and Penn Treebank. As such, the vector length for a single word included a total of 172,418 dimensions. The vector of a ternary compound was then obtained by appending the vector of each constituent, which resulted in a ternary compound vector of 517,254 dimensions. This combined vector was the input of the machine-learning technique of logistic regression, which achieved an overall accuracy of 83.3% in the same sample of ternary compounds collected by Lazaridou A. et al. (2013).

For the unsupervised method by Ménard P.A. and Barrière C. (2014), the usage of different resources for the bracketing of compounds of three and more constituents was compared, namely the English Google Web N-grams (Lin D. et al., 2010), English Google Books Ngrams (Michel J.B. et al., 2010), and open linked data DBpedia (Hellmann S. et al., 2009). The association measures chi-squared, PMI, and Dice, and the number of valid DBPedia paths were also analyzed. Their algorithm created an initial list containing all of the word pairs from a compound, which were then sorted in descending order of association scores. A second list of dependencies, which defined the complete bracketing of the compound, was constructed from the first list. For ternary compounds, the method with the English Google Books N-grams and the PMI achieved the highest overall accuracy, with a value of 81.47% on a sample of 2,889 ternary compounds from the Penn Treebank corpus (79.2% left-branched, and 20.8% right-branched).

Similarly, for the bracketing of compounds of three and more constituents, Barrière C. and Ménard P.A. (2014) applied the unsupervised method of Ménard P.A. and Barrière C. (2014), but relied on a word association model that combined the lexical,

relational, and coordinate nature of the associations between all pairs of words within a compound. The information for their word association model was collected from Wikipedia. The system reached an overall accuracy of 73.16%, with a sample of 4,749 compounds of three and more constituents from the Penn Treebank corpus, but the specific accuracy for the subset of ternary compounds was not provided.

León-Araúz P. et al. (2021) developed an unsupervised, knowledge-rich method for bracketing specialized ternary compounds in the domain of wind energy. The authors used 12 variables, mainly related to the surface and paraphrase variables proposed by Nakov P. and Hearst M. (2005), which measured frequency counts in a specialized corpus on wind energy. The counts were collected by means of CQL (Corpus Query Language) queries in the Sketch Engine corpus manager. A total of 34 specific CQL queries were designed for the extraction of occurrences of each of the linguistic structures underlying the 12 variables. Based on the results, the authors formulated 16 rules to decide on the bracketing of a ternary compound. Hence, the final bracketing structure was decided by applying the majority vote strategy to the votes of the individual rules. As such, the CQL queries and rules permitted the implementation of a system to automate the compound bracketing task for users such as translators and terminologists. The method achieved an overall accuracy of 86.4%, with a sample of 103 ternary compounds from the wind energy domain (67% left-branched, and 33% right-branched).

In short, previous research focused on semantic information provided by the components of an MWT. The number of variables used for prediction ranged from 12 to 517,254 features. These variables were mostly based on n-gram statistics, and semantic information of the MWT components stored in linguistic resources such as WordNet. The overall accuracy of the prediction models ranged from 72.60% to 95.40%.

Our approach, however, was based on semantic information that previous research has not as yet considered. This semantic information was encoded in both the co-text of a ternary compound (i.e., the sentence where the ternary compound was used as an argument) and the ternary compound seen as a unit (i.e., its semantic role). The set of predictor variables consisted of only three (i.e., the semantic relation, predicate lexical domain, and semantic role of the MWT), whereas previous research employed a minimum of 12 variables (León-Araúz P. et al., 2021).

4 Materials

A set of 1,694 sentences, in which a named river (e.g., Mississippi River) was an argument of the predicate of the sentences, were semantically analyzed and annotated. These sentences were extracted from a subcorpus of English texts on Coastal Engineering, comprising roughly 7 million tokens and composed of specialized texts (scientific articles, technical reports, and PhD dissertations), and semi-specialized texts (textbooks and encyclopedias on Coastal Engineering). This subcorpus is part of the English EcoLexicon Corpus (23.1 million words) (see León-Araúz P. et al. (2018) for a detailed description).

5 Semantic Approach for MWT Bracketing

Since the semantic information in a sentence firmly guides its syntactic parsing (Fillmore C.J., 1968; Lazaridou A. et al., 2013), one could assume that the correct bracketing of an MWT, when used as an argument in a sentence, can be predicted from the semantic information encoded in that sentence. In other words, the context, which constrains the factors that drive understanding (Leech G., 1981), helps to resolve the structural disambiguation of the ternary compound.

As semantic information in a sentence, this pilot study explored the contribution of three semantic variables to the prediction of ternary compound bracketing. These variables were the lexical domain of the verb, semantic role of the ternary compound, and semantic relation of the ternary compound to the named river. From the 1,694 sentences semantically analyzed and annotated, 188 sentences contained 190 ternary compounds as arguments. This sample of 190 ternary compounds, along with the values of the abovementioned three semantic variables annotated in their corresponding sentences, were employed for the training and testing of two supervised models to predict whether a ternary compound was right-branched or left-branched.

5.1 Annotation of the Semantic Variables

A set of 1,694 sentences from the corpus, where 294 different rivers are mentioned, were annotated by three terminologists from the LexiCon research group of the University of Granada (Spain). They performed the semantic annotation of the predicate-argument structure of a sentence by assigning a: (1) lexical domain to the predicate; (2) semantic role to the arguments of the predicate; (3) semantic relation to the link between the named river and the other arguments in the sentence; and (4) bracketing (left or right) to the ternary compounds used as arguments in the sentence. The values of these four semantic variables are shown in Table 1.

Semantic variables annotated	Values
Lexical domain of the predicates (8 values)	CHANGE, MOVEMENT, EXISTENCE, POSSESSION, POSITION, MANIPULATION, ACTION, COGNITION
Semantic roles of the arguments (13 values)	AGENT, RESULT, PATIENT, THEME, LOCATION, RECIPIENT, INSTRUMENT, TIME, RATE, MANNER, DESCRIPTION, CONDITION, PURPOSE
Semantic relation between the ternary compound and the named river (30 values)	*type_of, part_of, made_of, delimited_by, located_at, takes_place_in, phase_of, affects, causes, result_of, attribute_of, has_function, studies, measures, effected_by, improves, worsens, creates, becomes, gives, gives_to, receives, receives_from, drains, has_path, transfers, discharges_into, places, controls, applied_to*
Bracketing of the ternary compounds in the sentences (2 values)	RIGHT, LEFT

Table 1: Semantic variables annotated in the set of sentences, and their values.

The most frequent verbs in the corpus are general language verbs (e.g., *accumulate*, *pollute*, *increase*, *discharge*, *supply*, *drain*), which are also used in specialized texts and thus reflect how environmental entities interact. In this sense, such verbs are susceptible to classification in the lexical domains proposed by Faber P. and Mairal R. (1999), within the Functional Lexematic Model. These lexical domains were used to annotate the predicates of our set of sentences, and shown in Table 1.

Specialized knowledge representation includes semantic properties that help to describe the nature of entities and processes. These semantic properties are reflected as the relations between a predicate and its arguments, which are typical semantic roles. The semantic roles used to annotate the arguments in our set of sentences largely coincided with those specified by Kroeger P.R. (2005: 54-55), and Thompson P. et al. (2009), and summarized in Table 1.

Conceptual description of specialized concepts includes their relational behavior. These relations, depicted by Faber P. et al. (2009) for environmental concepts, with additional non-hierarchical relations specific to named rivers (Rojas-Garcia J., forthcoming), were all used to annotate the semantic relation between the arguments in our set of sentences, and collected in Table 1.

The inter-annotation agreement coefficient, *Cohen's kappa* (κ), showed a very good agreement for all the annotator pairs (κ>90%, *p*-value<0.05) in the annotation of the semantic roles, relations, and bracketing according to Krippendorff K.'s (2012) recommendations for text content analysis. Notwithstanding, the disagreements in the original annotations were resolved based on discussion between the annotators to reach a consensus on the definitive annotations of semantic roles, relations, and bracketing.

For the initial annotation of predicates with lexical domains, the inter-annotation agreement was lower for all the annotator pairs (84%<κ<88%, *p*-value<0.05), indicating that this variable lent itself to alternative, though plausible, interpretations. A review of the differences between annotators showed that the lexical domains of MOVEMENT and POSSESSION were more prone to confusion. The issues fundamentally arose from verbs that could potentially belong to more than one lexical domain (e.g., *drain* and *discharge*), as Faber P. and Mairal R. (1999) already proved. To arrive at a consensus on the definitive annotations of lexical domains, the factorization of meaning from the Functional Lexematic Model framework was applied to verbs to resolve disagreements between the annotators.

5.2 Description of the Sample of MWTs

A selection of 10 sentences from the sample, which incorporated ternary compounds as arguments, is provided in Table 2. For each of those 10 sentences, Table 3 shows the values of the following four annotated variables: (1) lexical domain of the predicate (*LexDom*); (2) semantic role of the ternary compound (*SemRol_mwt*); (3) semantic relation between the ternary compound and the named river (*SemRel*); and (4) bracketing of the ternary compound (*Bracketing*), which was the variable to be predicted.[2]

The distribution of bracketing structures within the MWT sample was reasonably balanced between left-branching (110 MWTs, 58% of the sample), and right-branching (80 MWTs, 42% of the sample). Table 4 summarizes the counts for the sample data, disaggregated by lexical domain and bracketing structure of the MWTs, and describes the distribution of the 190 MWTs across these variables. Some conclusions could be drawn from the characteristics of the sample: (1) sentences whose predicate belonged to the lexical domains of MOVEMENT, ACTION, POSITION, MANIPULATION, and COGNITION included ternary compounds which were only right-branched; and (2) sentences whose predicate belonged to the lexical domain of POSSESSION incorporated ternary compounds which were only left-branched.

[2] The whole dataset of MWTs, the values of the annotated variables, and the corpus will be available on the website of the LexiCon research group of the University of Granada (Granada, Spain) (http://lexicon.ugr.es/).

Sentences from the Sample with Ternary Compounds as Arguments
(1) Blackstone River draining into Narragansett Bay has been extensively dammed, and although not well quantified, models *show* **decreasing sediment load** in the **Blackstone River**.
(2) The dramatical sediment load variation in the **Pearl River**, with the almost unchanged **water discharge level**, *represents* an example of such effect that human activities can have on river deltas.
(3) **Muddy silt deposition** in the **Clyne River** discharging into the Swansea Bay *would increase*.
(4) **Rising sea levels** *change* **Salinas River Estuary** and could thus potentially alter sediment supplies and process patterns.
(5) The **Salinas River** no longer *contributes* substantial **beach size sand** to the Littoral Cell because the river gradient has greatly decreased with sea level rise, reducing the flow rate.
(6) The **River Murray** flows across Tertiary formations to *enter* coastal lagoons behind the **dune calcarenite barriers** of Encounter Bay.
(7) Not all the sediments drained by the **Dee River** *participate* to **coastal sediment transport**.
(8) The field site for this study is the **Zuidgors salt marsh**, *located* in the **Western Scheldt estuary** in The Netherlands.
(9) **Natural sediment supply** within this region *is defined* by the **Ventura River** that drains large watersheds.
(10) The **average discharge rate** of beach size sand in the **Salinas River** *is estimated* at approximately 65,000 cubic yards per year.

Table 2: Selection of 10 sentences (from the sample of 188 sentences), which included 10 ternary compounds as arguments.

MWT	LexDom	SemRol_mwt	SemRel	Bracketing
decreasing [sediment load]	EXISTENCE	DESCRIPTION	*attribute_of*	RIGHT
[water discharge] level	EXISTENCE	THEME	*attribute_of*	LEFT
[muddy silt] deposition	CHANGE	PATIENT	*takes_place_in*	LEFT
rising [sea level]	CHANGE	AGENT	*worsens*	RIGHT
[beach size] sand	POSSESSION	THEME	*gives*	LEFT
dune [calcarenite barrier]	MOVEMENT	AGENT	*has_path*	RIGHT
coastal [sediment transport]	ACTION	DESCRIPTION	*affects*	RIGHT
Zuidgors [salt marsh]	POSITION	THEME	*located_at*	RIGHT
natural [sediment supply]	MANIPULATION	PATIENT	*controls*	RIGHT
average [discharge rate]	COGNITION	THEME	*attribute_of*	RIGHT

Table 3: Semantic annotations and variables for a set of 10 MWTs out of the 190 MWTs that comprised the sample. The semantic information in the rows corresponds to the respective sentences in Table 2.

Lexical Domain	LEFT-branched MWTs	RIGHT-branched MWTs	Total
MOVEMENT	0	10	10 (5.3%)
POSSESSION	30	0	30 (15.8%)
CHANGE	20	10	30 (15.8%)
EXISTENCE	60	20	80 (42.0%)
ACTION	0	10	10 (5.3%)
POSITION	0	10	10 (5.3%)
MANIPULAT.	0	10	10 (5.3%)
COGNITION	0	10	10 (5.3%)
Total	**110 (58%)**	**80 (42%)**	**190 (100%)**

Table 4: Description of the sample of ternary compounds.

5.3 Supervised Models

Regarding the supervised models for classification, *binary decision tree* and *random forest* were tested to predict ternary compound bracketing. Since variables in our dataset were categorical, both tree-based models were adopted because they can efficiently manage qualitative variables (James G. et al., 2015: 315).

A decision-tree model is simple and readily interpretable because the set of prediction rules is graphically summarized in a tree, typically drawn upside down, in the sense that the terminal nodes or leaves, which convey the predictions, are at the bottom of the tree. However, it is usually not competitive with other predictive models.

For that reason, we also experimented with a random forest model, which produces a large number of decision trees, and then combines them to reach a single consensus prediction. Namely, each tree in the ensemble (or forest) casts a vote for the bracketing of an MWT, which is finally classified into the bracketing structure that has the most votes. Random forest models thus lead to remarkable improvements in prediction accuracy, at the expense of loss in interpretation since it is difficult to obtain insight as to how the model makes the predictions.

5.4 Data Splitting

For the construction and evaluation of the models, the dataset with the 190 MWTs was divided into two: (1) the training dataset to create the models (with 133 MWTs, 70% of the original dataset), and (2) the test dataset to qualify model performance (57 MWTs, 30% of the original dataset).

For both the training and test datasets to have the same distribution in the outcome variable (i.e., *Bracketing*) as the original dataset (i.e., 58% left-branched MWTs, and 42% right-branched MWTs), stratified random sampling was conducted, which randomly sampled observations within the classes LEFT and RIGHT of the *Bracketing* variable in the original dataset.

5.5 Model Performance Measures

The quality of the two models (decision tree and random forest) was assessed by analyzing how well they performed on the test dataset, which was hidden from the model-building process for evaluation purposes. As such, the predictions of the models were compared to the true classes of the test dataset (i.e., the true bracketing structures LEFT and RIGHT, recorded in the *Bracketing* variable of the test dataset), and performance measures were calculated.

A widely used performance measure is *overall accuracy*, which provides the percentage of correctly classified instances. However, this measure has some drawbacks in imbalanced datasets, or datasets whose outcome variable exhibits a significant disproportion among the number of instances of each class.

According to Fernández A. et al. (2018: vii), the learning process of most classification algorithms, including decision tree and random forest, is often biased toward the majority-class instances, and minority-class ones are thus not well modelled into the final system. Consequently, in imbalanced scenarios, the accuracy measure may mask a poor classification performance in the minority class. Unfortunately, as already seen in the literature review, there is much research on bracketing prediction that still uses overall accuracy with severely bracketing-imbalanced datasets. Therefore, despite the fact that our dataset was only slightly bracketing-imbalanced, we preferred to use, in addition to accuracy, other measures that were not sensitive to disparities in the class proportions to evaluate classification performance. Such measures were the *area under the ROC curve*, and the *F_1-score* (Fernández A. et al., 2018: 52-55).

The *receiver operating characteristic (ROC) curve* is a function of the sensitivity and specificity of a two-class predictive model to evaluate its trade-off between both measures. *Sensitivity* is the fraction of the minority-class instances (in our case, the right-branched MWTs) that are correctly classified, whereas *specificity* refers to the proportion of the majority-class instances (in our case, the left-branched MWTs) that are correctly classified. Hence, the *a*rea *u*nder the ROC *c*urve (henceforth referred to as AUC) is a method for combining sensitivity and specificity into a single value. AUC ranges from 0 to 1. The higher the AUC, the better the performance of the model at distinguishing between the two classes.

The F_1-score is the harmonic mean between the precision and recall of a predictive model. *Precision* is the fraction of correctly classified minority-class instances among the instances classified as belonging to the minority class, whereas *recall* is the same as sensitivity. Thus, the F_1-score evaluates the trade-off between correctness and coverage in classifying minority-class instances.

5.6 Construction of the Predictive Models

The predictors *SemRel*, *LexDom*, and *SemRol_mwt* were used to construct two predictive models with the *caret* package (Kuhn M., 2021) for the R programming language.

For the random forest, 7-fold cross-validation in the training dataset was used to evaluate its performance in training. Although 10 folds are conventionally employed, we chose 7 folds, a divisor of 133, so that the number of instances in all folds would be the same (i.e., 19 instances). During the process of tuning parameters, the AUC performance measure was chosen to be maximized. Accordingly, the random forest model attained in training an AUC value equal to 1.0 when: (1) the splits in the trees were allowed to use one predictor of a subset of one predictor; and (2) the number of trees in the forest was, surprisingly, only three trees. In the test dataset, the random forest also achieved an AUC value equal to 1.0. Consequently, the three predictors were capable of correctly predicting bracketing in the test dataset with a random forest model.

Similarly, for the decision tree, 7-fold cross-validation in the training dataset was employed to evaluate its performance in training. During the process of tuning parameters, the AUC performance measure was also chosen to be maximized. Therefore, the decision-tree model yielded in training the greatest AUC, equal to 0.9545, when: (1) the cost-complexity parameter (*cp*) was equal to *cp*=0.8392857; and (2) the splitting criterion for predictors was the *information gain*, and not the *Gini index*. In the test dataset, the decision-tree model achieved an AUC value also equal to 0.9545, which indicated a very satisfactory performance.

Table 5 provides further performance measures, in the training and test datasets, for the random forest and decision-tree models.

Predictors: *SemRel*, *LexDom*, and *SemRol_mwt*					
	Decision Tree Model				
Dataset	**AUC**	**Precision**	**Recall**	**F_1**	**Accura.**
Train	0.9545	0.8952	1.000	0.9430	0.9474
Test	**0.9545**	0.8889	1.000	**0.9412**	**0.9474**
	Random Forest Model (3 ensembled decision trees)				
Train	1.0000	1.0000	1.0000	1.0000	1.0000
Test	**1.0000**	1.0000	1.0000	**1.0000**	**1.0000**

Table 5: Performance measures of the models for bracketing prediction with the predictors semantic relation, lexical domain, and semantic role.

Since the decision-tree model reached a significant AUC in the test dataset (AUC=0.9545), its only prediction rule, graphically summarized in Figure 1, is worth mentioning.

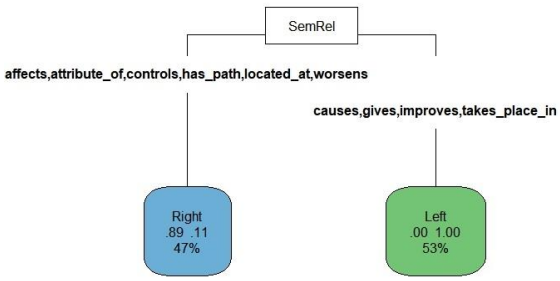


Figure 1: Classification tree for bracketing prediction, inferred by the decision-tree model trained with the predictors semantic relation, lexical domain, and semantic role of the MWTs.

In our constrained context (i.e., specialized ternary compounds from Coastal Engineering, used in sentences where a named river was mentioned), the classification tree of the model, displayed in Figure 1, can be interpreted as follows.

SemRel was the most important factor in determining *Bracketing*, and the only predictor selected by the decision-tree model. In our opinion, the predictive power of the semantic relation between an MWT and another argument in the same sentence is so high that the model was obliged to reject the use of the predictors *LexDom* and *SemRol_mwt* to avoid overfitting to training data.

As such, the ternary compounds whose semantic relation to the other argument, filled with a named river in our case, belonged to the group formed by *causes*, *gives*, *improves*, and *takes_place_in* (right-hand branch in the classification tree) accounted for 53% of the sample; these MWTs were all left-branched and correctly classified. It thus seemed that these four semantic relations forced the use of only left-branched MWTs.

In contrast, the ternary compounds whose semantic relation to the other argument fell into the group formed by *affects*, *attribute_of*, *controls*, *has_path*, *located_at*, and *worsens* (left-hand branch in the classification tree) comprised 47% of the sample, and could be right- or left-branched; under these conditions, the model correctly classified all the right-branched MWTs (89%), but misclassified the true left-branched MWTs (11%) as right-branched.

An analysis of the errors made by the decision-tree model revealed that, both in the training and test datasets, those left-branched MWTs with the values *SemRel=attribute_of*, *LexDom*=EXISTENCE, and *SemRol_mwt*=THEME (e.g., *water discharge level*, in row 2 of Table 3), were all misclassified as right-branched.

5.7 Baseline Models

The results of our semantic approach were compared to those of four baseline models, namely: (1) adjacency model with the point-wise mutual information (PMI) association measure, as defined by Marcus M. (1980); (2) adjacency model with the chi-squared association measure; (3) dependency model with PMI; and (4) dependency model with chi-squared. These non-supervised models, widely used in the literature on bracketing prediction, were applied to the whole sample of 190 MWTs.

Table 6 shows that the two predictive models, explained in this paper, outperformed the baseline models. Furthermore, the dependency model achieved better performance than the adjacency model, and the chi-squared association measure yielded better results than PMI.

Models	Precision	Recall	F_1
Adjacency model with PMI	0.6444	0.7250	0.6823
Adjacency model with chi-squared	0.6623	0.7375	0.6979
Dependency model with PMI	0.6818	0.7500	0.7143
Dependency model with chi-squared	0.7011	0.7625	0.7305
Decision tree model	0.8889	1.0000	**0.9412**
Random forest model	1.0000	1.0000	**1.0000**

Table 6: Comparison of the decision-tree and random forest models to four baseline models.

5.8 Comparison of the Models

Despite the promising results, it is obvious that further investigation is necessary to acquire a more in-depth understanding of the influence of the semantic variables in this study on ternary compound bracketing. Therefore, the following statements should be considered scope-bounded because they were derived from a restricted framework in which this research was conducted, namely specialized ternary compounds from Coastal Engineering used in sentences mentioning named rivers.

As far as the selection of the best model is concerned, there are convincing arguments in favor of either model. Since the random forest model had an error-free performance, it could be used to implement a system for bracketing ternary compounds.

Nevertheless, the performance of the decision-tree model was also fairly good. It also has the advantage of interpretability and visualization, which affords linguistic insights into how ternary compound bracketing is governed by semantic information encoded in a sentence. Since the binary decision-tree model only needed the *SemRel* predictor to achieve a highly satisfactory level of performance, practical applications for automatic bracketing could employ

solely the semantic relation between a ternary compound and another argument in the same sentence.

6 Discussion

Although the comparison of our study with previous research in the literature review is far from ideal, it still serves as an indication of the performance of our semantic approach.

For bracketing prediction, previous research focused on semantic information provided by the components of an MWT. The number of variables that they used for prediction ranged from 12 to 517,254 features. These variables were mostly based on n-gram statistics, which could arguably capture some semantic information encoded in frequent co-occurrences of MWT components (Lazaridou A. et al., 2013: 1909). Other research studies relied on semantic information of the MWT components stored in linguistic resources such as WordNet. The overall accuracy of the prediction models ranged from 72.60% to 95.40%.

Our semantic approach, however, was based on semantic information that previous research has not as yet considered. The semantic information was encoded in both the co-text of a ternary compound (i.e., the sentence where the ternary compound was used as an argument) and the ternary compound seen as a unit (i.e., its semantic role). The set of variables consisted of only three (semantic relation, lexical domain, and semantic role of the MWT), whereas previous research employed a minimum of 12 variables (León-Araúz P. et al., 2021). This set of three variables yielded, in the test dataset, an error-free performance with a random forest model, whereas the highest overall accuracy achieved in previous research was 95.40% with support vector machine (Pitler E. et al., 2010), a less interpretable predictive model.

7 Conclusions

A set of 1,694 sentences, in which a named river was an argument of the predicate of the sentences, were semantically analyzed and annotated with the lexical domain of the predicates, the semantic role of the arguments, and the semantic relation between the arguments. Those semantic annotations were analyzed to see whether the bracketing of a ternary compound, when used as an argument in a sentence, can be predicted from the semantic information encoded in that sentence.

The semantic relation of the MWT to another argument in the same sentence, the lexical domain of the predicate, and the semantic role of the MWT were able to predict the bracketing of the 190 ternary compounds used as arguments in a sample of 188 semantically annotated sentences (out of the 1,694 annotated sentences). A random forest model, with three ensembled decision trees, achieved in the test dataset an AUC equal to 100% (overall accuracy of 100%). When a decision tree was trained, the model only needed the semantic relation to yield, in the test dataset, an AUC equal to 95.45% (overall accuracy of 94.74%). Hence, the semantic relation of an MWT to another argument in the same sentence proved enormous capability to predict ternary compound bracketing.

Therefore, this pilot study showed that the semantic information in a sentence, encoded in the semantic relation of the MWT to another argument in the same sentence, the lexical domain of the predicate, and the semantic role of the MWT, contributed substantially to compound parsing. Given the beneficial effects of multiword-term bracketing on overall accuracy of sentence parsers (Vadas D. and Curran J.R., 2008), and machine translation systems (Green N., 2011), this result potentially suggests a novel research direction in the integration of such semantic variables into syntactic parsers and machine translation applications, in line with Agirre E. et al. (2008), Girju R. et al. (2005), and Kim S.N. and Baldwin T. (2013).

Evidently, it is not as yet clear whether such semantic variables are also able to predict the bracketing of MWTs of four or more constituents. This issue is thus deferred for further investigation.

Finally, notwithstanding the promising results, they should be considered scope-bounded because of the small size of the MWT sample and the restricted framework in which the analysis has been conducted, namely specialized ternary compounds from Coastal Engineering used in sentences that mentioned named rivers. In future research, a wider framework shall be established to acquire a more profound understanding of the influence of the semantic variables focused in this study on multiword-term bracketing.

Acknowledgements

This research was carried out as part of projects PID2020-118369GB-I00, "Transversal Integration of Culture in a Terminological Knowledge Base on Environment" (TRANSCULTURE), funded by the Spanish Ministry of Science and Innovation; and A-HUM-600-UGR20, "Culture as Transversal Module in a Terminological Knowledge Base on the Environment" (CULTURAMA), funded by the Andalusian Ministry of Economy, Knowledge, Business, and University.

References

Agirre, E., T. Baldwin, and D. Martinez (2008). Improving parsing and PP attachment performance with sense information. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 317-325). ACL.

Barrière, C., and P.A. Ménard (2014). Multiword noun compound bracketing using Wikipedia. In *Proceedings of the First Workshop on Computational Approaches to Compound Analysis (ComAComA 2014)* (pp. 72-80). ACL.

Bergsma, S., E. Pitler, and D. Lin (2010). Creating robust supervised classifiers via web-scale n-gram data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 865-874). ACL.

Brants, T., and A. Franz (2006). Web 1T 5-gram Version 1. Linguistic Data Consortium.

Faber, P., and R. Mairal (1999). *Constructing a Lexicon of English Verbs*. Mouton de Gruyter.

Faber, P., P. León-Araúz, and J.A. Prieto (2009). Semantic relations, dynamicity, and terminological knowledge bases. *Current Issues in Language Studies, 1*, 1-23.

Faruqui, M., and C. Dyer (2015). Non-distributional word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics* (pp. 464-469). ACL.

Fellbaum, C.A. (1998). Semantic network of English: The mother of all WordNets. *Computers and the Humanities, 32*, 209-220.

Fernández, A., S. García, M. Galar, R.C. Prati, B. Krawczyk, and F. Herrera (2018). *Learning from Imbalanced Data Sets*. Springer.

Fillmore, C.J. (1968). The case for case. In E. Bach, and R. Harms (Eds.), *Universals in Linguistic Theory* (pp. 1-89). Holt, Rinehart, and Winston.

Girju, R., D.I. Moldovan, M. Tatu, and D. Antohe (2005). On the semantics of noun compounds. *Computer Speech and Language, 19*(4), 479-496.

Green, N. (2011). Effects of noun phrase bracketing in dependency parsing and machine translation. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Proceedings of Student Session* (pp. 69-74). ACL.

Hellmann, S., C. Stadler, J. Lehmann, and S. Auer (2009). DBpedia live extraction. In R. Meersman, T. Dillon, and P. Herrero (Eds.), *On the Move to Meaningful Internet Systems (OTM 2009)* (Vol. 5871, pp. 1209-1223). Springer. Lecture Notes in Computer Science.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2015). *An Introduction to Statistical Learning*. Springer.

Kim, S.N., and T. Baldwin (2013). A lexical semantic approach to interpreting and bracketing English noun compounds. *Natural Language Engineering, 19*(3), 385-407.

Krippendorff, K. (2012). *Content Analysis: An Introduction to its Methodology*. Sage.

Kroeger, P.R. (2005). *Analyzing Grammar: An Introduction*. Cambridge University Press.

Kuhn, M. (2021). *caret: Classification and Regression Training*. R package version 6.0-90.

Lapata, M., and F. Keller (2004). The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL (HLT-NAACL 2004)* (pp. 121-128). ACL.

Lauer, M. (1994). *Conceptual Association for Compound Noun Analysis*. CoRR.

Lauer, M. (1995). Corpus statistics meet the noun compound: Some empirical results. In *Proceedings of the 33rd Annual Meeting of the ACL* (pp. 47-54). ACL.

Lazaridou, A., E.M. Vecchi, and M. Baroni (2013). Fish transporters and miracle homes: How compositional distributional semantics can help NP parsing. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)* (pp. 1908-1913). ACL.

Leech, G. (1981). *Semantics: The Study of Meaning*. Penguin.

León-Araúz, P., A. San Martín, and A. Reimerink (2018). The EcoLexicon English corpus as an open corpus in Sketch Engine. In *Proceedings of the 18th EURALEX International Congress* (pp. 893-901). Euralex.

León-Araúz, P., M. Cabezas-García, and P. Faber (2021). Multiword-term bracketing and representation in terminological knowledge bases.

In *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2021 Conference* (pp. 139-163). Lexical Computing CZ.

Lin, D., K.W Church, H. Ji, S. Sekine, D. Yarowsky, S. Bergsma, K. Patil, E. Pitler, R. Lathbury, V. Rao, K. Dalwani, and S. Narsale (2010). New tools for web-scale n-grams. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (pp. 2221-2227). ELRA.

Marcus, M. (1980). *A Theory of Syntactic Recognition for Natural Language*. MIT Press.

Marcus, M.P., M.A. Marcinkiewicz, and B. Santorini (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313-330.

Ménard, P.A., and C. Barrière (2014). Linked open data and web corpus data for noun compound bracketing. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)* (pp. 702-709). ELRA.

Michel, J.B., Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, T.G.B. Team, J.P. Pickett, D. Holberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, and E.L. Aiden (2010). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176-182.

Nakov, P., and M. Hearst (2005). Search engine statistics beyond the n-gram: Application to noun compound bracketing. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL)* (pp. 17-24). ACL.

Pitler, E., S. Bergsma, D. Lin, and K.W. Church (2010). Using web-scale n-grams to improve base NP parsing performance. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)* (pp. 886-894). ACL.

Pustejovsky, J., P. Anick, and S. Bergler (1993). Lexical semantic techniques for corpus analysis. *Computational Linguistics*, *19*(2), 331-358.

Resnik, P.S. (1993). *Selection and Information: A Class-based Approach to Lexical Relationships*. Ph.D. Thesis. University of Pennsylvania.

Roget, P.M. (1852). *Roget's Thesaurus of English Words and Phrases*. Available in Project Gutemberg. https://www.gutenberg.org/ebooks/10681.

Rojas-Garcia, J. (forthcoming). Semantic representation of context for the inclusion of named rivers in a terminological knowledge base. *Frontiers in Psychology*.

Ruppenhofer, J., M. Ellsworth, M.R.L. Petruck, C.R. Johnson, and J. Scheffczyk (2010). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute.

Thompson, P., S.A. Iqbal, J. McNaught, and S. Ananiadou (2009). Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics*, *10*, 349.

Vadas, D., and J.R Curran (2007). Large-scale supervised models for noun phrase bracketing. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING-2007)* (pp. 104-112). PACLING.

Vadas, D., and J.R. Curran (2008). Parsing noun phrase structure with CCG. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 335-343). ACL.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 153-164 recibido 30-03-2022 revisado 19-05-2022 aceptado 23-05-2022

Evaluating Contextualized Vectors from both Large Language Models and Compositional Strategies

Evaluando vectores contextualizados generados a partir de grandes modelos de lenguaje y de estrategias composicionales

Pablo Gamallo, Marcos Garcia, Iria de-Dios-Flores
Centro de Investigación en Tecnoloxías Intelixentes (CITIUS)
Universidade de Santiago de Compostela, Galiza
{pablo.gamallo, marcos.garcia.gonzalez, iria.dedios}@usc.gal

Abstract: In this article, we compare contextualized vectors derived from large language models with those generated by means of dependency-based compositional techniques. For this purpose, we make use of a word-in-context similarity task. As all experiments are conducted for the Galician language, we created a new Galician evaluation dataset for this specific semantic task. The results show that compositional vectors derived from syntactic approaches based on selectional preferences are competitive with the contextual embeddings derived from neural-based large language models.
Keywords: Large Language Models, Contextualized Vectors, Comositionality, Semantic Similarity, Selection Preferences, Syntactic Dependencies.

Resumen: En este artículo, comparamos los vectores contextualizados derivados de grandes modelos de lenguaje con los generados mediante técnicas de composición basadas en dependencias sintácticas. Para ello, nos servimos de una tarea de similitud de palabras en contextos controldados. Como se trata de una experimentación orientada a la lengua gallega, creamos un nuevo conjunto de datos de evaluación en gallego para esta tarea semántica específica. Los resultados muestran que los vectores composicionales derivados de enfoques sintácticos basados en restricciones de selección son competitivos con los *embeddings* contextuales derivados de los modelos de lenguaje de gran tamaño basados en arquitecturas neuronales.
Palabras clave: Grandes Modelos de Lenguaje, Vectores Contextualizados, Composicionalidad, Similitud Semántica, Restricciones de Seleccion, Dependencias Sintácticas.

1 Introduction

Large Language Models (LLMs) are a disruptive breakthrough in Artificial Intelligence that have received an increasing amount of attention in many Natural Language Processing (NLP) tasks. As in the case of classical models, it is possible to use two different approaches to evaluate LLMs: intrinsic and extrinsic evaluations. Intrinsic evaluation consists of using a metric to evaluate the language model itself, without considering any task in which it may be involved. Extrinsic evaluation consists of evaluating the models by employing them in a downstream NLP task. This strategy allows us to compare how their final representation affects the accomplishment of the target task.

Perplexity is one of the most popular metrics for intrinsically evaluating language models. It measures how good a language model is at predicting real sentences. Although perplexity measurements allow researchers to assess the quality of a model in a fast and inexpensive way, it is not considered a fair metric to compare models because the final value is highly dependent on the models' size and vocabulary. In addition, while this metric can be easily applied to classical language models, it is not well-defined for auto-encoding LLM (Salazar et al., 2020), such as masked language models such as BERT (Devlin et al., 2019).

Most commonly, LLMs are evaluated on several NLP tasks by making use of extensive and comprehensive benchmarks (e.g., GLUE (Wang et al., 2019)). However, extrinsic evaluation also has some drawbacks. First, it is a costly and computationally slow process,

ISSN 1135-5948. DOI 10.26342/2022-69-13

since it requires supervised fine-tuning (i.e., training a new model with annotated examples to adapt it to the task). Second, hyperparameters for fine-tuning are likely to have an important influence on the results of the evaluations (Shibayama et al., 2020). And third, the most comprehensive datasets to evaluate LLMs are only available for either English or a dozen of mid-resource languages (Lin et al., 2021), but not for low-resource languages such as Galician.

As an alternative to fine-tuning, which adjusts the vector weights with new annotated data, it is possible to optimize a pre-trained language model for many different tasks by making use of prompt tuning, which is a sort of zero-shot learning approach based on the optimization of the model by embedding the description of the task in the input. So LLMs can also be externally evaluated through the evaluation of their prompted-based tasks.

Another way to evaluate LLMs is to do so on the basis of some of their components, for instance word embeddings. LLMs transform input sentences into contextual vectors of each token constituent. These sensitive context word embeddings are seen as components that are dynamically derived from the LLM. Contextual embeddings can be evaluated in a manner analogous to the way non-contextual and static word embeddings are evaluated. While the latter are evaluated intrinsically on subtasks searching for word similarity and analogy completion out of context, the former can be evaluated by means of tasks that measure both in-context word similarity and sentence similarity. For LLMs, these tasks are simpler and faster than extrinsic evaluations, since they do not require supervision or fine-tuning, and allow us to directly check the quality of the model that generated the contextual embeddings. It should be noted that, even though this type of evaluation is known as intrinsic evaluation of embeddings, it is not actually intrinsic for the LLM from which the embeddings are derived. To avoid terminological confusion, we will call it *vector-based evaluation* of LLMs.

Importantly, contextual embeddings can be generated not only from LLMs, but also by compositional techniques that combine static embeddings, as described in numerous works on compositional distributional semantics (Baroni, 2013; Weir et al., 2016; Gamallo et al., 2019; Wijnholds, Sadrzadeh, and Clark, 2020). In some of these approaches, static embeddings representing the meaning of words in a sentence are combined by syntactic dependencies in an entirely compositional manner, resulting in contextualized vectors of each constituent word (Gamallo et al., 2019; Weir et al., 2016).

The aim of this work is to compare contextual embeddings generated from LLMs with those generated by using syntax-based compositional techniques. All embeddings will be evaluated in a word-in-context similarity task. To do so a very specific dataset is needed because the compositional techniques, due to their linguistic complexity, can only be evaluated on controlled and simple syntactic constructions (e.g. adjective-noun, noun-verb, noun-verb-noun, etc). For this purpose, we created a syntactically controlled dataset in Galician language. In sum, the main contributions of the paper are the following:

- Creation of a new Galician dataset to perform word-in-context similarity tasks.
- Vector-based evaluation (via contextualized word embeddings) of four different LLMs, namely three BERT monolingual models for Galician, and the official multilingual one (mBERT).
- Evaluation of dependency-based compositional vectors generated from Galician Wikipedia.
- Comparison of the performance of all these dynamic and contextually sensitive embeddings against the same dataset.

The rest of the article is organized as follows. The next section introduces some related work (2). Then, the different types of language models, both LLMs and dependency-based, are defined in Section 3. The results and the dataset used in the evaluation are described and analyzed in Section 4. Finally, the conclusions are presented in Section 5.

2 Related work

The first well-known datasets to evaluate contextualized vectors in controlled syntactic constructions are those described in Mitchell and Lapata (2008; 2010). The authors did

not actually use the term *contextualized vectors* for what they called the representation of the meaning of sentences in vector space by means of vector composition. In their work, the meaning of phrases or sentences is represented as the combination of constituent word vectors together with arithmetic operations such as addition and component-wise multiplication. The main drawback of this approach is that it is not fully compositional because word order and syntactic functions are not taken into account. The dataset created by Mitchell and Lapata (2008) in order to evaluate vector composition contains pairs of intransitive English sentences (subject-verb constructions) differing only in the verb. In Mitchell and Lapata (2010) the dataset contains pairs of verb-object constructions differing also in the verb.

Later, Grefenstette et al. (2011a) and Kartsaklis and Sadrzadeh (2013) built very similar evaluation datasets, always for English. These datasets also consist of pairs of sentences, all of which are subject-verb-object transitive constructions that differ only in the verb. Yet, unlike the previous work by Mitchell and Lapata, the semantic approaches that were evaluated on these datasets were full compositional models based on functional words represented as high-dimensional tensors (Baroni, Bernardi, and Zamparelli, 2014). The main concern with these approaches is that they require several high-order tensor representations of verbs with several arguments, something which is computationally inefficient.

To facilitate linguistic preprocessing, all sentences in those datasets are presented as sequences of lemmas, for instance *ball_ricochet* instead of *the ball ricocheted*. Thus, they are not true sentence but n-grams of lemmas representing controlled syntactic constructions.

The increasing development of context-sensitive word embeddings derived from neural language models has marginalized syntax-based and compositional semantic models. One of the main reasons for the low interest in these models is the difficulty to adapt them to open phrases and sentences with any type of syntactic construction. Purely compositional models, due to their linguistic complexity, are so far only successfully applied to datasets with controlled syntactic expressions. In contrast, as context-sensitive embeddings derived from LLMs are built in open syntactic environments, most datasets available and used in shared tasks are composed of open text without syntactic constraints (Pilehvar and Camacho-Collados, 2019; Armendariz et al., 2020). However, it is worth mentioning that the syntactically controlled datasets cited above have been used to compare the two contextual approaches, that is, both compositional embeddings built by means of dependencies and contextual embeddings derived from LLMs, e.g., Wijnholds et al. (2020) and Gamallo et al. (2021) for English, and Gamallo et al. (2021) for Portuguese and Spanish.

There are recent studies which also take advantage of syntactically controlled datasets (such as BiRD (Asaadi, Mohammad, and Kiritchenko, 2019)) to probe the compositional abilities of LLMs. In this respect, Yu and Ettinger (2020) found that Transformer-based models mostly rely on word content, and therefore miss additional information provided by compositional operations.

Finally, recent approaches (Nguyen et al., 2020; Bai et al., 2021) use syntactic information to improve self-attention mechanism, resulting in interesting attempts to include compositional semantic strategies to build the contextualized meaning of words from LLMs.

3 Contextualized Word Vectors from Galician Language Models

Contextualized word vectors can be derived from different types of language models following distributional-based strategies. In our work we explore contextualized word vectors from two types of models for the Galician language, described below: (a) BERT-based LLMs, and (b) transparent models with syntactic dependencies.

3.1 Contextualized Word Vectors from BERT-based Models

Besides the official multilingual model (mBERT, with 12 hidden layers) provided by Devlin et al. (2019), we evaluate the following monolingual models: Bertinho-base, with 12 layers (Vilares, Garcia, and Gómez-Rodríguez, 2021), and two models of Bert-Galician ('base' and 'small') released by Garcia (2021), with 12 and 6 layers, respectively. Concerning the size of the training corpus of each model, mBERT and Bertinho-base were

trained on the Wikipedia, which contains about 42M tokens, while the two versions of Bert-Galician were trained on a larger corpus with about 500M tokens.

To obtain the contextualized vector of a word in the input sentence, we use the standard approach of adding the last four layers, as they have been found to provide more context-specific representations (Ethayarajh, 2019; Vulić et al., 2020). When the tokenizer divides a word into several sub-words (or affixes), only the first subword is considered since it represents the lexical stem of the full token.

We also generate sentence embeddings from LLMs by making use a pooling strategy. This is the same strategy used by Sentence-BERT for English (Reimers and Gurevych, 2019). The main difference with regard to Sentence-BERT is that the Galician pretrained models of our experiments are not fine-tuned with annotated collections of semantically similar pairs of sentences. The basic pooling strategy used to generate our sentence embeddings consists of computing the mean of all output vectors.

3.2 Contextualized Word Vectors from a Galician Dependency-Based Model

3.2.1 Selectional Preferences

Dependency-based distributional models, also known as structured vector spaces, allow us to directly deal with issues related to semantic compositionality and selectional preferences between syntactically related words. To build such a syntax-based model in a transparent way, we opt for a count-based strategy with explicit and sparse dimensions representing lexical-syntactic contexts of words. For instance, given the dependency $(obj, catch, ball)$, representing the heading verb *catch* occurring with dependent noun *ball* in the direct object relation (obj), we extract two lexical-syntactic contexts: either being a dependent noun occurring with *catch* in *obj* relation, or being a verbal head occurring with *ball* in *obj* relation.

The high number of dimensions (lexical-syntactic contexts) of the vector space is reduced by selecting the N most relevant contexts per word (Biemann and Riedl, 2013; Padró et al., 2014; Gamallo, 2017), where N is a global, arbitrarily defined constant whose usual values range from 100 to 1000 (Padró et al., 2014). The relevance value of a context with regard to a word is computed by means of a lexical association measure (e.g., pointwise mutual information, loglikelihood, etc.). This is an explicit, transparent, and static representation of word meaning, very similar to the predictive-based and also static (i.e. out of context) representation known as word embeddings (Mikolov, Yih, and Zweig, 2013).

In order to build contextualized word vectors from these dependency-based representations, we follow the concept of selectional preference formalized in Erk and Padó (2008), which states that the two words related by a dependency relation impose restrictions on each other. Let $A(obj, catch, ball)$ denote the lexical association A between verbal head *catch* and dependent noun *ball* via relation *obj* in a parsed corpus, then the selectional preferences, noted h and d, imposed by the two lemmas on each other in relation *obj* are computed in equations 1 and 2.

$$\vec{h}_{ball}(obj) = \sum_{h:A(obj,h,ball)>\theta} \vec{h} \qquad (1)$$

$$\vec{d}_{catch}(obj) = \sum_{d:A(obj,catch,d)>\theta} \vec{d} \qquad (2)$$

Where $h : A(obj, h, ball) > \theta$ is the set of heading verbs (e.g., *catch, throw, organize...*) that have *ball* as *obj* with a lexical association value higher than threshold θ, and $d : A(obj, catch, d) > \theta$ is the set of dependent nouns (e.g., *ball, baseball, cold, drift...*) occurring with *catch* via *obj* with an association value higher than θ. Note that the former represents the paradigmatic class of those relevant verbs having *ball* as direct object, while the latter is the paradigmatic class of relevant nouns appearing as direct objects of *catch*. In both cases, the selectional preferences imposed by the two related lemmas result in two new compositional vectors, $\vec{h}_{ball}(obj)$ and $\vec{d}_{catch}(obj)$, created by the iterative sum of the static vectors, respectively noted $\vec{h}$ and $\vec{d}$, of the paradigmatic classes (see equations 1 and 2).

Once the selectional preferences are built, they are combined by component-wise multiplication with the static vectors of both the head and dependent lemmas, giving rise to

two new contextualized vectors: the vector of the head lemma, $\vec{catch}_{(obj,h,ball)}$, contextualized with the selectional preferences of *ball* (equation 3), and the vector of the dependent lemma, $\vec{ball}_{(obj,catch,d)}$, contextualized with the selectional preferences of *catch* (equation 4).

$$\vec{catch}_{(obj,h,ball)} = \vec{catch} \odot \vec{h}_{ball}(obj) \quad (3)$$

$$\vec{ball}_{(obj,catch,d)} = \vec{ball} \odot \vec{d}_{catch}(obj) \quad (4)$$

At the end of this compositional process, the two contextualized vectors represent the in-context meaning of the two related words, which are more precise than the out-of-context meaning of the initial static vectors: *catch* means in the context of *ball* some event similar to *grab*, and not to *contract* as in *catch a disease*, while *ball* means in the context of *catch* a spherical object and not and dancing event as in *attend a ball*. In sum, these two contextualized vectors represent discriminated and disambiguated word senses.

3.2.2 Incremental Contextualization

So far we have defined the process of compositional semantics between two dependent words, but this process can be extended to the sentence level. Given the dependency parse tree of an input sentence, the contextualization of all constituent words in the sentence is the result of applying the compositional operations carried out by all dependencies identified in the parse tree in an iterative and incremental way. Thus, at the end of the process, each word of the sentence, including the root one, is assigned a contextualized vector. The order in which compositional operations are applied is not predetermined and the incremental and iterative process can go either from left-to-right or from right-to-left.

3.2.3 Galician Model

To build the language model and their corresponding vector space, the Galician Wikipedia (dump file of November 2019) was parsed with LinguaKit (Gamallo et al., 2018).[1] The Linguakit module used for this purpose is *dep*(endencies), which in turn makes use of the PoS tagger and lemmatizer modules. Since it is a small corpus containing about 42.7 million tokens, we used lemmas as the main lexical unit. Lemmas appearing less than 100 times were filtered out, and lexico-syntactic contexts with frequency less than 50 were removed. Then, for each lemma, we selected the 500 most relevant lexico-syntactic contexts by means of loglikelihood as lexical association measure. The final model resulted in a non-zero matrix of about 50k different lemmas and over 33k different contexts. In total, the vector space consists of a non-zero matrix with about 4.251 million word-context pairs. All static vectors of out-of-context Galician words are derived from this language space. See Gamallo (2017) for more details on how dependency-based vectors are built.

The software used to dynamically build contextualized word vectors from the Galician static vectors is freely available.[2] This is an improved upgrade we implemented on the basis of an older version that was fully described in Gamallo (2019). For the Galician corpus, the θ parameter was set to 0. Previous experiments did not show any improvement by assigning positive values to this parameter. It means that the second selection of relevant contexts made by this parameter is not justified with small corpus sizes such as the one used here.

Although the compositional strategy is designed to work with any type of sentence, due to the difficulty of the task, the implemented version only applies to linguistic expressions with a fixed and predefined syntactic structure (e.g., adjective-noun, subject-verb-object, and son on).

4 Evaluation

In order to compare contextual embeddings generated from BERT-based models with those generated with the compositional dependency-based strategy, we use a word-in-context similarity task in Galician. For this purpose, we created a syntactically controlled Galician dataset with subject-verb-object sentences.

4.1 Test Dataset

Following the structure of the English dataset described in Grefenstette and Sadrzadeh (2011a), a new Galician dataset with 192 sentence pairs of subject-verb-object sentences

[1] https://github.com/citiususc/Linguakit

[2] https://github.com/gamallo/DepFunc

was built.[3]

As it is not possible to make a direct translation of the original English sentences, since the selection preferences are very different from one language to another, we chose analogous examples with 68 different polysemous verbs and 149 different nouns in subject and object position. All sentences consist of just one basic nominal phrase as subject, a verb as predicate, and a basic nominal phrase as direct object.

In each pair, one transitive sentence consisting of a verb with its subject and direct object is compared to another transitive sentence combining the same subject and object with a semantically related verb that is chosen to be either appropriate or inappropriate in the same context. For instance, *a empresa compra un político* ('the company buys a politician') is semantically appropriate and very close to *a empresa suborna un político* ('the company bribes a politician') as *comprar* ('buy') is a very close synonym of *subornar* ('bribe') in this context, where the subject is a person or organization and the object is also a person or organization. However, the same pair of verbs have a very dissimilar behavior in a different context, e.g., *o director compra unha acción / ??o director suborna unha acción* ('the director buys a share' / '??the director bribes a stock'), as the verb *subornar* ('bribe') cannot be applied on objects that are not provided with the human feature. The selectional preferences imposed by that verb are not fulfilled by the direct object.

Unlike the original English dataset from which it is inspired, we created complete sentences, and not just triples of lemmas. However, all sentences were also lemmatized to enable to be evaluated with the dependency-based approach. Most verbs and their arguments were adapted (and not literally translated) to Galician from the English original dataset.

Three native speakers of Galician (and expert linguists) were asked to rate the degree of semantic correctness and similarity of each sentence pair using a 1 to 7 Likert scale. The average scores per annotator are 4.22, 3.45 and 3.94, with the following standard deviations: 1.96, 2.02 and 1.97, respectively. In order to measure the reliability of the ratings provided by the annotators, we calculated an intraclass correlation coefficient (ICC) using the *irr* package in R (Gamer et al., 2019). The agreement ICC was 0.71, indicating a high reliability among raters. Then, the average of the three scores per pair was computed.

As in many intrinsic evaluations of word embeddings, we compute Spearman correlation between human scores (the average of the three evaluators) and the predictions returned by the systems. Both human evaluators and systems should provide high scores to semantically similar sentence pairs with a high degree of semantic correctness.

4.2 Types of Sentence Similarity

Contextualized word embeddings are powerful semantic artifacts that can be used to measure the similarity between two sentences from different points of view. In the following subsections, we define different types of sentence similarity depending on which is the most representative constituent of the sentence.

4.2.1 BERT Sentence Similarity

As all constituent words are fully contextualized, we assume that any of them can represent the whole sentence semantically from a specific point of view. For example, in the sentence *the president signed the decree*, in addition to the verb, the contextualized subject refers to the president who signed the decree, while the contextualized direct object designates the decree that is signed by the president. So, in a transitive sentence, each of the three contextualized word vectors (subject, verb, or object) might be used to compute similarity at the sentence level (and not just at the word level). Moreover, it is also possible to build a new vector representing the whole sentence by combining the embeddings of its constituent words. In total, we can build the following four vectors:

BERT - verb : Contextualized vector of the verb head, resulting from adding the 4 last layers.

BERT - subj : Contextualized vector of the subject word, resulting from adding the 4 last layers.

BERT - obj : Contextualized vector of the direct object, resulting from adding the 4 last layers.

[3]The dataset is available with the software (Cf. footnote 2).

BERT - sentence : Mean of all output vectors.

Note that for English, Sentence-BERT (Reimers and Gurevych, 2019) generates fixed sized vectors of sentences in a way that is similar to our BERT-sentence strategy. There are, however, two significant differences: Sentence-BERT was derived from BERT-large (with 24 layers) and was fine-tuned with two very large dataset collections: SNLI (Bowman et al., 2015) and MultiNLI (Williams, Nangia, and Bowman, 2018) containing 1 million sentence pairs which were annotated for semantic tasks such as inference, contradiction, and entailment. So, while Sentence-BERT is a fine-tuned model trained with a supervised technique on annotated corpora, BERT-sentence is a fully unsupervised model.

4.2.2 Dependency-Based Sentence Similarity

This strategy builds compositional vectors in an incremental way. Thus, it is sensitive to the order of application of the identified syntactic dependencies. The semantic meaning of *The company buys the politician* can be interpreted either from left to right (see 5 below) or from right to left (see 6), according to the order in which the two dependencies of the sentence are applied:

$$(nsubj, buy, company), (obj, buy, politician) \quad (5)$$

$$(obj, buy, politician), (nsubj, buy, company) \quad (6)$$

Then, considering the direction of the compositional process, several compositional vectors representing the meaning of the transitive sentence are built:

left-to-right - verb : This builds the compositional vector of the verb head *buy*. It results from being contextualized first by the selectional preferences imposed by the nominal subject *company* and then by the selectional preferences of the direct object *politician*.

left-to-right - obj : This builds the compositional vector of the direct object *politician*. It results from being contextualized by the preferences imposed by *buy* previously combined with the subject *company*.

left-to-right - sentence : The addition of the two previous left-to-right values (head and dep).

right-to-left - verb : This builds the compositional vector of the verb head *buy*. It results from being contextualized first by the selectional preferences imposed by the direct object *politician* and then by the selectional preferences of the subject *company*.

right-to-left - subj : This builds the compositional vector of the subject *company*. It results from being contextualized by the preferences imposed by *buy* previously combined with the direct object *politician*.

right-to-left - sentence : The addition of the two previous right-to-left values (head and dep).

Note that, in the left-to-right direction, the object is fully contextualized by the verb and the subject. By contrast, the subject is not contextualized by the object, so that this partially contextualized sense of the subject is not used to represent the sentence. The same occurs for the direct object in the right-to-left compositional processes. The incremental direction is not relevant in BERT LLMs because the BERT-like strategy relies on bidirectional scanning by jointly conditioning on both left and right context in all layers. Hence, any constituent word is contextualized by the other words in both left-to-right and right-to-left direction.

4.3 Results

All the models and their different vector configurations were evaluated using the Galician dataset described above.

Table 1 shows the results of the four BERT models for the four types of contextualized vectors introduced in subsection 4.2.1, by using Spearman correlation between the system scores and the human evaluators. We observe that the most significant element of the meaning of the sentence is the contextualized sense of the verb -something expected because the verb is the syntactic root. The verb provides even better results than the sentence method, as a representative of the whole sentence, in three of the four models. By contrast, the subject tends to be the least significant constituent. Perhaps this might

be explained by the fact that in transitive constructions the object is mostly determined by the verb (through more restrictive selection preferences) than by the subject.

If we focus on the comparison of the four LLMs, we observe that the best one is clearly BERT-base (57 for the verb), followed by BERT-small (46). The big differences between these two LLMs and Bertinho-base (21) and mBERT (25) are quite remarkable.

Table 2 shows the results obtained with different contextualized vectors derived from the dependency-based model (see subsection 4.2.2). In the first row, the table also shows a non-compositional baseline strategy just comparing the similarity of verb vectors out of context.

As it was the case with BERT models, the verb is also the most representative constituent of the meaning of the sentence: it achieves 47 and 41 correlation in the two directions, compared to only 18 and 15 for subject and object respectively. It follows that the verbal root, once contextualized by the sense of the arguments, can be taken as the meaning of the whole sentence.

Although they are not totally comparable, we also show in the last rows of the table the best values obtained by compositional systems applied to the English dataset described in (Grefenstette and Sadrzadeh, 2011b) and from which we have created the Galician one. The values obtained on the Galician dataset by using BERT-Base-Galician outperform all the compositional methods for English, including the highest score, 54, obtained by the system described in (Wijnholds, Sadrzadeh, and Clark, 2020).

Models	ρ
BERT-base-Galician - sentence	54
BERT-base-Galician- verb	**57**
BERT-base-Galician - subj	33
BERT-base-Galician - obj	49
BERT-small-Galician- sentence	**46**
BERT-small-Galician - verb	43
BERT-small-Galician - subj	28
BERT-small-Galician - obj	36
mBERT - sentence	21
mBERT - verb	**25**
mBERT - subj	23
mBERT - obj	24
Bertinho-base - sentence	9
Bertinho-base - verb	**21**
Bertinho-base - subj	6
Bertinho-base - obj	9

Table 1: Spearman correlation between different configurations of BERT and human judgments on 192 subject-verb-object sentence pairs.

Models	ρ
baseline - verb	28
left-to-right - sentence	37
left-to-right - verb	**47**
left-to-right - obj	15
right-to-left - sentence	32
right-to-left - verb	41
right-to-left - subj	18
Hashimoto and Tsuruoka (2014)	43 (en)
Polajnar et al. (2015)	35 (en)
Wijnholds et al. (2020)	54 (en)

Table 2: Spearman correlation between different configurations of the compositional dependency-based method and human judgments on 192 subject-verb-object sentence pairs (in lemmas). The table also shows a baseline based on just comparing verbs out-of-context (first row) and some related evaluations on a quite similar dataset for English (three last rows).

4.4 Discussion

The analysis of the results presented in the previous section leads us to draw some conclusions about the strategies compared in the experiments.

BERT-base-Galician and BERT-small-Galician models clearly outperform both Bertinho-base and mBERT in the proposed semantic task. It is also important to point out that the best scores of both Bertinho-base and mBERT are still below the baseline (28).

Concerning the dependency-based strategy, its results are comparable to those of BERT-small-Galician, even if they are far from the higher correlation given by BERT-base-Galician. Let us note that the training corpus of the dependency-based method, as well as that of Bertinho-base and mBERT (Galician part) is just the Galician Wikipedia, and this is much smaller than the training corpus used for the two BERT-Galician models: ≈42 million tokens vs. ≈500 million.

As it was already reported, the contextualized sense of the verbal root is the most

representative meaning of the sentence for most strategies. It behaves better than the other constituents (subject and object), and even than computing a global meaning for the whole sentence. This is a very relevant observation as most systems computing the sentence meaning do not know which is the root word because they do not rely on a dependency tree, and so they make use of the vectors of all constituent words.

And finally, we must point out that the correlation values obtained here and in other related experiments for other languages are in low to medium ranges. This shows that this is a semantic task of great complexity that still requires improved language models, perhaps not larger or computationally deeper, but of higher quality and with deeper linguistic knowledge.

5 Conclusions

In this article, we evaluated and compared the performance of contextualized vectors built with LLMs and a fully compositional strategy based on syntactic dependencies and selectional preferences. The use of selectional preferences to build contextualized vectors is a linguistically motivated attention strategy focused on selecting only syntactically relevant contextual elements. It can be seen, therefore, as a mechanism of attention driven by syntactic information.

According to the results obtained, the compositional strategy turned out to be competitive when compared to several configurations of BERT in a specific task focused on sentence similarity in Galician. It should be noted that the computational cost of training compositional models is much lower than that of neural-based LLMs. In addition, the syntax-based vectors we have used for the compositional approach are more transparent and interpretable than those derived from the Transformer architecture. Transparent models make it easier to explain the errors and successes committed in a particular task, since it is possible to explicitly list the syntactic contexts involved in vector composition.

However, the dependency-based strategy has important weaknesses. First, as it mainly relies on syntactic parsing, it has a vulnerable exposure to parser errors. Second, this strategy cannot be easily adapted to syntactically open sentences.

In order to overcome these drawbacks, in future work we will define and implement a syntax-based model allowing us to build fully contextualized vectors for open sentences. This will enable to apply the compostional method to any sentence in as similar way as Transformers do.

Acknowledgements

This research was funded by the project "Nós: Galician in the society and economy of artificial intelligence", agreement between Xunta de Galicia and University of Santiago de Compostela, and grant ED431G2019/04 by the Galician Ministry of Education, University and Professional Training, and the European Regional Development Fund (ERDF/FEDER program), and Groups of Reference: ED431C 2020/21. In addition: Ramón y Cajal grant (RYC2019-028473-I) and Grant ED431F 2021/01 (Galician Government).

References

Armendariz, C. S., M. Purver, S. Pollak, N. Ljubešić, M. Ulčar, M. Robnik-Šikonja, I. Vulić, and M. T. Pilehvar. 2020. SemEval-2020 task 3: Graded word similarity in context (GWSC). In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Asaadi, S., S. Mohammad, and S. Kiritchenko. 2019. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 505–516, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bai, J., Y. Wang, Y. Chen, Y. Yang, J. Bai, J. Yu, and Y. Tong. 2021. Syntax-BERT: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3011–3020, Online. Association for Computational Linguistics.

Baroni, M. 2013. Composition in distributional semantics. *Language and Linguistics Compass*, 7:511–522.

Baroni, M., R. Bernardi, and R. Zamparelli. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology (LiLT)*, 9:241–346.

Biemann, C. and M. Riedl. 2013. Text: Now in 2d! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95.

Bowman, S. R., G. Angeli, C. Potts, and C. D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-2019, Volume 1*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Erk, K. and S. Padó. 2008. A structured vector space model for word meaning in context. In *2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008*, pages 897–906, Honolulu, HI.

Ethayarajh, K. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *EMNLP/IJCNLP (1)*, pages 55–65. Association for Computational Linguistics.

Gamallo, P., M. Garcia, C. Piñeiro, R. Martinez-Castaño, and J. C. Pichel. 2018. LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Gamallo, P. 2017. Comparing explicit and predictive distributional semantic models endowed with syntactic contexts. *Language Resources and Evaluation*, 51(3):727–743.

Gamallo, P. 2019. A dependency-based approach to word contextualization using compositional distributional semantics. *Language Modelling*, 7(1):53–92.

Gamallo, P. 2021. Compositional distributional semantics with syntactic dependencies and selectional preferences. *Applied Sciences*, 11(12).

Gamallo, P., M. P. Corral, and M. Garcia. 2021. Comparing dependency-based compositional models with contextualized word embedding. In *13th International Conference on Agents and Artificial Intelligence (ICAART-2021)*.

Gamallo, P., S. Sotelo, J. R. Pichel, and M. Artetxe. 2019. Contextualized translations of phrasal verbs with distributional compositional semantics and monolingual corpora. *Computational Linguistics*, 45(3):395–421.

Gamer, M., J. Lemon, I. Fellows, and P. Singh, 2019. *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1.

Garcia, M. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3625–3640, Online, August. Association for Computational Linguistics.

Grefenstette, E. and M. Sadrzadeh. 2011a. Experimental support for a categorical compositional distributional model of meaning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 1394–1404.

Grefenstette, E. and M. Sadrzadeh. 2011b. Experimenting with transitive verbs in a discocat. In *Workshop on Geometrical Models of Natural Language Semantics (EMNLP 2011)*.

Kartsaklis, D. and M. Sadrzadeh. 2013. Prior disambiguation of word tensors for constructing sentence vectors. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1590–1601.

Lin, B. Y., S. Lee, X. Qiao, and X. Ren. 2021. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. *CoRR*, abs/2106.06937.

Mikolov, T., W.-t. Yih, and G. Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia.

Mitchell, J. and M. Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 236–244, Columbus, Ohio.

Mitchell, J. and M. Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439.

Nguyen, X.-P., S. Joty, S. C. H. Hoi, and R. Socher. 2020. Tree-structured attention with hierarchical accumulation.

Padró, M., M. Idiart, A. Villavicencio, and C. Ramisch. 2014. Nothing like good old frequency: Studying context filters for distributional thesauri. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 419–424.

Pilehvar, M. T. and J. Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Reimers, N. and I. Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Salazar, J., D. Liang, T. Q. Nguyen, and K. Kirchhoff. 2020. Masked language model scoring. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Shibayama, N., R. Cao, J. Bai, W. Ma, and H. Shinnou. 2020. Evaluation of pretrained BERT model by using sentence clustering. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 279–285, Hanoi, Vietnam, October. Association for Computational Linguistics.

Vilares, D., M. Garcia, and C. Gómez-Rodríguez. 2021. Bertinho: Galician BERT Representations. *Procesamiento del Lenguaje Natural*, 66:13–26.

Vulić, I., E. M. Ponti, R. Litschko, G. Glavaš, and A. Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online, November. Association for Computational Linguistics.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR 2019*.

Weir, D. J., J. Weeds, J. Reffin, and T. Kober. 2016. Aligning packed dependency trees: A theory of composition for distributional semantics. *Computational Linguistics*, 42(4):727–761.

Wijnholds, G., M. Sadrzadeh, and S. Clark. 2020. Representation learning for type-driven composition. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 313–324, Online. Association for Computational Linguistics.

Williams, A., N. Nangia, and S. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yu, L. and A. Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online, November. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 165-176 recibido 31-03-2022 revisado 11-05-2022 aceptado 13-05-2022

An Overview of Drugs, Diseases, Genes and Proteins in the CORD-19 Corpus

Una visión general de los Fármacos, Enfermedades, Genes y Proteínas en el corpus CORD-19

Carlos Badenes-Olmedo, Álvaro Alonso, Oscar Corcho
Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
{carlos.badenes, oscar.corcho}@upm.es
{alvaro.alonsoc}@alumnos.upm.es

Abstract: Several initiatives have emerged during the COVID-19 pandemic to gather scientific publications related to coronaviruses. Among them, the COVID-19 Open Research Dataset (CORD-19) has proven to be a valuable resource that provides full-text articles from the PubMed Central, bioRxiv and medRxiv repositories. Such a large amount of biomedical literature needs to be properly managed to facilitate and promote its use by health professionals, for example by tagging documents with the biomedical entities that appear on them. We created a biomedical named entity recognizer (NER) that normalizes (NEN) the drugs, diseases, genes and proteins mentioned in texts with the codes of the main standardization systems such as MeSH, ICD-10, ATC, SNOMED, ChEBI, GARD and NCBI. It is based on fine-tuning the BioBERT language model independently for each entity type using domain-specific datasets and an inverse index search to normalize the references. We have used the resultant BioNER+BioNEN system to process the CORD-19 corpus and offer an overview of the drugs, diseases, genes and proteins related to coronaviruses in the last fifty years.
Keywords: ner, normalization, bioentities, document retrieval.

Resumen: Durante la pandemia del COVID-19 han surgido varias iniciativas para recopilar publicaciones científicas relacionadas con el coronavirus. Entre ellos, el conjunto de datos de investigación abierta sobre COVID-19 (CORD-19) ha demostrado ser un recurso valioso que proporciona el texto completo de artículos extraídos de los repositorios PubMed Central, bioRxiv y medRxiv. Una cantidad tan grande de literatura biomédica debe gestionarse adecuadamente para facilitar y promover su uso por parte de los profesionales de la salud, por ejemplo, etiquetando documentos con las entidades biomédicas que aparecen mencionadas. Hemos creado un reconocedor biomédico de entidades nombradas (NER) que normaliza (NEN) los fármacos, enfermedades, genes y proteínas mencionados en textos con los códigos de los principales sistemas de estandarización como MeSH, ICD-10, ATC, SNOMED, ChEBI, GARD y NCBI. Se basa en afinar el modelo de lenguaje BioBERT de forma independiente para cada tipo de entidad utilizando conjuntos de datos específicos de dominio y una búsqueda de índice inverso para normalizar las referencias. Hemos utilizado el sistema BioNER+BioNEN resultante para procesar el corpus CORD-19 y ofrecer una visión general de los fármacos, enfermedades, genes y proteínas relacionados con el coronavirus en los últimos cincuenta años.
Palabras clave: identificación de entidades, normalización, bio-entidades, recuperación de documentos.

1 Introduction

Several initiatives have emerged during the COVID-19 pandemic to gather scientific publications related to coronaviruses. The COVID-19 Data Portal[1], maintained by the EU, or the Humandata[2], focused on COVID-

[1] https://www.covid19dataportal.org
[2] https://data.humdata.org/event/covid-19

ISSN 1135-5948. DOI 10.26342/2022-69-14

19 cases around the world, are some examples. The Allen Institute for Artificial Intelligence created the COVID-19 Open Research Dataset (CORD-19)(Wang et al., 2020). It is a continuously growing corpus with all publicly available COVID-19 and coronavirus-related research (e.g. SARS, MERS, etc.) published during the last fifty years, with a huge increase in the last two years. This dataset provides full-text research papers in PDF and JSON format, which can be used as a source of information to extract knowledge related to the infection and disease. At the time of this study (January 2022), it is composed of 334,572 scientific articles retrieved from PubMed Central, a corpus maintained by the World Health Organization (WHO), bioRxiv and medRxiv pre-prints.

Such a large amount of biomedical literature needs to be properly managed to facilitate and promote its use by health professionals. Natural Language Processing (NLP) facilitates document analysis through the extraction of key information from the underlying texts and turning them into structured knowledge that can be understood by humans (Pyysalo et al., 2007). One of the main NLP tasks is the recognition of relevant entities found in texts, what is commonly known as Named Entity Recognition (NER)(Nadeau and Sekine, 2007). This task enables the exploration of texts guided by key terms and the discovery of relationships between them. The NER task identifies meaningful terms in a domain, called *named entities*, and classifies them into predefined entity classes (Li et al., 2020). In the biomedical domain, these entities are medical concepts such as drugs, diseases, or gene mutations, and the task is more specifically known as BioNER. Entities can also be classified according to existing taxonomies to avoid ambiguity, in a task that is commonly known as Entity Linking or Named Entity Normalization (NEN) and, when applied in the biomedical domain, BioNEN (Campos, Matos, and Oliveira, 2012).

The main objective in BioNEN is to use controlled and curated biomedical vocabularies such as Medical Subject Headings (MeSH) [3] codes or the Anatomical Therapeutic Chemical (ATC)[4] classification system, to reduce ambiguities and to extend the information about the entities. Once the entity recognition and normalization tasks are applied in biomedical literature, a set of normalized concepts can be used by Information Retrieval processes, such as the creation of efficient search algorithms, content classification, or Knowledge-Graph construction among others (Chatterjee et al., 2021). These processes play a key role in subsequent NLP tasks such as Question-Answering, Relation Extraction, Knowledge-base population, or Semantic search (Nadeau and Sekine, 2007).

However, the biomedical language entails some challenges in identifying entities (Zhou et al., 2004): (1) *highly specialized terms* (i.e. most of the terms are exclusive of these kinds of texts, making it difficult to reuse general domain knowledge to identify and classify specific domain concepts), (2) *sharing of nouns* (e.g. "5kb and 17kb viruses" refers to "5kb viruses and 17kb viruses"), and (3) *non-standardized naming convention* (e.g. "$N-acetyl-\beta-D-glucosamine$", "$N-Acetylglucosamine$", and "$C_{18}H_{15}NO_6$" refers to the same concept).

This article describes how we performed BioNER and BioNEN tasks on the CORD-19 corpus, and our analysis of the presence of diseases, drugs, genes and proteins in their texts. Our main contributions are:

- A BioNER+BioNEN system based on independently fine-tuned BioBERT models to identify diseases, drugs and genes/proteins from technical texts.[5]
- A collection of scientific texts tagged with normalized terms and codes of diseases, drugs, and genes/proteins[6]. (Badenes-Olmedo, Alonso, and Corcho, 2022)
- A statistical analysis of the presence of biomedical entities in the January 2022 edition of the CORD-19 corpus.

The paper is structured as follows: Section 2 review the state-of-the-art methods to identify biomedical entities and present our approach. The normalization process that we have followed is described in section 3. Section 4 details how the CORD-19 corpus has been processed, and show and discuss the results. Final remarks and future work are presented in section 5.

[3] https://www.nlm.nih.gov/mesh

[4] https://www.whocc.no

[5] https://github.com/drugs4covid/bio-ner

[6] https://doi.org/10.5281/zenodo.6532473

2 *Biomedical Named Entity Recognition*

NER tasks usually follow the pipeline showed in Fig. 1. The text is firstly pre-processed depending on the requirements of subsequent processes (e.g. word cleaning, stemming, verb tense normalization, etc). Afterwards, a representation of the words which compose a text span is made, what serves as an input to a NER model which performs the classification of these features to assign tags to the words. Sometimes, in order to refine the results, a post-processing step is also required to extend or group the entities. Biomedical-NER (BioNER) specializes the NER classification task for the medical domain and, sometimes, particularizes the techniques used to characterize texts. A BioNER method, depending on the type of technique used to classify terms, can be organized into: *Rule/Dictionary-based* (i.e requires domain knowledge to define patterns of the different sorts of named entities to characterize them), *Machine Learning-based* (i.e. discovers rules through automatic patterns and reduces the need for domain knowledge) and Hybrid approaches (i.e. combines methods to leverage benefits from different approaches) (Li et al., 2020) (Perera, Dehmer, and Emmert-Streib, 2020)(Yadav and Bethard, 2019).

2.1 Our Approach

We have created a hybrid system for the recognition and normalization of biomedical entities based on state-of-the-art methods. Our model specializes the BioBERT (Lee et al., 2020) model pretrained with millions of scientific and biomedical articles, with additional training corpora to extend the BioNER task and to cover the BioNEN task from multiple external standardization databases.

The entity classes considered for our system were the most widely used classes in BioNER modelling and the ones with a higher number of corpora available for a fine-tuning process (see Table 1). It is important to note that these biomedical entities can be mentioned in different ways and this further makes it more difficult to achieve a correct recognition and normalization. Variations can be trivial names (e.g. *water*), technical names (.e.g *lung infection with Mycoplasma pneumoniae* to refer to *Bacterial Pneumonia*), brands (e.g. *Veklury®*), systematic IUPAC names (e.g. *2,5,5-trimethyl-2-hexene*),

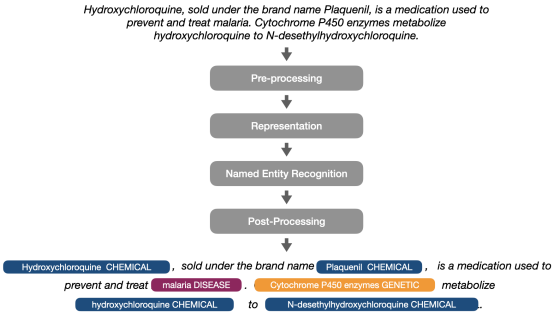


Figure 1: NER pipeline.

generic names (e.g. *Benzenes*), molecular formulas (e.g. CH_3), abbreviated forms (e.g. *DMA* for *dimethylacetamide*) and identifiers of curated databases such as ChEBI[7] (e.g. *145994*).

For each entity class (i.e. disease, drugs and gene/proteins), a different BioNER model was created (see Fig. 2). One model was fine-tuned to recognize disease entities, another for chemical (i.e. drugs) entities and another for genes/proteins. We adopted this strategy because it has proven to behave better for fine-tuned tasks than combining several entity classes in the same task in only one model. The more specific the model is, the better results will be usually obtained for a specific task (Gururangan et al., 2020). Separate models capture better patterns within each of the entity classes allowing to maximize its tagging performance, resulting in a system with the better model possible for each of the entities. Our system offers slightly lower performance than BioBERT model because we jointly use several datasets for fine-tuning. The aim is to increase the ability to identify as many entities as possible, even at the cost of penalizing the accuracy of the model, since our pipeline incorporates an additional normalization step where the entities will be filtered out. The post-processing tasks are based on an inverse index search. The architecture of the system is described in Fig. 2 and further details about each of the components are revised along the following sections.

2.2 Datasets

We have added an untrained fully-connected layer on top of a BioBERT model to perform the fine-tuning. At least three fine-tuning processes have been done to cover the

[7] `https://www.ebi.ac.uk/chebi`

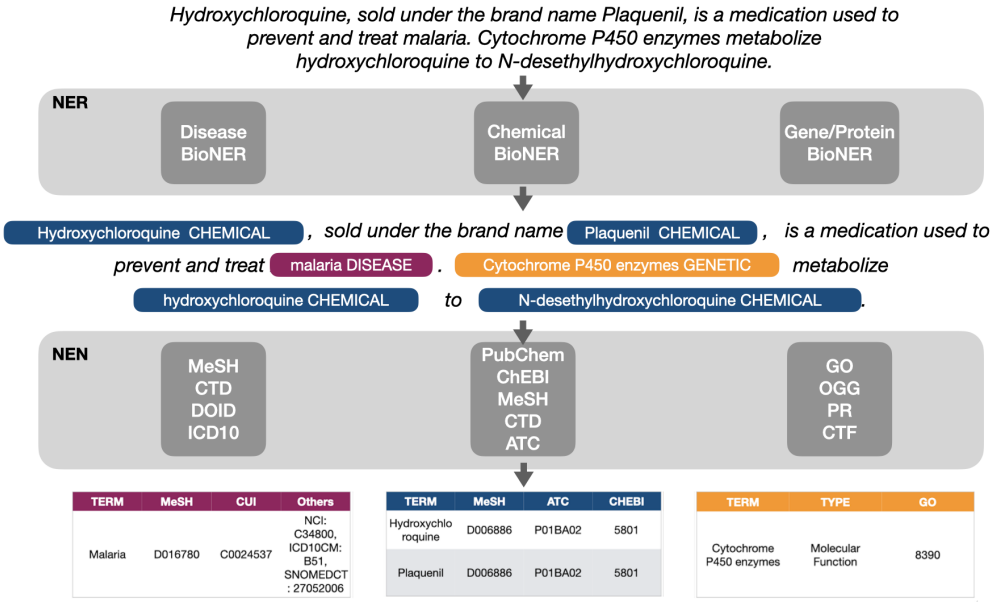


Figure 2: Overview of our BioNER+BioNEN architecture.

three different entities (i.e diseases, drugs and genes/proteins). The corpora used for each training process were selected from existing datasets (see Table 1) according to criteria based on data volume and quality. Furthermore, since the techniques used to identify entities slightly differ between the existing datasets despite being the same target entity, we use two corpora for each entity class to supply the model with a better generalization capacity in situations where a never-seen text is used (Lee et al., 2020).

2.2.1 Diseases Dataset

The BC5CDR-Diseases dataset (Li et al., 2016), with around 13,000 annotations, and the NCBI-Diseases dataset (Doğan, Leaman, and Lu, 2014), with almost 7,000, were the corpora used to recognize diseases. These datasets are the most widely used in BioNER tasks for disease entities and most models provide results for each of them, including BioBERT which obtained an F1-score of 89.71 for NCBI-Diseases and 87.15 for B5CDR-Diseases. Our model, created by combining both datasets during the fine-tuning process, offers a slightly lower performance with a F1-score of 87.4 and 85.8, respectively. This is likely because of the hyperparameter search intensity and because the number of epochs done is lower.

2.2.2 Drugs Dataset

For chemical entities, the two selected datasets were BC4CHEMD (Krallinger et al., 2015) and BC5CDR-Chemicals (Li et al., 2016) with around 80,000 and 15,000 entities respectively. The largest annotated corpus, BioSemantics (Akhondi et al., 2014), was not considered since it is based on patents which could slightly differ from biomedical articles that are the kind of texts in which our system is focused on. The selected datasets were also the most widely adopted corpora for NER tasks in chemical entities and most models provide performance results for them. BioBERT obtained state-of-the-art results in BC4CHEMD with an F1-score of 92.36 and the second best result for BC5CDR with 93.47, which is almost the same than the state-of-the-art result obtained by BlueBERT (Peng, Yan, and Lu, 2019) which was 93.5 . Our system obtained F1 results of 91.7 for BC4CHEMD and 92.99 for BC5CDR-Chemicals.

2.2.3 Gene and Proteins Dataset

Gene and protein entities were jointly considered since they belong to similar semantic types. This consideration is widely adopted in most existent corpus, which consider them together (Goyal, Gupta, and Kumar, 2018). The pair of selected datasets were JNLPBA (Kim et al., 2004) and BC2GM (Smith et al., 2008), which offer around 35,000 and 25,000 annotations respectively. The CRAFT corpus (Bada et al., 2012), which is the largest Gene/Protein NER corpus, was discarded since most models report results

Year	Reference	Corpus Name	Entities	# Annotations	# Tokens
2004	(Kim et al., 2004)	JNLPBA	Genes/Proteins	35460	597333
			Cell Lines	4332	
2008	(Smith et al., 2008)	BC2GM	Genes/Proteins	24583	508257
2012	(Bada et al., 2012)	CRAFT	Chemicals	8137	560000
			Genes/Proteins	49961	
			Species	7449	
			Cell Lines	5760	
2013	(Pafilis et al., 2013)	Species-800	Species	3646	195197
2013	(Ohta et al., 2013)	BioNLP13CG	Species	21683	129878
			Anatomy		
			Genes/Proteins		
2013	(Segura Bedmar, Martínez, and Herrero Zazo, 2013)	SemEval2013 - DrugBank	Chemicals	15745	≈ 65000
2013	(Segura Bedmar, Martínez, and Herrero Zazo, 2013)	SemEval2013 - Medline	Chemicals	2746	≈ 20000
2013	(Pyysalo et al., 2015)	BioNLP13PC	Genes/Proteins	15901	108356
			Chemicals		
2014	(Bagewadi et al., 2014)	mi-RNA	Genes/Proteins	1006	65998
			Species	726	
			Diseases	2123	
2014	(Akhondi et al., 2014)	BioSemantics	Chemicals	386110	5690518
2014	(Pyysalo and Ananiadou, 2014)	AnatEM	Anatomy	13000	250000
2014	(Doğan, Leaman, and Lu, 2014)	NCBI Disease	Diseases	6881	174487
2015	(Goldberg et al., 2015)	LocText	Species	276	22550
2015	(Krallinger et al., 2015)	BC4CHEMD	Chemicals	79842	2235435
2016	(Li et al., 2016)	BC5CDR	Diseases	12694	323281
			Chemicals	15411	
2016	(Kaewphan et al., 2016)	CLL	Cell Lines	341	6547
		Gellus		640	278910
2020	(Legrand et al., 2020)	PGxCorpus	Diseases	635	≈ 35000
			Chemicals	1718	
			Genes/Proteins	1708	

Table 1: Corpora with biomedical entities.

based on those corpus and a comparison between them can be established. BioBERT reported state-of-the-art results on BC2GM results with a F1 of 84.72 and in JNLPBA results (77.59) were slightly worse than state-of-the-art which were reported by PubMedBERT with a F1 of 80.06. Results from our fine-tuning model were a bit worse with 83.0 and 76.0 for BC2GM and JNLPBA respectively. Results on this joint entity class are significantly worse than other entity classes, perhaps due to the broad range of subentity classes which take part within this class. This makes the amount of linguistic variability larger, and hence harder to capture than the former entity classes.

Once the models have been fine-tuned, we require some additional steps before having a homogeneous representation of the entities (see Fig. 2). The following section details the entity normalization process and the additional tasks required in our NER+NEN pipeline (Fig. 1).

3 Entity Normalization

The normalization process has been addressed through an inverted index search. Each entity is associated with a set of related terms extracted from external coding systems. Once the medical term is recognized, we search for entities that contain that term in any of their related fields, and we sort that set of candidates based on the BM25 ranking function (Robertson et al., 1994). Those with fewer related terms will have greater relevance. Each type of entity has its own database (i.e index). This way, indexes can be built separately with curated and related terms that helps to map concepts with terms and codes (see Table 2). Multiple sources were taken into account in each of the entity classes, mainly from BioPortal ontologies[8], but also from the Comparative Toxicogenomics Database[9] and PubChem[10].

[8] `http://bioportal.bioontology.org`

[9] `http://ctdbase.org`

[10] `https://pubchem.ncbi.nlm.nih.gov`

Type	Entities	Codes	Sources
Diseases	126573	5	4
Drugs	344238	7	5
Genes	946584	3	4

Table 2: Resources used for normalization.

For each entity, regardless of whether it is a drug, disease or gene/protein, the following information was collected: (1) a *term* or *description* of the underlying concept (e.g. "*Hydroxychloroquine*") ; (2) a list of *synonyms* that holds all possible related words present for a given term (e.g. '*Oxichloroquine*', '*Polirreumin*'); (3) a *semantic type* (e.g. '*Pharmacologic Substance*') and (4) a list of *identifiers* based on MeSH, CUI, ATC, or any other more specific database cross references (e.g. *mesh_id:D006886*, *cid:3652*, *atc:P01BA02*). The range of possibilities to refer to the same element (i.e by code, term or synonym) allow choosing the one with the higher score between different search criteria (e.g. terms or synonyms; strict or similar matches) and filtering criteria (e.g. based on word order, or single terms). The result with the higher score is considered.

3.1 Diseases

Four different sources were merged in the same index to normalize disease terms based on the mappings between their codes and the medical terms used to represent them.

MeSH - Diseases: Medical Subject Headings[11] is a thesaurus with hierarchical and controlled vocabulary produced by the National Library of Medicine (NLM). This thesaurus includes thousands of terms regarding to several semantic types with disease-related terms among them. BioPortal includes an ontology version of this thesaurus from which we have extracted disease-related terms attending to the UMLS Semantic Type each term belongs to.

CTD - Diseases: CTD's MEDIC disease vocabulary is a modified subset of the "Diseases" branch of the NLM's MeSH, combined with genetic disorders from the Online Mendelian Inheritance in Man[12] (OMIM) database. These terms have been merged with the previous ones through an outer join on MeSH IDs.

DOID: The Human Disease Ontology (Schriml et al., 2012) is a comprehensive knowledge base of inherited, developmental and acquired human diseases. It integrates terms from a wide range of medical vocabularies such as MeSH, SNOMED, NCI, or OMIM, and has been used to extend terms which were not previously captured by the other sources. The way this was done is through an outer join on MeSH IDs.

ICD-10-CM: The International Classification of Diseases is a hierarchical classification listed by the World Health Organization (WHO), in which are encoded a wide range of signs, symptoms, abnormal findings, causes of damage, diseases, and/or other disease-related terms. The ICD-10-CM is the 10th version of this classification with a Clinical Modification of the source. Since this classification is used in its proper BioPortal ontology, further mapping concepts are added, which is the case of Unified Medical Language System identifiers (CUIs). The way this source extends the previous sources is through this CUI since not MeSH IDs are included. For that purpose, an outer join on this id was done.

3.2 Drugs

Five sources were considered to merge chemical terms in a shared index. The main objective was to capture the wide range of possible chemical mentions that this entity class can support

PubChem: PubChem is the world largest chemistry open database maintained by the National Institute of Health (NIH). Among the classification systems offered to organize the chemical entities, we used the MeSH hierarchy for our database. Approximately 130000 terms were considered which is expected to have the most widely adopted chemical terms within all the collection.

ChEBI: ChEBI is a chemical database mainly focused on small chemical components of molecular entities and therefore it complements other types of terms considered in the rest of sources. Any biological or synthetical component present in biological organisms is aimed to be captured on this database. An outer join on *InChIKey* was used for connecting these terms with the ones present in the previous source. *InChIKey* is a hashed key of *InChI*, an International Identifier for chemicals, which offers an IUPAC identifier for an standardized codification of

[11] https://www.nlm.nih.gov/mesh

[12] https://www.omim.org

chemicals.

MeSH - Chemicals: MeSH also includes thousands of terms regarding to chemical-related terms. The ontology version in BioPortal has been used to extract chemical-related terms attending to the UMLS Semantic Type each term belongs to. Since PubChem already includes MeSH terms, this source has been just used to add MeSH IDs and extend information from the previous terms. This source was combined with the previous ones through checking if the term is found either on term field or on the synonyms list. If it is not found, it has been appended to chemical terms.

CTD - Chemicals: Database that incorporates terms from multiple chemical sources and therefore it has been used for complementing previously existent processed terms. It also helps to extend the retrieved information about previously considered terms. Non previously found terms have been appended from this source.

ATC: Classification of pharmacological substances organized in therapeutic levels. The ontology version of BioPortal has been the source considered for ATC since it incorporates further information and relations with other terms. Information regarding ATC level and ATC code was added to the previously considered terms. If the term is not present, it has been appended.

3.3 Genetics

This entity class is composed of a broad semantic type since it includes both gene and proteins-related terms. They are close semantic types and even in some occasions the use of the same expressions is diffuse. This has led to a wide range of terms within this entity class in which four large and complementary sources were merged in the same index to cover the biggest amount of entity variability possible.

GO: The knowledgebase underlying the Gene Ontology (Ashburner et al., 2000) is the largest source for the functions of genes and therefore it has been used aiming to capture terms related to genetic mechanisms.

OGG: The Ontology of Genes and Genomes (He, Liu, and Zhao, 2014) collects genes and genomes of certain organisms such as humans, virus and bacteria. Mappings to multiple sources are found in the BioPortal ontology.

	Entities	Coverage (%)	Normalization (%)
Diseases	18,355	49.4	4.2
Drugs	55,120	15.1	22.3
Genes/Proteins	79,063	16.6	9.1

Table 3: CORD-19 statistics (January-2022). Total number of appearances (*Entities*), diversity (*Coverage*) and standardization (*Normalization*) ratio.

PR: The Protein Ontology (Natale et al., 2017) contains a wide range of protein-related entities along with relations between them. This source contains a large amount of terms that covers the protein part.

CTD - Genes: It contains a vocabulary retrieved from multiple sources with a great variety of genes in multiple species. It has been used to extend the gene terms which were not previously captured, appending non-retrieved genes.

4 CORD-19 Entities

The BioNER+BioNEN system described in this paper was used to identify and normalize the drugs, diseases and genetic-related terms mentioned in the CORD-19 corpus (January 2022 Edition). The recognition process was time consuming (approximately 48 days) in a server composed by a 32 CPU-cores Intel Xeon with 256GB RAM. The lack of GPUs made the process considerably slower (i.e. 1173hours at a rate of 0,4s/task) since it requires matrix computation for the transformer-based language models, one for each biomedical concept. The source code is publicly available [13].

Entity recognition and normalization was done for each paragraph of the scientific article. A first group of labels is created to identify the medical terms as they appear in the text (i.e. *diseases_ss*, *chemicals_ss*, *genetics_ss*), and in a standardized way (i.e. *disease_terms_ss*, *chemical_terms_ss*, *genetic_terms_ss*). In the case of diseases and genes/proteins, a predefined category is also established during the normalization process (i.e. *disease_types_ss*, *genetic_types_ss*). The following group of labels contains the codes for each of the classification systems described in Section 3 (i.e. *mesh_codes_ss*, *atc_codes_ss*, *cid_codes_ss*, *doid_codes_ss*, *cui_codes_ss*, *icd10_codes_ss*, *icd9_codes_ss*,

[13] https://github.com/drugs4covid/cord-19

gard_codes_ss, *snomed_codes_ss*, *nci_codes_ss*, *ncbi_codes_ss*, *uniprot_codes_ss*). The suffix *_ss* in all tags indicates that the format is a textual list (i.e. string sequence).

Table 3 shows some statistics about entity classes once the corpus was processed. As expected, almost half of the paragraphs contain at least one mention of a disease or symptom (see column *Coverage*), while drugs, genes or proteins appear less frequently. This is strongly influenced by the criteria used by Allen AI to create the CORD-19 corpus, as they filter articles that contain coronavirus-related terms in their title or abstract. This guide the content of the article and also explains why the variety of disease and symptom entities (see column *Entities*) is far inferior to drugs and genetic information. However, what is striking is the high rate of standard terms (according to our model) used to refer to drugs, with respect to the rest of biomedical entities. Column *Normalization* shows the ratio of entities mentioned in the text using any of the terms extracted from the classification systems described in section 3. We think that there is more flexibility in scientific texts to refer to symptoms or diseases than to drugs or active ingredients, with respect to the standards (e.g. ATC, MeSH, ICD-10 or SNOMED mainly). Regarding genetic information, perhaps the cause lies in the precision in the recognition of the boundary that defines the entity, being sometimes eliminated part of the chemical expression of the entity itself.

Table 4 shows the most widely captured entities according to the following classification systems: ICD-10, MeSH, ChEBI, ATC, MedGen (CUID), GARD, NCBI and SNOMED. Jointly with the code and description of the entity, the occurrences of these words are given (column *Ratio*). This allows us to have an idea about the relevance of the concept in the corpus with respect to the rest of the concepts of the same classification system. In top positions we can find general concepts related to respiratory difficulties. As we go down in the top, more specific terms begin to appear. In the systems that cover diseases such as MeSH or ICD-10, we can find as the most relevant concept the COVID-19 disease, as expected, and the related symptoms (e.g. U07.1 in ICD-10, D000086382 in MeSH or C5203670 in MedGen). The systems more oriented to chemicals identify substances related to respiratory disorders (e.g. Dioxyegn in ChEBI or Oxygen in ATC). And the systems focused on genetic and protein information show, with similar relevance, the pathways of the coronavirus (e.g. Angiotensin converting enzyme 2, Interleukin-6 or Interferon in NCBI).

Thanks to the normalization process that we incorporate in our entity recognition system, we can use the hierarchies defined in the underlying classification system to establish more or less general labels. For example, the Anatomical Therapeutic Chemical (ATC) classification system, which is supported by the World Health Organization (WHO) and widely used in hospital pharmacies to identify drug components, organizes active substances according to the organ or system on which they act and their therapeutic, pharmacological, and chemical properties. Drugs are classified into groups at five different levels. The first one corresponds to main groups, the second one to pharmacological or therapeutic subgroups, the third and the fourth one are chemical-pharmacological-therapeutic subgroups and the last one is the chemical substance. Once the code of a drug has been identified in this classification system, we can extend the labels of the text with those groups of the hierarchy, enabling additional ways of exploiting the results of the annotation process.

In the following experiment we want to take advantage of the labels generated by our system to find evidence about the anatomical behavior of drugs used to treat coronavirus. We do not know a priori which groups of drugs are related in this domain, and we assume that an evidence implies the joint presence of several groups in the same paragraph. Since the ATC classification system is hierarchical and establishes 14 anatomic groups at the first level of drug organization, we can create a matrix with the paragraphs where drugs are mentioned and the anatomic groups to which they belong. Figure 3 shows the correlation between each of these anatomical groups based on analysis of mentions of drugs in texts. It can be seen how the highest correlation exists between drugs associated with *Sensory organs* and *Anti-infectives for systemic use*. This may be due to the fact that many of the anti-infective active substances used systemically (i.e. orally or intravenously) are

		Entity	
	Ratio	**Code**	**Description**
ICD-10	51.1	U07.1	COVID-19
	8.0	J12.81	Pneumonia due to SARS-associated coronavirus
	3.7	J11.1	Influenza due to unidentified influenza virus with other respiratory manifestations
	3.3	A79.0	Trench fever
	3.2	F53.0	Postpartum depression
MeSH	34.0	D000086382	COVID-19
	11.1	D000085343	Latent Infection
	4.5	D018352	Coronavirus Infections
	3.9	D003643	Death
	3.7	D045169	Severe Acute Respiratory Syndrome
ChEBI	8.8	15379	Dioxygen
	5.2	33708	Amino-acid Residue
	3.8	172234	TG(14:1(9Z)/22:5(7Z,10Z,13Z,16Z,19Z)/24:1(15Z))
	3.2	30879	Alcohol
	2.9	5801	Hydroxychloroquine
ATC	14.6	V03AN01	Oxygen
	6.4	V04CA02	Glucose
	4.9	P01BA02	Hydroxychloroquine
	3.0	A12AA	Calcium
	2.8	V03AN04	Nitrogen
MedGen (CUI)	40.6	C5203670	COVID-19
	13.2	C0872054	Latent Infection
	6.4	C1175175	Severe acute respiratory syndrome
	4.3	C3714514	Infection
	3.0	C0003467	Anxiety
GARD	26.9	9237	SARS
	11.7	5698	Acute respiratory distress syndrome
	5.9	6427	Farmer's lung
	4.1	2035	Lymphatic filariasis
	2.9	6254	Dengue fever
NCBI	6.6	59272	Angiotensin converting enzyme 2
	5.5	100628202	Interleukin-6
	4.5	100304604	Interferon
	4.3	101180090	Immunoglobulin G level
	4.0	7124	tumor necrosis factor
SNOMED	57.4	840539006	Disease caused by 2019 novel coronavirus (disorder)
	6.3	398447004	Severe acute respiratory syndrome (disorder)
	4.2	155559006	Influenza (disorder)
	4.1	266391003	Pneumonia and influenza or pneumonia (disorder)
	3.8	82214002	Trench fever (disorder)

Table 4: Presence (*Ratio*) of the most frequent entities (*Code*) organized by coding system.

also categorized within the sensory organs group, for example the Ciprofloxacin, since they can also be administered by the otic or ophthalmic route. Thanks to the tags created by our system, it is sufficient to filter the paragraphs labeled with the ATC codes 'S' (i.e *Sensory organs*) and 'J' (i.e. *Anti-infectives*) to find the candidates for evidence. The other most notable correlation is between *Anti-infectives for systemic use* and *Anti-parasitic products*. It could be explained because the anti-infective drugs used for parasites are classified as anti-parasitic products, and the active substances most used experimentally for the treatment of coronavirus were found within these categories, such as Lopinavir/Ritonavir (anti-infective) and Hydroxychloroquine (anti-parasitic). Again, we can take advantage of the tags in our system to find texts in the articles that help us validate this assumption.

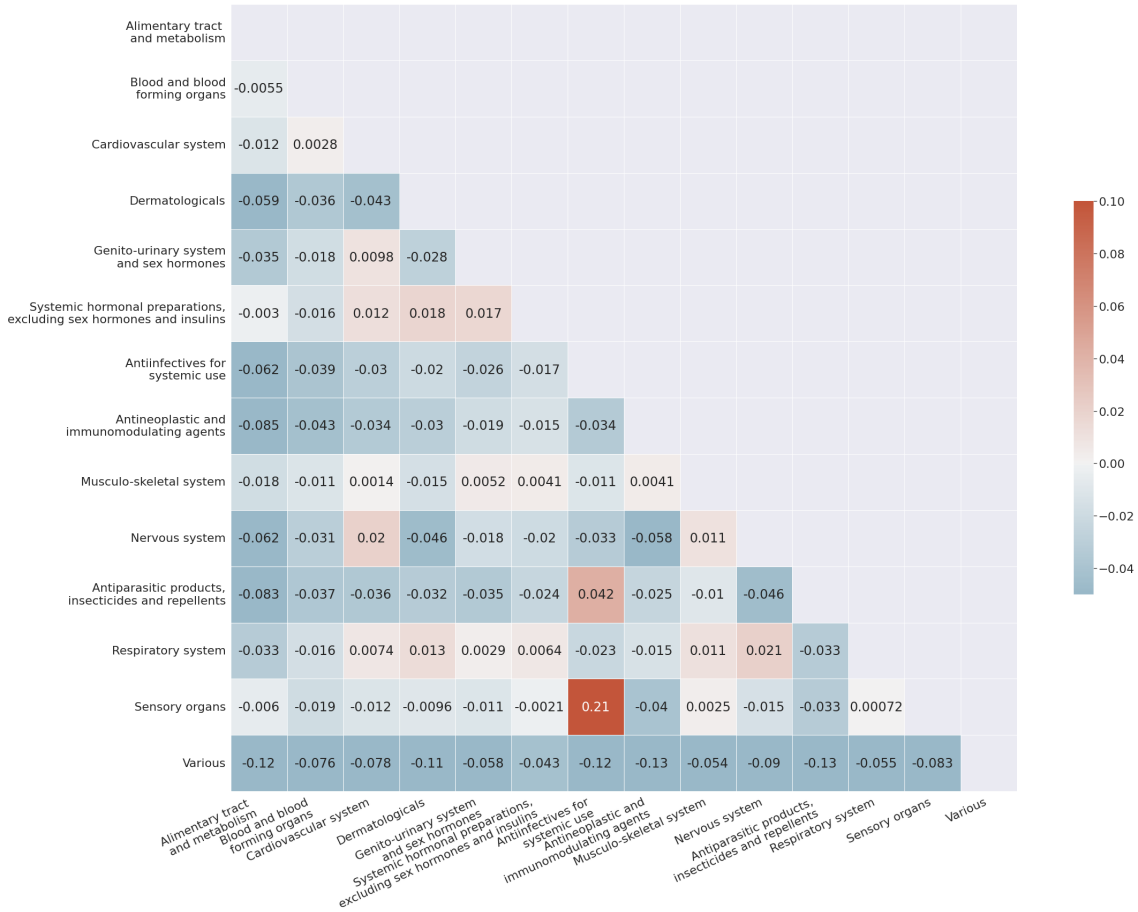


Figure 3: Correlation matrix of ATC data at Anatomical group level.

5 *Conclusions*

We have created a corpus with the diseases, drugs, genes, and proteins mentioned in the paragraphs of the articles in the January edition of the CORD-19 corpus. It contains not only the biomedical entities, but also their normalized references based on several curated databases such as MeSH, ICD-10, ATC, ChEBI or SNOMED. The generated corpus is publicly available and is updated periodically to take up changes in the CORD-19 dataset.

An analysis has been carried out on this corpus to measure the presence and degree of normalization of each type of biomedical entity. As expected, practically half of the paragraphs contain some reference to a disease or symptom. However, only 4% of them were mentioned using any of the standard codes or alias. The behavior in genes and proteins is similar although much lower in terms of presence. Drugs are the least present and most varied type of entity in the corpus. The correlation between the anatomical groups of the drugs has also been measured to value the usefulness of the tags created. The procedure to easily extract the evidence, i.e. paragraphs where the groups are mentioned, is also described.

Our biomedical named entity recognizer created to produce the tags is also described. It is based on the pre-trained BioBERT language model and combines three different models each of them specialized in the recognition of a different biomedical entity: disease, drug and gene/protein. In the future we want to explore the ability of the tags to produce knowledge, either to organize entities or to discover relationships that may arise between them, and to take advantage of the knowledge acquired to create a Spanish BioNER+BioNEN model.

Acknowledgments

Work supported by the *DRUGS4COVID++* project , financed by *Ayudas Fundación BBVA a equipos de investigación científica SARS-CoV-2 y COVID-19*.

References

Akhondi, S. A., A. G. Klenner, C. Tyrchan, A. K. Manchala, K. Boppana, D. Lowe, M. Zimmermann, S. A. Jagarlapudi, R. Sayle, J. A. Kors, et al. 2014. Annotated chemical patent corpus: a gold standard for text mining. *PloS one*, 9(9):e107477.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Bada, M., M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. A. Baumgartner, K. B. Cohen, K. Verspoor, J. A. Blake, et al. 2012. Concept annotation in the craft corpus. *BMC bioinformatics*, 13(1):1–20.

Badenes-Olmedo, C., A. Alonso, and O. Corcho. 2022. Drugs, Diseases, Genes and Proteins in the CORD-19 Corpus, March.

Bagewadi, S., T. Bobić, M. Hofmann-Apitius, J. Fluck, and R. Klinger. 2014. Detecting mirna mentions and relations in biomedical literature. *F1000Research*, 3.

Campos, D., S. Matos, and J. L. Oliveira. 2012. Biomedical named entity recognition: a survey of machine-learning tools. *Theory and Applications for Advanced Text Mining*, 11:175–195.

Chatterjee, A., C. Nardi, C. Oberije, and P. Lambin. 2021. Knowledge graphs for covid-19: An exploratory review of the current landscape. *Journal of Personalized Medicine*, 11(4).

Doğan, R. I., R. Leaman, and Z. Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Goldberg, T., S. Vinchurkar, J. M. Cejuela, L. J. Jensen, and B. Rost. 2015. Linked annotations: a middle ground for manual curation of biomedical databases and text corpora. In *BMC proceedings*, volume 9, pages 1–3. BioMed Central.

Goyal, A., V. Gupta, and M. Kumar. 2018. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21–43.

Gururangan, S., A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

He, Y., Y. Liu, and B. Zhao. 2014. Ogg: a biological ontology for representing genes and genomes in specific organisms. In *ICBO*, pages 13–20. Citeseer.

Kaewphan, S., S. Van Landeghem, T. Ohta, Y. Van de Peer, F. Ginter, and S. Pyysalo. 2016. Cell line name recognition in support of the identification of synthetic lethality in cancer from text. *Bioinformatics*, 32(2):276–282.

Kim, J.-D., T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75. Citeseer.

Krallinger, M., O. Rabal, F. Leitner, M. Vazquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D. Ji, D. M. Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Legrand, J., R. Gogdemir, C. Bousquet, K. Dalleau, M.-D. Devignes, W. Digan, C.-J. Lee, N.-C. Ndiaye, N. Petitpain, P. Ringot, et al. 2020. Pgxcorpus, a manually annotated corpus for pharmacogenomics. *Scientific data*, 7(1):1–13.

Li, J., Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, and Z. Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Li, J., A. Sun, J. Han, and C. Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Nadeau, D. and S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Natale, D. A., C. N. Arighi, J. A. Blake, J. Bona, C. Chen, S.-C. Chen, K. R. Christie, J. Cowart, P. D'Eustachio, A. D. Diehl, et al. 2017. Protein ontology (pro): enhancing and scaling up the representation of protein entities. *Nucleic acids research*, 45(D1):D339–D346.

Ohta, T., S. Pyysalo, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, S. Ananiadou, and J. Tsujii. 2013. Overview of the pathway curation (pc) task of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 67–75.

Pafilis, E., S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen. 2013. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6):e65390.

Peng, Y., S. Yan, and Z. Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Perera, N., M. Dehmer, and F. Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in Cell and Developmental Biology*, 8:673.

Pyysalo, S. and S. Ananiadou. 2014. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868–875.

Pyysalo, S., F. Ginter, J. Heimonen, J. Bjorne, J. Boberg, J. Jarvinen, and T. Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC bioinformatics*, 8(1):1–24.

Pyysalo, S., T. Ohta, R. Rak, A. Rowley, H.-W. Chun, S.-J. Jung, S.-P. Choi, J. Tsujii, and S. Ananiadou. 2015. Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013. *BMC bioinformatics*, 16(10):1–19.

Robertson, S. E., S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at trec-3. In *TREC*.

Schriml, L. M., C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. 2012. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946.

Segura Bedmar, I., P. Martínez, and M. Herrero Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics.

Smith, L., L. K. Tanabe, R. J. nee Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, et al. 2008. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(2):1–19.

Wang, L. L., K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, et al. 2020. CORD-19: The Covid-19 Open Research Dataset. *arXiv preprint arXiv:2004.10706*.

Yadav, V. and S. Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

Zhou, G., J. Zhang, J. Su, D. Shen, and C. Tan. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190, May.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 177-188 recibido 31-03-2022 revisado 03-06-2022 aceptado 06-06-2022

Transformers for Lexical Complexity Prediction in Spanish Language

Transformers para la Predicción de la Complejidad Léxica en Lengua Española

Jenny Ortiz-Zambrano[1], César Espin-Riofrio[1], Arturo Montejo-Ráez[2]
[1]Universidad de Guayaquil
[2]Universidad de Jaén
{jenny.ortizz,cesar.espinr}@ug.edu.ec
amontejo@red.ujaen.es

Abstract: In this article we have presented a contribution to the prediction of the complexity of simple words in the Spanish language whose foundation is based on the combination of a large number of features of different types. We obtained the results after run the fined models based on Transformers and executed on the pre-trained models BERT, XLM-RoBERTa, and RoBERTa-large-BNE in the different datasets in Spanish and executed on several regression algorithms. The evaluation of the results determined that a good performance was achieved with a Mean Absolute Error (MAE) = 0.1598 and Pearson = 0.9883 achieved with the training and evaluation of the Random Forest Regressor algorithm for the refined BERT model. As a possible alternative proposal to achieve a better prediction of lexical complexity, we are very interested in continuing to carry out experimentations with data sets for Spanish, testing state-of-the-art Transformer models.
Keywords: Lexical Complexity, Prediction, Encodings, Transformers.

Resumen: En este artículo hemos presentado una contribución a la predicción de la complejidad de palabras simples en lengua española cuyo fundamento se basa en la combinación de un gran número de características de distinta naturaleza. Obtuvimos los resultados después de ejecutar los modelos afinados basados en Transformers y ejecutados sobre los modelos pre-entrenados BERT, XLM-RoBERTa y RoBERTa-large-BNE en los diferentes conjuntos de datos en español y corridos con varios algoritmos de regresión. La evaluación de los resultados determinó que se logró un buen desempeño con un Error Absoluto Medio (MAE) = 0.1598 y Pearson = 0.9883 logrado con el entrenamiento y evaluación del algoritmo Random Forest Regressor para el modelo BERT afinado. Como posible propuesta alternativa para lograr una mejor predicción de la complejidad léxica, estamos muy interesados en seguir realizando experimentaciones con conjuntos de datos para español probando modelos de Transformer de última generación.
Palabras clave: Complejidad Léxica, Predicción, Incrustaciones de Palabra, Transformadores.

1 Introduction

A common assumption is that people who are familiar with the vocabulary of a text can often understand the meaning of the words, even if they have difficulty with grammatical structures (Uluslu, 2022). The task of detecting in the content of the documents the words that are difficult or complex to understand by the people of a given group is known as Complex Word Identification - CWI (Rico-Sulayes, 2020) and it is a task that constitutes a fundamental step in many applications related to natural language, such as Text Simplification. Automatic lexical simplification can then become an effective method of making the text accessible to different audiences (Uluslu, 2022).

Deep learning and its revolutionary tech-

ISSN 1135-5948. DOI 10.26342/2022-69-15

nology constitute the new state of the art in various Natural Language Processing (NLP) tasks (Singh and Mahmood, 2021), in which lexical complexity prediction (LCP) is no exception (Nandy et al., 2021). It is important to point out that, after the comparison and analysis of other approaches versus deep learning approaches, a viable path of possible solutions is generated for low-resource languages where deep models are not always available or work as well as those of deep learning in English language. Likewise, it should be taken into account that the computational requirements for the application of deep learning models turn out to be significantly higher compared to those used in traditional approaches (Bender et al., 2021).

The field of NLP has shown incredible progress in the last two years, this is particularly due to the Transformer architecture (Vaswani et al., 2017) that takes advantage of large amounts of unlabeled text corpus (Canete et al., 2020). Deep learning models show significant improvement over shallow machine learning models with the rise of transfer learning and pretrained language models. The deep learning pretrained language models, BERT and XLM-RoBERTa, are considered state-of-the-art in many NLP tasks (Yaseen et al., 2021).

We present our approach aimed at predicting the complexity score for single words in the Spanish language, since resources are scarce and are not as numerous as those available for the English language.

Our model leverages the combination of advanced NLP techniques of deep learning models based on Transformers: BERT (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2019), RoBERTa-large-BNE (Gutiérrez-Fandiño et al., 2021) and pre-trained word embeddings together with a set of textual complexity features made by hand (Hand-Crafted Features). For this, we use the corpus in Spanish CLexIS^2 corpus proposed by (Zambrano and Montejo-Ráez, 2021). Our challenge is achieving to improve the lexical complexity prediction implementing a fine-tuned model on a previously trained model, for which, we follow the research done by (Rojas and Alva-Manchego, 2021).

The models used achieve a good performance shown in the results with a MAE = 0.1592 and a Person correlation 0.9883 for the identification of simple complex words.

2 Related Work

In past decades, the application of very simple metrics such as calculating the number of syllables in words (Mc Laughlin, 1969) or verifying whether the word was part of a specific list to classify it as easy or complex (Dale and Chall, 1948) were the techniques that were applied in text legibility tasks.

After, the systems based on the characterization of words (using contextual, lexical and semantic characteristics) and the application of a Random Forest classifier (Breiman, 2001) to determine whether a word is complex or not are presented. A total of 45 handwritten features were computed in these systems, and each word was modeled as a feature vector. Surface functions (three functions), dependency tree functions (eight functions), Corpus-based functions (fifteen functions), WordNet functions (eleven functions), and WordNet and corpus frequency functions (four functions) were applied. The best result obtained was a Precision value of 0.186, a Recall of 0.673, a G-score of 0.750 99 and an F-score of 0.292.

The investigations in the last years are directed to the Identification of Complex Words - CWI. The objective of these applications is to be able to predict the complexity of words based on the construction of their features, as exposed in the work carried out by (Shardlow, Cooper, and Zampieri, 2020) presenting their approach on a set of features of word embeddings from Glove, InferSent, and various linguistic features obtained as predictive sources of lexical complexity, such as word frequency, word length, or number of syllables. Then, they trained a linear regression model using different subsets of functions, obtaining as a result an MAE = of 0.0853.

(Shardlow et al., 2021) developed a system for predicting word complexity for the shared LCP task hosted on SemEval 2021 where task organizers distributed to participants the CompLex corpus (Shardlow, Cooper, and Zampieri, 2020) but in its augmented version. The task was located on the Lexical Semantics track, which consisted of predicting the complexity value of words in context.

(Ortiz-Zambrano and Montejo-Ráez, 2021) Carried out a machine learning approach that was based on 15 linguistic features obtained at the word level and their environment. Trained a supervised

random forest regression algorithm on the set of features. Several runs were made with different values to observe the performance of the algorithm. The best results achieved were a MAE = 0.07347, MSE = 0.00938 and RMSE = 0.096871.

In our approach, we review the use of word embeddings from the pre-trained and fine-tuned models, and compare them to a broader list of linguistic features at the lexical level. Our objective is to provide an exhaustive evaluation that shows more clearly, the executions carried out on several different data sets in the Spanish language, how the lexical features together with the deep encodings contribute to the prediction of lexical complexity.

3 Materials and Method

This section briefly details about the pre-trained models and their application for the generation of encodings at both the sentence and word levels. Likewise, the data sets that have been used in the different experiments are presented. Finally, the different classification algorithms and the applied features are shown.

3.1 Dataset

The CLexIS^2 corpus was elaborated with the transcripts of the recorded classes of the professors of the degrees of Computer Systems Engineering and Software Engineering, two degrees that belong to the Faculty of Mathematical Sciences of the University of Guayaquil (Ecuador) (Zambrano and Montejo-Ráez, 2021).

CLexIS^2 has become a resource of great interest and importance, due to the fact that there are few resources in Spanish available for NLP researchers[1], and some of them do not usually contain annotations that facilitate the development of NLP models (Davidson et al., 2020). For its construction, the collection presented in the ALexS 2020 competition at IberLEF 2020 (Ortiz-Zambranoa and Montejo-Ráezb, 2020) was taken as a reference.

Annotated words as complex have an associated level of complexity, computed based on the number of annotators that agreed to consider it as a complex word. Therefore, the task we are facing here can be faced as a regression problem, so error metrics will be used to evaluate different systems.

Table 1 shows some statistics on different type of words present in the CLexIS^2 dataset.

3.2 Transformer based language models

The models were taken from the *Transformers*[2] library.

- The pre-trained BERT model that we chose was the one that the Spanish community uses mostly in research work to date, which is bert-base-uncased (BETO) (Canete et al., 2020).

  BERT-base model has the number of layers L=12, the hidden size H=768, the number of self-attention heads A=12,and Total Parameters=110M.

  BERT-large model has the number of layers=24, the hidden size=1024, the number of self-attention heads=16, and Total Parameters=335M (Conneau et al., 2019).

- The -RoBERTa model applied was xlm-roberta-base (Conneau et al., 2019).

  The XLM-RoBERTa-base model has the number of layers L=12, the hidden size H=768, the number of self-attention heads A=12,and Total Parameters=270M.

  XLM-RoBERTa-large model has the number of layers L=24, the hidden size H=1024, the number of self-attention heads A=16, and Total Parameters=550M (Conneau et al., 2019).

- The RoBERTa-large-BNE model used was PlanTL-GOB-ES/roberta-large-bne being the largest Spanish-specific model to date (Gutiérrez-Fandiño et al., 2021).

  XLM-RoBERTa-large is a transformer-based masked language model for the Spanish language. It is based on the RoBERTa large model[3].

  The RoBERTa-large-BNE model has the number of layers L=24, the hidden size H=1024, the number of self-attention heads A=16,and Total Parameters=335M(Conneau et al., 2019).

[1] CLexIS^2 - https://osf.io/kfpc9/?view$_o nly$ = 18ae61a2049a48cb91c6773d53fb8ac2

[2] https://huggingface.co/

[3] https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne

Number of	Count
Sum. of content words (verbs, adjectives and nouns)	153,885
Different content words	200,785
Rare words (low frequency in CREA corpus (Saggion et al., 2015))	143,464
Sentences	9,756
Complex sentences	4,101
Total words	300,420

Table 1: Volumetrics for CLexIS[2].

3.3 Experiments design

Our purpose is to demonstrate how the combination of different types of features contribute to a better performance in predicting lexical complexity. We base our proposal on several of the works presented at the International Workshop on Semantic Evaluation - SemEval-2021 (Shardlow et al., 2021) where a total of 198 teams were presented, of which 54 teams officially sent their executions[4]; but the work that most attracted us due to its methodology was the experimentation carried out by (Zaharia, Cercel, and Dascalu, 2021) about *Combining Deep Learning and Hand-Crafted Features for Lexical Complexity Prediction*.

The figure 1 presents the workflow of the process executed to obtain of the Lexical Complexity Prediction. First, we chose the data sets for training were: the first data set was made up of the linguistic features made by hand - Hand-Crafted Features (HCF) and the second data set was made up of the Transformers encodings from the models: BERT in Spanish, multilingual XLM-RoBERTa and RoBERTa-grande-BNE. Next, we applied a fitted model on top of the previously trained model to demonstrate how running the fitted model on the previously trained model contributed to more accurate LCP results as see figure 1.

Finally, the different supervised learning algorithms were executed on the training data set to evaluate which of them achieved the best prediction score. Triple cross-validation was performed to ensure that the partitions contained independent data for training and testing. We have used some metrics that were applied to the results of the experiments presented in Sem-Eval 2021 (Shardlow et al., 2021), which are appropriate for evaluating continuous and classified data, such as: MAE, MSE, RMSE and Pearson's correlation.

[4]https:// semeval.github.io/SemEval2021/tasks

3.3.1 Features

- **Hand-Crafted Features - HCF**

To obtain the morphological aspects of the text, we perform several experiments applying a total of 23 linguistic features and combine them with the word and sentence embeddings of previously trained deep learning models.

We have considered the 15 HCF proposed by (Ortiz-Zambrano and Montejo-Ráez, 2021) and added a sets of features computed from POS categories counts (Vettigli and Sorgente, 2021), (Liebeskind, Elkayam, and Liebeskind, 2021), giving a total of 23 Hand-Crafted Features. We used the Spacy library together with the model *es_core_news_sm* to extract these features. All these features were normalized with a z-score transformation before passing them to the learning algorithm.

1. *Absolute frequency*: the absolute frequency.
   The frequency of words is a measure that serves as an indicator of lexical complexity. If in common parlance a word occurs frequently, it is more likely to be recognized (Rayner and Duffy, 1986) and (Shardlow, Cooper, and Zampieri, 2020).

2. *Relative frequency*: the relative frequency of the target word.

3. *Word length*: the number of characters of the token. The length of the word was calculated in number of its characters. It is often the case that longer length words are more difficult to process and can therefore be considered *complex*. (Shardlow, 2013) (Shardlow, Cooper, and Zampieri, 2020) (Paetzold and Specia, 2016).

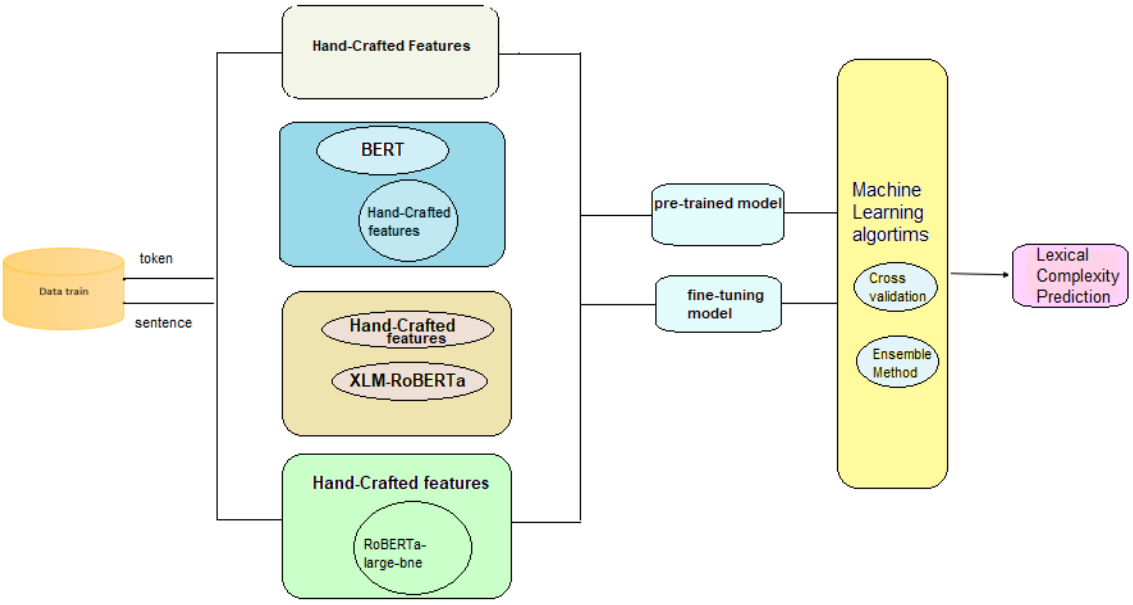


Figure 1: Representation of the workflow to obtain the Lexical Complexity Prediction.

4. *Number of syllables*: the number of syllables. A good estimate of complexity is the number of syllables contained in a word (Shardlow, 2013) (Ronzano et al., 2016) (Shardlow, Cooper, and Zampieri, 2020) (Paetzold and Specia, 2016).
5. *Target word position* (token-position): the position of the target word in the sentence. Position of the word (Word-Position) (Shardlow, 2013) (Ronzano et al., 2016).
6. *Number of words in the sentence* : number of words in the sentence. Words in sentence (NumSentenceWords) (Shardlow, 2013) (Ronzano et al., 2016).

   Based on the work proposed by (Ronzano et al., 2016) in the exploring linguistic features for lexical complexity prediction.
7. *Part Of Speech (POS)*: the Part Of Speech category.
8. *Relative frequency of the previous token*: the relative frequency of the word before the token.
9. *Relative frequency of the word after the token*: the relative frequency of the word after the token.
10. *Length of previous word*: the number of characters in the word before the token.
11. *Length of the after word*: the number of characters in the word after the token.
12. *Lexical diversity - MTDL*: the lexical diversity of the target word in the sentence.

    Additionally, the following WordNet features were also considered for each target word, as in the works carried out by (Gooding and Kochmar, 2018):
13. *Number of synonyms.*
14. *Number of hyponyms.*
15. *Number of hyperonyms.*

We follow the recommendations of (Paetzold and Specia, 2016), (Ronzano et al., 2016), (Gooding and Kochmar, 2018), (Liebeskind, Elkayam, and Liebeskind, 2021), (Desai et al., 2021) with the aim of improving results, generating 8 new features originating from the POS, which were:

1. PROPN - Number of pronouns within the sentence.
2. AUX - Number of auxiliaries within the sentence.
3. VERB - Number of verbs within the sentence.
4. ADP - Number of adverbs within the sentence.
5. NOUN - Number of nouns within the sentence.
6. NN - Number of Nouns, singular or massive.

7. SYM - Number of symbols within the sentence.
8. NUM - Number of numbers within the sentence.

- **BERT vector:** The bert-base-uncased model from the Hugging Face transformer library (Wolf et al., 2020) was applied. We took all the 768-dimensional numerical representation produced by the pre-trained and fine-tuned BERT model (Devlin et al., 2018) and added the twenty-three Hand-Crafted Features obtaining a dataset with a total of 1559 linguistic features of different nature.
- **XLM-RoBERTa vector:** As in the case of the BERT model, we take all the 768-dimensional numerical representation produced by the pre-trained RoBERTa model (Conneau et al., 2019) in the different combinations of sentence and target word encodings, for both the pre-trained model and the model fine-tuned, reaching a total of 1559 linguistic characteristics of different nature.
- **RoBERTa-large-BNE vector:** Regarding this model, we take all the 1024-dimensional numerical representation produced by the pre-trained RoBERTa-large-model model (Gutiérrez-Fandiño et al., 2021), in the same way that they were applied in the previous models, the data sets were made up of for the different combinations of sentence and target word encodings, for both the pre-trained model and the fine-tuned model, reaching a total of 2071 linguistic characteristics of different nature.

3.3.2 Machine Learning Algorithms

Similar to the work done by (Zaharia, Cercel, and Dascalu, 2021) in the case of the algorithms, the training and evaluation of the different c ombinations o f t he s ets w as carried out with a total of eight supervised algorithms for the regression, these are:

1. AdaBoost - AB (Paetzold, 2021).
2. Desicion Tree - DT (Shardlow, Evans, and Zampieri, 2021).
3. Gradient Boosting - GB (Vettigli and Sorgente, 2021).
4. Stochastic Gradient - SG (Bottou, 2010).
5. Nearest Neighbors - KNN (Liebeskind, Elkayam, and Liebeskind, 2021).
6. Support Vector Machines - SVM (Liebeskind, Elkayam, and Liebeskind, 2021).
7. Passive Aggressive - PA (Crammer et al., 2006).
8. Random Forest - RF) (Zaharia, Cercel, and Dascalu, 2021), (Desai et al., 2021).

Several experiments were carried out for each of the datasets where different configurations were explored for each of the algorithms. We apply the default values for the algorithms except for the case for tree-based algorithms, achieving to determine the best hyper-parameters with the following number of nodes:

- AdaBoost with 100 nodes.
- Random forest with 241 nodes.
- Gradient Boosting algorithm with 350 nodes.

4 Results

4.1 Features Sets

We build several datasets composed of the combination of the features described above to run them on the pre-trained models. The table 2 table presents the description of the abbreviations that will be used for a better understanding of the features applied to the data sets. The detail below:

- The *Hand-Crafted Features* with the features coming from the 768-dimensional vector of the initial [CLS] token as sentence embeddings ($BERT_{sent}$).
- *The Hand-Crafted Features* with the 768-dimensional vector corresponding to the target token as word embeddings ($BERT_{word}$).
- The *Hand-Crafted Features* with encodings of the [CLS] token and encodings of the target token.
- The encodings of the [CLS] token.
- The encodings of the target token.
- The encodings of the [CLS] token with the encodings of the target token.

Features identifier	Description
HCF	Hand-Crafted (linguistic) Features.
BERT_{sent}	Sentence encodings from BERT models.
BERT_{word}	Token encodings from BERT models.
XLMR_{sent}	Sentence encodings from XLM-RoBERTa model.
XLMR_{word}	Token encodings from RoBERTa model.
RBNE_{sent}	Sentence encodings from RoBERTa-large-BNE model.
RBNE_{word}	Token encodings from RoBERTa-large-BNE model.

Table 2: Description of the Feature sets.

For the evaluation of the trained and fine-tuned models, those that were widely applied to LCP for the shared LCP task hosted in SemEval 2021: Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and Pearson correlation (Shardlow et al., 2021).

4.2 BERT model *pre-trained*

The table 3 shows the eight best performances corresponding to different combinations of features described in section 4.1 executed with BERT pre-trained.

As we can see in the three best results in predicting lexical complexity were achieved by the ABR and SVR algorithms. The best performance was achieved by the ABR - AdaBoost algorithm presenting the best prediction for the Spanish language with a MAE = 0.1632 and a Pearson = 0.999 in the execution with the data set made up of the combination of the features generated at the sentence level and at the word level - $\text{BERT}_{sent}\oplus$ BERT_{word}.

4.3 BERT model *fine-tuned*

We have applied the fine-tuned BERT model on top of the pre-trained BERT model for the purpose of the results. The table 4 shows the eight best executions, positioning RFR - Random Forest Regressor algorithm and the GBR - Gradient Boosting Regressor algorithm in the first places.

The best performance was obtained with the dataset composed of the combination of the features with target word encodings together with sentence encodings from BERT fine-tuned. The same combination of features achieved the best performance in the pre-trained model, but with lower results.

It should be noted that the RFR algorithm does not appear within the top eight places in the execution of the pre-trained model, but it achieves its best result when the model is refined, placing first and third within the three best executions tuned. RFR presented the best prediction for the Spanish language with a MAE = 0.1592 and a Pearson = 0.988 combining $\text{BERT}_{sent}\oplus$ BERT_{word}.

4.4 XLM-RoBERTa model *pre-trained*

Similar to the BERT model, the top eight sites were taken from all the runs that were done on the different data sets. The results of the best place for the pre-trained XLM-RoBERTa model were achieved by the ABR - AdaBoots algorithm with a MAE = 0.1623 and a Pearson = 0.9973 result of the combination of the features with target word and sentence encodings together with the HCF - $\text{XLMR}_{sent}\oplus$ $\text{XLMR}_{word}\oplus$ HCF, as can be seen in the table 5. It can be clearly shown that the pre-trained XLM-RoBERTa model has a better performance compared to the pre-trained BERT model, achieving a better prediction of Lexical Complexity.

4.5 XLM-RoBERTa model *fine-tuned*

We also highlight that in the execution of the XLM-RoBERTa tuned model, it achieved a significant improvement compared to the results of the pre-trained model, reaching a MAE = 0.1601 and a Pearson = 0.998 as result of the combination of the features with target word encodings together HCF - $\text{XLMR}_{word}\oplus$ HCF. See table 6.

Comparing the results of the BERT and XLM-RoBERTa both tuned models, BERT tuned is so far the one that has an important performance achieved by a MAE = 0.1592 with the execution of the RFR algorithm combining $\text{BERT}_{sent}\oplus$ BERT_{word}.

BERT *model pre-trained* with CLexIS^2					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
BERT$_{sent}\oplus$ **BERT**$_{word}$	**ABR**	**0,1632**	**0,0502**	**0,2343**	**0,9999**
BERT$_{word}$	SVR	0,1634	0,0432	0,2074	0,9023
BERT$_{word}$	ABR	0,1643	0,0512	0,2332	0,9947
BERT$_{word}$	GBR	0,1653	0,0494	0,0447	0,6977
BERT$_{word}\oplus$ **HCF**	GBR	0,1655	0,0454	0,2088	0,7040
BERT$_{sent}\oplus$ **BERT**$_{word}\oplus$ **HCF**	GBR	0,1659	0,0418	0,2074	0,7147
BERT$_{sent}\oplus$ **BERT**$_{word}$	GBR	0,1694	0,0444	0,2039	0,7167
BERT$_{sent}\oplus$ **BERT**$_{word}\oplus$ **HCF**	ABR	0,1699	0,0554	0,2334	0,9939

Table 3: Results of the model BERT pre-trained with features of different nature.

BERT *model fine-tuned* with CLexIS^2					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
BERT$_{sent}\oplus$ **BERT**$_{word}$	**RFR**	**0,1592**	**0,0379**	**0,1982**	**0,9883**
BERT$_{sent}\oplus$ **BERT**$_{word}\oplus$ **HCF**	GBR	0,1600	0,0367	0,1979	0,8202
BERT$_{sent}\oplus$ **BERT**$_{word}\oplus$ **HCF**	RFR	0,1610	0,0401	0,1988	0,9987
BERT$_{sent}\oplus$ **BERT**$_{word}\oplus$ **HCF**	ABR	0,1610	0,0506	0,2242	0,9998
BERT$_{sent}\oplus$ **BERT**$_{word}\oplus$ **HCF**	ABR	0,1621	0,0487	0,2300	0,9999
BERT$_{word}\oplus$ **HCF**	GBR	0,1622	0,0430	0,1984	0,8983
BERT$_{word}$	SVR	0,1622	0,0429	0,2018	0,9183
BERT$_{word}$	GBR	0,1632	0,0472	0,0429	0,7083

Table 4: Results of the model BERT tuned with features of different nature.

4.6 RoBERTa-large-BNE model *pre-trained*

The novelty of this research is to have incorporated the executions with the pre-trained model RoBERTa-large-BNE and its adjusted model. The eight best results are displayed in the table 7. The best position were achieved by the ABR-AdaBoost algorithm with a MAE = 0.1609 and a Person = 0.6754 combining the sentence and word encodings together with the HCF - RBNE$_{sent}\oplus$ RBNE$_{word}\oplus$ HCF.

It should be noted that the pre-trained model RoBERTa-large-BNE is the one that achieves a better prediction for lexical complexity in the Spanish language compared to the pre-trained models BERT and XLM-RoBERTa. See table 9.

4.7 RoBERTa-large-BNE model *fine-tuned*

Executing the RoBERTa-large-BNE tuned model, the results are encouraging, there is an improvement compared to the results of the pre-trained model. The table 8 displays the first places reached by the GBR-Gradient Boosting Regressor and SVR-Super Vector Regressor algorithms. It presents a low improvement, achieving in its performance a MAE = 0.1609 and a Pearson = 0.6754 combining the sentence and word encodings together with the HCF - RLBNE$_{sent}\oplus$ RLBNE$_{word}\oplus$ HCF, and the second and third places prove it in comparison with the pre-trained model.

It should be noted that the tuned model BERT is the one that achieves a better prediction for lexical complexity in the Spanish language compared to the tuned models XLM-RoBERTa and RBNE. See table 10.

It can be seen that the fined models based on Transformers make an important contribution to the Prediction of Lexical Complexity in the Spanish language. The table 11 presents the best five best results of all the experiments carried out with the models, both pre-trained and fined. It is important to mention that the Hand-Crafted Features, being such simple features because they are only based on the frequency of the words and several manual calculations, have been shown to contribute to improving the level of prediction of the complexity of the words.

5 Discussion

We have applied the BERT, RoBERTa, and RoBERTa-large-BNE models for our research in predicting lexical complexity in Spanish. We have closely followed the methodology applied in several of the works presented in the LCP task of the SemEval 2021 International Conference (Shardlow et al., 2021) which has allowed us to achieve very important results that demonstrate a relevant contribution in

XLM-RoBERTA *model pre-trained* with CLexIS^2					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HCF**	ABR	0,1623	0,0527	0,2270	0,9973
XLMR$_{word}\oplus$ **HCF**	ABR	0,1623	0,0513	0,2273	0,9973
XLMR$_{sent}\oplus$ **HCF**	ABR	0,1630	0,0524	0,2293	0,9973
XLMR$_{word}\oplus$ **HCF**	GBR	0,1653	0,0433	0,2073	0,4848
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HCF**	GBR	0,1658	0,0434	0,2074	0,4874
XLMR$_{sent}\oplus$ **HCF**	GBR	0,1663	0,0433	0,2082	0,4807
XLMR$_{word}\oplus$ **HCF**	SVR	0,1680	0,0483	0,2194	0,3095
XLMR$_{word}\oplus$ **HCF**	RFR	0,1690	0,0445	0,2093	0,9803

Table 5: Results of the model XLMR pre-trained with features of different nature.

XLM-RoBERTA *model fine-tuned* with CLexIS^2					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
XLMR$_{word}\oplus$ **HCF**	ABR	0,1601	0,0501	0,2251	0,9987
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HCF**	ABR	0,1620	0,0526	0,2268	0,9987
XLMR$_{sent}\oplus$ **HCF**	ABR	0,1620	0,0519	0,2287	0,9979
XLMR$_{sent}\oplus$ **HCF**	GBR	0,1630	0,0420	0,2062	0,4790
XLMR$_{word}\oplus$ **HCF**	GBR	0,1638	0,0429	0,2034	0,4800
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HCF**	GBR	0,1652	0,0430	0,2069	0,4930
XLMR$_{word}\oplus$ **HCF**	SVR	0,1660	0,0482	0,2172	0,3083
XLMR$_{word}\oplus$ **HCF**	RFR	0,1669	0,0427	0,2013	0,9849

Table 6: Results of the model XLMR tuned with features of different nature.

RoBERTa-large-BNE *model pre-trained* with CLexIS^2					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	ABR	0,1609	0,0421	0,2047	0,6754
RBNE$_{sent}\oplus$ **RBNE**$_{word}$	ABR	0,1675	0,0556	0,2347	0,9952
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	GBR	0,1691	0,0434	0,2073	0,6607
RBNE$_{word}$	ABR	0,1693	0,0563	0,2360	0,9948
RBNE$_{word}$	GBR	0,1696	0,0447	0,2101	0,6400
RBNE$_{sent}\oplus$ **RBNE**$_{word}$	GBR	0,1698	0,0447	0,2102	0,6450
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	SVR	0,1708	0,0507	0,2224	0,2363
RBNE$_{sent}\oplus$ **HCF**	SVR	0,1708	0,0507	0,2224	0,0857

Table 7: Results of the model RBNE pre-trained with features of different nature.

RoBERTa-large-BNE *model fine-tuned* with CLexIS^2					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	GBR	0,1609	0.0421	0.2047	0.6754
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	SVR	0,1630	0,0435	0,2070	0,4883
RBNE$_{word}\oplus$ **HCF**	SVR	0,1666	0,0466	0,2136	0,4220
RBNE$_{word}$	ABR	0,1677	0.0551	0,2336	0,9952
RBNE$_{sent}\oplus$ **RBNE**$_{word}$	SVR	0,1684	0.0472	0.2152	0.4425
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	GBR	0,1686	0,0432	0,2067	0,6854
RBNE$_{word}$	SVR	0,1686	0,0468	0,2146	0,5021
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	ABR	0,1689	0,0558	0,2351	0,9951

Table 8: Results of the model RBNE tuned with features of different nature.

The *Spanish Language Models* pre-trained **Best Result**			
Model	**Features**	**Alg**	**MAE**
RBNE	RBNE$_{sent}\oplus$ RLBNE$_{word}\oplus$ HCF	ABR	0.1609
XLMR	XLMR$_{sent}\oplus$ XLMR$_{word}\oplus$ HCF	ABR	0.1623
BERT	BERT$_{sent}\oplus$ BERT$_{word}$	ABR	0.1632

Table 9: Best results models pre-trained.

The *Spanish Language Models* fine-tuned Best Result			
Model	**Features**	**Alg**	**MAE**
BERT	$\text{BERT}_{sent}\oplus$ BERT_{word}	RFR	0,1592
XLMR	$\text{XLMR}_{word}\oplus$ HCF	ABR	0,1601
RBNE	$\text{RBNE}_{sent}\oplus$ $\text{RBNE}_{word}\oplus$ HCF	GBR	0,1609

Table 10: Best results models fine-tuned.

Summary of best results on the CLexIS^2 corpus						
Model	**Features**	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
$\textbf{BERT}_{fine-tuned}$	$\textbf{BERT}_{sent}\oplus$ $\textbf{BERT}_{word}$	RFR	0,1592	0,0379	0,1982	0,9883
$\textbf{BERT}_{fine-tuned}$	$\textbf{BERT}_{sent}\oplus$ $\textbf{BERT}_{word}\oplus$ **HCF**	GBR	0,1600	0,0367	0,1979	0,8202
$\textbf{XLMR}_{fine-tuned}$	$\textbf{XLMR}_{word}\oplus$ **HCF**	ABR	0,1601	0,0501	0,2251	0,9987
$\textbf{RBNE}_{fine-tuned}$	$\textbf{RBNE}_{sent}\oplus$ $\textbf{RBNE}_{word}$	GBR	0,1609	0,0421	0,2047	0,6754
$\textbf{BERT}_{fine-tuned}$	$\textbf{BERT}_{sent}\oplus$ $\textbf{BERT}_{word}\oplus$ **HCF**	RFR	0,1610	0,0401	0,1988	0,9987

Table 11: Summary of best results on the CLexIS^2 dataset.

the area of Lexical Simplification for Spanish.

We observe that according to the results of the final evaluation, especially in terms of fine-tuning, the Spanish language fined models made an important contribution to the prediction of lexical complexity by outperforming the proposal presented after the execution of the manual features-HCF. In the case of the RoBERTa-large-BNE model, we have found a performance that exceeds the rest of the models after the execution of the pre-trained model and even remains within the three best executions in the results of the tuned models, such as the proposals presented by (Gutiérrez-Fandiño et al., 2021)

6 Conclusions and Further Work

In this article, we have presented a contribution to predict the complexity of simple words in the Spanish language, combining a large number of features of different types. We consider that, after the multiple experimentations that we carried out, it allowed us to know the maximum performance for the different combinations of the data sets by applying the regression algorithms.

In our experiments, we obtained the results after the execution of several previously trained transformer-based models on several datasets in Spanish, combining features of different nature. The application of the fine-tuned models to generate features (embeddings) achieved a better performance of explored machine learning algorithms, which led to a MAE = 0.1598 and a Pearson of 0.9883 achieved with the evaluation and training of the Random Forest Regressor algorithm for the tuned model BERT.

Additional features can boost pre-trained models to levels of performance close to those of fine-tuned models alone, so it could be a feasible approach when there are not enough computational resources for such a downstream training.

As a possible alternative proposal to achieve a better prediction of lexical complexity, we are very interested in continuing to carry out experimentations on data sets for Spanish, testing state-of-the-art Transformer models. To this end, extrinsic evaluation will be overcome, comparing the best systems on this specific task with the possibilities of integrating external features like the ones proposed in this work.

Acknowledgements

We appreciate Jostin Daniel Escobar Suarez, Joel Stalin Sorroza Contreras, Diego Gabriel Bernal Yucailla y Diana Geovanna Aroca Pincay graduate of the Computer Systems Engineering degree from the University of Guayaquil, for their valuable contribution to the development of our work.

This work is partially funded by grants P20_00956 (PAIDI 2020) and 1380939 (FEDER Andalucía 2014-2020) from the Andalusian Regional Government.

References

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the

dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, pages 177–186.

Breiman, L. 2001. Random forests. *Machine learning*, 45(1):5–32.

Canete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Crammer, K., O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Dale, E. and J. S. Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Davidson, S., A. Yamada, P. F. Mira, A. Carando, C. H. S. Gutierrez, and K. Sagae. 2020. Developing nlp tools with a new corpus of learner spanish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7238–7243.

Desai, A., K. North, M. Zampieri, and C. M. Homan. 2021. Lcp-rit at semeval-2021 task 1: Exploring linguistic features for lexical complexity prediction. *arXiv preprint arXiv:2105.08780*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Gooding, S. and E. Kochmar. 2018. Camb at cwi shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, and M. Villegas. 2021. Spanish language models. *arXiv preprint arXiv:2107.07253*.

Liebeskind, C., O. Elkayam, and S. Liebeskind. 2021. Jct at semeval-2021 task 1: Context-aware representation for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 138–143.

Liu, X., P. He, W. Chen, and J. Gao. 2019. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*.

Mc Laughlin, G. H. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Nandy, A., S. Adak, T. Halder, and S. M. Pokala. 2021. cs60075_team2 at semeval-2021 task 1: Lexical complexity prediction using transformer-based language models pre-trained on various text corpora. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 678–682.

Ortiz-Zambrano, J. A. and A. Montejo-Ráez. 2021. Complex words identification using word-level features for semeval-2020 task 1. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 126–129.

Ortiz-Zambranoa, J. A. and A. Montejo-Ráezb. 2020. Overview of alexs 2020: First workshop on lexical analysis at sepln. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*.

Paetzold, G. 2021. Utfpr at semeval-2021 task 1: Complexity prediction by combining bert vectors and classic features. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 617–622.

Paetzold, G. and L. Specia. 2016. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th Interna-*

tional Workshop on Semantic Evaluation (SemEval-2016), pages 969–974.

Rayner, K. and S. A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Rico-Sulayes, A. 2020. General lexicon-based complex word identification extended with stem n-grams and morphological engines. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain*.

Rojas, K. R. and F. Alva-Manchego. 2021. Iapucp at semeval-2021 task 1: Stacking fine-tuned transformers is almost all you need for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 144–149.

Ronzano, F., L. E. Anke, H. Saggion, et al. 2016. Taln at semeval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1011–1016.

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Shardlow, M. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109.

Shardlow, M., M. Cooper, and M. Zampieri. 2020. Complex: A new corpus for lexical complexity prediction from likert scale data. *arXiv preprint arXiv:2003.07008*.

Shardlow, M., R. Evans, G. H. Paetzold, and M. Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. *arXiv preprint arXiv:2106.00473*.

Shardlow, M., R. Evans, and M. Zampieri. 2021. Predicting lexical complexity in english texts. *arXiv preprint arXiv:2102.08773*.

Singh, S. and A. Mahmood. 2021. The nlp cookbook: Modern recipes for transformer based deep learning architectures. *IEEE Access*, 9:68675–68702.

Uluslu, A. Y. 2022. Automatic lexical simplification for turkish. *arXiv preprint arXiv:2201.05878*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vettigli, G. and A. Sorgente. 2021. Compna at semeval-2021 task 1: Prediction of lexical complexity analyzing heterogeneous features. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 560–564.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Yaseen, T. B., Q. Ismail, S. Al-Omari, E. Al-Sobh, and M. Abdullah. 2021. Just-blue at semeval-2021 task 1: Predicting lexical complexity using bert and roberta pretrained language models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 661–666.

Zaharia, G.-E., D.-C. Cercel, and M. Dascalu. 2021. Upb at semeval-2021 task 1: Combining deep learning and handcrafted features for lexical complexity prediction. *arXiv preprint arXiv:2104.06983*.

Zambrano, J. A. O. and A. Montejo-Ráez. 2021. Clexis2: A new corpus for complex word identification research in computing studies. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1075–1083.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 189-196 recibido 29-03-2022 revisado 20-05-2022 aceptado 27-05-2022

Building a comparable corpus and a benchmark for Spanish medical text simplification

Construcción de un corpus comparable y un recurso de referencia para la simplificación de textos médicos en español

Leonardo Campillos-Llanos,[1] **Ana R. Terroba Reinares,**[2]
Sofía Zakhir Puig,[1] **Ana Valverde-Mateos**[3], **Adrián Capllonch-Carrión**[4]
[1]ILLA - Consejo Superior de Investigaciones Científicas (CSIC)
[2]Fundación Rioja Salud
[3]Unidad de Terminología Médica, Real Academia Nacional de Medicina de España
[4]Centro de Salud Retiro, Hospital General Universitario Gregorio Marañón
{leonardo.campillos,sofia.zakhir}@csic.es, arterroba@riojasalud.es,
avalverde@ranm.es, adrian.capllonch@salud.madrid.org

Abstract: We report the collection of the CLARA-MeD comparable corpus, which is made up of 24 298 pairs of professional and simplified texts in the medical domain for the Spanish language (>96M tokens). Texts types range from drug leaflets and summaries of product characteristics (10 211 pairs of texts, >82M words), abstracts of systematic reviews (8138 pairs of texts, >9M words), cancer-related information summaries (201 pairs of texts, >3M tokens) and clinical trials announcements (5748 pairs of texts, 451 690 words). We also report the alignment of professional and simplified sentences, conducted manually by pairs of annotators. A subset of 3800 sentence pairs (149 862 tokens) has been aligned each by 2 experts, with an average inter-annotator agreement *kappa* score of 0.839 (±0.076). The data are available in the community and contributes with a new benchmark to develop and evaluate automatic medical text simplification systems.
Keywords: Comparable corpus. Medical text simplification. Biomedical natural language processing.

Resumen: Se describe la recogida del corpus comparable CLARA-MeD, formado por 24 298 pares de textos profesionales y simplificados de dominio médico en lengua española (>96M palabras). Los tipos de textos varían desde prospectos médicos y fichas técnicas de medicamentos (10 211 pares de textos, >82M palabras), resúmenes de revisiones sistemáticas (8138 pares de textos, >9M palabras), resúmenes de información sobre el cáncer (201 pares de textos, >3M palabras) y anuncios de ensayos clínicos (5748 pares de textos, 451 690 palabras). También presentamos el alineamiento de frases técnicas y simplificadas, realizado a mano por pares de anotadores. Un subconjunto de 3800 pares de frases (149 862 tokens) se han emparejado, con un acuerdo medio entre anotadores con valor $kappa = 0.839$ (±0.076). Los datos están disponibles en la comunidad y este nuevo recurso permite desarrollar y evaluar sistemas de simplificación automática de textos médicos.
Palabras clave: Corpus comparable. Simplificación de textos médicos. Procesamiento de lenguaje natural biomédico.

1 Introduction

Text simplification is the task of *transforming a text into an equivalent which is more understandable* (Saggion et al., 2011). The application of natural language processing (NLP) techniques makes it possible to automate the simplication of texts across domains and tasks, ranging from legal and administrative texts (Scarton et al., 2018), language learning (Petersen and Ostendorf, 2007), users with special reading needs (Barbu et al., 2015) or health literacy (Kindig et al., 2004).

Corpus data are required for analysing text simplification strategies, developing and testing NLP systems. This work introduces

ISSN 1135-5948. DOI 10.26342/2022-69-16

a new resource made up of documents from the medical domain, which is available at: https://digital.csic.es/handle/10261/269887.

2 Background

Text simplification approaches are commonly conceived as a translation task—from the technical to laymen's register. Simplification involves operations at multiple linguistic levels: grammar (e.g. simpler word order, passive to active voice), discourse (e.g. split long sentences) or lexis (e.g. replace complex words with clearer synonyms). Some automatic simplification approaches rely on rule-based or lexicon-based modules to address each linguistic level. In contrast, data-driven approaches may use machine translation of monolingual (professional/simplified) corpora—at present, mostly via deep-learning-based methods (Van den Bercken et al., 2019; Sakakini et al., 2020; Devaraj et al., 2021; Martin et al., 2021).

Regardless of the approach, dedicated corpora are need: comparable resources (professional and simplified versions of a text) or, ideally, parallel corpora (texts with almost identical content in different registers).

In addition to corpora for the English language (Van den Bercken et al., 2019; Sakakini et al., 2020), simplification text collections exist for Brazilian Portuguese (Caseli et al., 2009), German (Klaper et al., 2013), Italian (Tonelli et al., 2016) or French (Grabar and Cardon, 2018; Gala et al., 2020). Other multilingual resources have been released in challenges for complex word identification (Yimam et al., 2017)

For Spanish, there is the EASIER corpus,[1] with different characteristics compared to the CLARA-MeD text collection. First, the EASIER corpus is a general domain resource; even though some sentences are related to health topics, the CLARA-MeD resource focuses only on the medical domain. Second, the EASIER corpus gathers 3977 sentences annotated with 8155 complex words, and 3396 sentences labeled with 7892 suggested synonyms. The CLARA-MeD corpus is not annotated with these data, but features a sentence-level alignment of 3800 pairs of technical and simplified sentences, following specific criteria (§3.6). Lastly, besides including parallel data, the CLARA-MeD corpus is larger in size. These data can be used to enlarge the number of aligned sentences or can be annotated in more detail in future versions.

Several methods have been applied to collect such type of corpora. Wikipedia and Simple Wikipedia have been aligned to obtain a parallel corpus in the general domain (Zhu et al., 2010). However, the correspondence of content between technical/simplified versions are often deficient (Xu et al., 2015). Moreover, this method can not be directly extended to languages without a simplified Wikipedia. In these cases, some teams (Palmero Aprosio et al., 2019; Rauf et al., 2020) have created synthetic corpora via translations from SimpleWikipedia or similar sources such as WikiLarge (Grabar and Cardon, 2018). Another method is manual simplification by domain or task specialists (Gala et al., 2020), which assures linguistic quality but requires an adequate team and is more time-consuming. Hybrid methods have also been conducted for medical English (Van den Bercken et al., 2019; Moramarco et al., 2021).

3 Methods and Sources

Figure 1 summarizes the workflow applied to create the corpus. The first stage involved collecting a comparable resource. We gathered medical texts with two versions (for professionals and laymen readers) from sources recently reported (Moreno-Sandoval et al., 2019). At the current stage, we have not used articles from the Medicine category in Wikipedia, given that there is not a Spanish version from Simple Wikipedia.

The second stage implied matching professional and simplified sentences from each comparable subcorpus. In the following, we detail the data sources to create this resource (§ 3.1-§ 3.4), and Section 3.6 explains the criteria and methods to align sentences.

3.1 Drug leaflets and summaries of product characteristics

The Medicine Online Information Center (Centro de Información de Medicamentos, CIMA) [2] is a drug-related service and knowledge database maintained by the Spanish Agency of Medicines and Medical Devices

[1] shorturl.at/lvCJU

[2] https://cima.aemps.es

Figure 1: Methods and stages to compile the corpus.

(AEMPS). CIMA provides all the information related to drugs prescribed in Spain through a search engine, the Nomenclator resource (a rich database of medical drugs), and pharmaceutical/pharmacovigilance reports. In addition, the information about each medical drug (indication, medical brand, dosage, unit of presentation, etc.) is provided in two types of documents: summaries of product characteristics (written in a professional register and aimed at healthcare professionals) and drug leaflets (with simpler structures or terms, and aimed at patients). We downloaded both types of data, and release only cleaned, noise-free texts with both versions available (10 211 pairs of texts, 82 907 317 words).[3]

3.2 Systematic reviews

The Cochrane Library[4] is an updated database of systematic reviews and metanalyses. This is the main collection of medical evidence (Sackett et al., 1996), mostly from publications reporting results of clinical trials. Healthcare professionals use this database to keep up to date with the latest evidence to apply in their clinical practice. The Cochrane Library is a multilingual resource, although not all reviews are available in all languages. Each review presents an abstract of the full text and also a summary in plain language, which is aimed at a non-specialist readership. We collected a total of 8138 pairs of documents (9 618 698 words).

3.3 Cancer-related information summaries

The National Cancer Institute (NCI) website presents a large volume of bilingual (English and Spanish) information.[5] Contents revolve around cancer types, disorders and symptoms, oncological therapies, pharmacological substances, genetics, screening, prevention, palliative care or patient-oriented counseling. Noteworthy, *Physician Data Queries* (*PDQ*) articles gather updated and evidence-based information about essential aspects of cancer in two versions: for professionals and patients. We collected 201 pairs of PDQ texts, and removed noisy information from the web pages (e.g. URLs, content menus, tables, etc). The cleaned texts amount up to 3 044 461 words.

3.4 Clinical trials announcements

The European Clinical Trials register (EudraCT)[6] gathers public data about all clinical trials conducted in the European Union, either at national or multinational level. Clinical trials announcements are published both in English and the European language corresponding to the countries involved in the trial. Trial announcements describe the trial protocols, patients and participants, interventions, indications and expected outcomes of the trial, among others. Two sections are written with equivalent contents between scientific (aimed at healthcare professionals) and popularization levels (aimed at patients or laymen users): the public and scientific title of the trial, and the public and scientific indication. To gather these data, we reused 700 texts from the Clinical Trials for Evidence Based Medicine in Spanish (CT-EBM-SP) corpus (Campillos-Llanos et al., 2021). We also downloaded more than 7500 announcements from the website, and selected only those with both a public and scientific version of the title and/or indication. After filtering out noisy or redundant data, we gathered 5784 pairs of texts (451 690 words).

3.5 Descriptive statistics

Table 1 includes excerpts of professional and simplified versions of each data source. Table 2 shows the word count of the comparable corpus, and Table 3, the average of sentences per text and average words per sentence (with standard deviation, SD). Texts

[3]The data from the Cochrane Library cannot be released without permission.

[4]https://www.cochranelibrary.com/

[5]https://www.cancer.gov

[6]https://www.clinicaltrialsregister.eu

Source	Professional version	Simplified version
CIMA	*La administración concomitante de metamizol con metotrexato u otros antineoplásicos puede aumentar la toxicidad sanguínea de los antineoplásicos particularmente en pacientes de edad avanzada.* 'Concomitant administration of metamizole with methotrexate or other antineoplastics may increase the blood toxicity of antineoplastics particularly in elderly patients.'	*Si se administra conjuntamente con metotrexato u otros medicamentos para el tratamiento de los tumores (antineoplásicos), puede potenciar los efectos tóxicos en sangre de los antineoplásicos, sobre todo en pacientes de edad avanzada.* 'If co-administered with methotrexate or other drugs for the treatment of tumors (antineoplastics), it may potentiate the toxic effects of antineoplastics in the blood, especially in elderly patients.'
Cochrane	*La administración de suplementos de vitamina D podría disminuir la necesidad de ventilación mecánica invasiva, pero la evidencia es incierta (evidencia de certeza baja).* 'Vitamin D supplementation may decrease need for invasive mechanical ventilation, but the evidence is uncertain (low-certainty evidence).'	*La vitamina D podría reducir la necesidad de conectar a los pacientes a un respirador para ayudarles a respirar, pero se desconoce la evidencia.* 'Vitamin D may reduce the need for patients to be put on a ventilator to help them breathe, but the evidence is uncertain.'
EudraCT	*Ensayo clínico aleatorizado, doble ciego, controlado con placebo, para evaluar la eficacia y seguridad de la vacuna COMIRNATY (vacuna COVID-19 ARNm, Pfizer-BioNTech) en personas con COVID persistente* 'Randomized, double-blind, placebo-controlled clinical trial to evaluate the efficacy and safety of the COMIRNATY vaccine (COVID-19 mRNA vaccine, Pfizer-BioNTech) in people with long COVID'	*El objetivo del estudio es analizar si la administración de una vacuna contra la infección COVID19 puede hacer disminuir los síntomas de COVID persistente.* 'The aim of the study is to analyze whether the administration of a vaccine against COVID19 infection can reduce the symptoms of long COVID.'
NCI	*El LH que se diagnostica durante el primer trimestre del embarazo no constituye un indicador absoluto de la necesidad de un aborto terapéutico.* 'HL that is diagnosed in the first trimester of pregnancy does not constitute an absolute indication for therapeutic abortion.'	*Cuando el linfoma de Hodgkin se diagnostica durante el primer trimestre del embarazo, no siempre significa que se aconsejará a la mujer que interrumpa el embarazo.* 'When Hodgkin lymphoma is diagnosed in the first trimester of pregnancy, it does not necessarily mean that the woman will be advised to end the pregnancy.'

Table 1: Samples of professional and simplified versions of texts from different data sources.

from CIMA (drug leaflets and summaries of product characteristics) and NCI (cancer-related summaries) are longer. Professional texts from all sources tend to be longer than simplified texts. Likewise, the average sentence length (number of words per sentence) is generally longer in the professional version of almost all sources (except for CIMA).

3.6 Parallel text alignment

We aligned 3800 professional-laymen sentences extracted from the CLARA-MeD corpus. Pairs of annotators with varied backgrounds (a computational linguist, a medical doctor and medical terminologists) matched scientific and simplified versions of a subset of sentences.

Gathering parallel sentences from the EudraCT announcements was straightforward. Each clinical trial announcement contains two versions (for patients/laymen users and healthcare professionals) of specific sections: the public and scientific title of the trial, and a public and scientific indication. We gathered the data from both versions, which yielded a preliminary noisy alignment of 5784 sentence pairs. Similar sentences were rejected, and two annotators per data batch conducted a manual revision. We followed a set of linguistic criteria to accept a sentence pair as adequate equivalences (§ 3.6.2). Problematic pairs of sentences were discussed to achieve a consensus.

For the other sources, we automated the alignment by extracting, for each professional sentence, the most similar simplified version (§ 3.6.1). After the semi-automatic alignment, we followed the same methodology for the manual revision, and two annotators per data batch checked the sentence pairs. We thereby filtered out the most reliable sentence pairs and assessed the inter-annotator agreement.

3.6.1 Semi-automatic alignment

To gather aligned pairs of sentences, we should combine, for each pair of texts, all professional sentences with all simplified ones. Nonetheless, the amount of candidate pairs collected in this way is unaffordable to be re-

Source	Text pairs	Professional	Simplified	Total
CIMA	10 211	55 463 410	27 443 907	82 907 317
Cochrane abstracts	8138	6 235 454	3 383 244	9 618 698
EudraCT	5748	255 902	195 788	451 690
NCI	201	2 093 569	955 480	3 049 049
Total	**24 298**	**64 048 335**	**31 978 419**	**96 022 166**

Table 2: Word count per source data of the comparable corpus.

Source		Professional	Simplified
CIMA	Avg sentences per text (±SD)	431.72 (±234.47)	210.03 (±71.96)
	Avg words per sentence (±SD)	12.36 (±2.48)	12.76 (±1.61)
Cochrane abstracts	Avg sentences per text (±SD)	34.06 (±9.05)	19.57 (±11.70)
	Avg words per sentence (±SD)	22.08 (±3.94)	22.02 (±4.39)
EudraCT	Avg sentences per text (±SD)	1.77 (±0.65)	1.75 (±0.62)
	Avg words per sentence (±SD)	26.47 (±11.20)	19.73 (±8.74)
NCI	Avg sentences per text (±SD)	505.53 (±492.57)	309.39 (±177.61)
	Avg words per sentence (±SD)	20.53 (±3.75)	15.47 (±1.49)

Table 3: Average (avg) sentences per text and average words per sentence (±standard deviation).

vised, and only very few pairs will be adequate alignments.

Previous work (Cardon and Grabar, 2020) has reduced this *search space* by leveraging parsing information and developing a machine-learning classifier using features such as sentence length, the Levenshtein distance or number of shared words between each version, among others. The alignment may also be automated by means of tools such as CATS (Štajner et al., 2018), which extracts similar sentences according to character n-grams, the average of word embeddings (WAVG) in the sentence or the continuous word alignment-based similarity analysis model. Another tool is MASSAlign (Paetzold et al., 2017), which aligns paragraphs or sentences by computing a TF-IDF-based similarity matrix of items between each pair of texts, and a vicinity procedure to complete the alignment iteratively.

Nonetheless, we experimented with a state-of-the-art procedure, BERT-based Sentence Embeddings (Reimers and Gurevych, 2019). With this method, we obtained an embedding representation of each sentence, and compute the cosine similarity between the professional and simplified version. We applied a threshold set empirically on the cosine similarity score ($cosine > 0.6$). We discarded sentences that were not similar enough, to obtain a subset to be revised manually later. We provide a companion python jupyter notebook to reproduce our method.[7]

3.6.2 Alignment criteria

We adopted the guidelines from Grabar and Cardon (2018) to align pairs of sentences (e.g. identical pairs are not used), and we added new rules. Table 4 summarizes our criteria.

We followed these criteria and rejected those sentences that each pair of annotators per data batch judged as bad alignments. Disagreements were discussed to achieve a consensus, and the medical practitioner solved specific questions about medical aspects of the contents. We aligned a total of 3800 sentence pairs (149 862 words). The average inter-annotator agreement between experts was of $kappa = 0.839$ (±0.076), which represents an *almost perfect agreement*.

4 Conclusions and future work

This work has presented the CLARA-MeD corpus of comparable (professional/laymen) medical texts in Spanish, a new contribution for analysing text simplification strategies and conduct experiments on text simplification tasks. A limitation of our work is the scarce data obtained to train and test data-intensive methods (e.g. deep-learning). More

[7] https://github.com/lcampillos/CLARA-MeD/

1. We prioritize aligning one-to-one sentences; however, in some cases, one simplified sentence needs to be aligned with two professional ones, and vice versa:

P: *Linfoma folicular recidivante/resistente* ('Relapsed/refractory follicular lymphoma')
S: *El linfoma folicular es un cáncer que afecta a los glóbulos blancos llamados linfocitos. El término recaída o refractaria indica una enfermedad que vuelve a crecer o no responde al tratamiento* ('Follicular lymphoma is a cancer that affects white blood cells called lymphocytes. The term relapsed or refractory indicates disease that grows again or does not respond to the treatment.')

2. Sentence pairs that only differ in punctuation or functional words (e.g. prepositions or adverbs) are not aligned if the simplified version does not have a simpler structure.

3. Simplified sentences that have unintelligible acronyms without their explanation or expansion are not used (we except widely-used acronyms: e.g. *SIDA*, 'AIDS'):

P: *Cancer colorectal (CCR)* ('Colorectal cancer (CRC)')
S: *El CCR es el desarrollo del cáncer desde el colon o el recto* (Not aligned)
('CRC is the development of cancer from the colon or rectum')

4. Professional and simplified sentences are not aligned if the simplified version presents a large loss of essential information that is present in the professional version:

P: *Tratamiento del síndrome de Hunter y deterioro cognitivo*
('Treatment of Hunter syndrome and cognitive impairment')
S: *Síndrome de Hunter: deficiencia de la enzima iduronato 2-sulfatasa* (Not aligned)
('Hunter syndrome-Iduronate-2-Sulfatase enzyme deficiency')

5. We do not align sentence pairs with incoherent data, imprecise information or contradictions between the professional and the simplified version:

P: *Diabetes mellitus tipo 1* ('Type 1 Diabetes Mellitus')
S: *Altos niveles de azucar (glucosa) en sangre* (Not aligned)
('High levels of sugar (glucose) in the blood')

6. Sentences consisting of paraphrases, definitions or explanations of technical terms are considered adequate simplified versions and can be aligned with the professional version:

P: *Colitis ulcerosa* ('Ulcerative Colitis')
S: *La colitis ulcerosa es una enfermedad inflamatoria intestinal que provoca inflamación en el revestimiento del intestino grueso con irritación e hinchazón* ('Ulcerative Colitis is a type of inflammatory bowel disease that causes the lining of the large intestine (colon) to become inflamed (irritated and swollen)') (Aligned)

7. Aligned sentences may have redundant information or elliptic words (e.g. prepositions), minor spelling, grammar or typographic errors, provided that the meaning is not distorted:

P: *Neumonía por SARS-CoV-2* ('SARS-CoV-2 infected patients with pneumonia')
S: *Neumonía COVID* ('COVID-Pneumonia') (Aligned)

Table 4: Alignment criteria with examples (*P*: professional; *S*: simplified).

drug-related documents from CIMA need to be cleaned and revised to be used. Moreover, more types of comparable data could be obtained from medical websites with two versions (professional-oriented and patient-oriented contents). In addition to this, the methodology applied by van der Bercken et al. (2019) could be useful to widen the coverage of Spanish medical texts from Wikipedia, and thus include this source in the cor-

pus. Sometimes, the sentence-level alignment might not be satisfactory because there is not always a one-to-one correspondence. In such cases, a paragraph-level alignment is a more suitable option (Devaraj et al., 2021). Lastly, the corpus is not annotated with complex words, which would be useful in complex word identification (CWI) tasks.

Even so, this is the first version of a benchmark for medical text simplification in the Spanish language. The high inter-annotator agreement values show a fine sentence-level alignment, which was assessed by linguists, a lexicographer and with the advice of a health professional. This ensures that these data are a quality benchmark for developing and testing medical text simplification systems.

Acknowledgements

We thank the reviewers for their valuable comments to improve this work. Project CLARA-MED (PID2020-116001RA-C33) funded by MCIN/AEI/ 10.13039/501100011033/, in project call: "Proyectos I+D+i Retos Investigación".

References

Barbu, E., Martín-Valdivia, M. T., Martinez-Camara, E., and Urena-López, L. A. (2015). Language technologies applied to document simplification for helping autistic people. *Expert Systems with Applications*, 42(12):5076–5086.

Campillos-Llanos, L., Valverde-Mateos, A., Capllonch-Carrión, A., and Moreno-Sandoval, A. (2021). A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC medical informatics and decision making*, 21(1):1–19.

Cardon, R. and Grabar, N. (2020). Construction d'un corpus parallèle à partir de corpus comparables pour la simplification de textes médicaux en français. *Traitement Automatique des Langues*, 61(2):15–39.

Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A., Gasperin, C., and Aluísio, S. M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. *Proc. of 10th CICLing*, 41:59–70.

Devaraj, A., Marshall, I., Wallace, B., and Li, J. J. (2021). Paragraph-level simplification of medical texts. In *Proc. of the NAACL 2021*, pages 4972–4984.

Gala, N., Tack, A., Javourey-Drevet, L., François, T., and Ziegler, J. C. (2020). Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Proc. of LREC 2020*, page 1353–1361.

Grabar, N. and Cardon, R. (2018). CLEAR - Simple corpus for medical French. In *Proc. of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9.

Kindig, D. A., Panzer, A. M., Nielsen-Bohlman, L., et al. (2004). *Health literacy: a prescription to end confusion*. Washington (DC): National Academies Press.

Klaper, D., Ebling, S., and Volk, M. (2013). Building a German/simple German parallel corpus for automatic text simplification. In *Proc. of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2013), Sofia, Bulgaria*.

Martin, L., Fan, A., de la Clergerie, É., Bordes, A., and Sagot, B. (2021). Muss: Multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Moramarco, F., Juric, D., Savkov, A., Flann, J., Lehl, M., Boda, K., Grafen, T., Zhelezniak, V., Gohil, S., Korfiatis, A. P., et al. (2021). Towards more patient friendly clinical notes through language models and ontologies. In *Proc. of the AMIA Annual Symposium*, pages 881–890.

Moreno-Sandoval, A., Torre-Toledano, D., Valverde-Mateos, A., and Campillos-Llanos, L. (2019). Estudio sobre documentos reutilizables como recursos lingüísticos en el marco del desarrollo del plan de impulso de las tecnologías del lenguaje. *Procesamiento del Lenguaje Natural*, 63:167–170.

Paetzold, G., Alva-Manchego, F., and Specia, L. (2017). Massalign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4.

Palmero Aprosio, A., Tonelli, S., Turchi, M., Negri, M., and Di Gangi Mattia, A.

(2019). Neural text simplification in low-resource conditions using weak supervision. In *Workshop on Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen)*, pages 37–44.

Petersen, S. E. and Ostendorf, M. (2007). Text simplification for language learners: a corpus analysis. In *Workshop on speech and language technology in education*. Citeseer.

Rauf, S. A., Ligozat, A.-L., Yvon, F., Illouz, G., and Hamon, T. (2020). Simplification automatique de texte dans un contexte de faibles ressources. In *Actes 6e conférence Traitement Automatique des Langues Naturelles (TALN), vol. 2*, pages 332–341.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal*, 312(7023):71–72.

Saggion, H., Gómez-Martínez, E., Etayo, E., Anula, A., and Bourg, L. (2011). Text simplification in simplext: Making texts more accessible. *Procesamiento del lenguaje natural*, (47):341–342.

Sakakini, T., Lee, J. Y., Duri, A., Azevedo, R. F., Sadauskas, V., Gu, K., Bhat, S., Morrow, D., Graumlich, J., Walayat, S., et al. (2020). Context-aware automatic text simplification of health materials in low-resource domains. In *Proc. of the 11th LOUHI Workshop*, pages 115–126.

Scarton, C., Paetzold, G., and Specia, L. (2018). Simpa: A sentence-level simplification corpus for the public administration domain. In *Proc. of LREC 2018*, pages 4333–4338.

Štajner, S., Franco-Salvador, M., Rosso, P., and Ponzetto, S. P. (2018). CATS: A tool for customized alignment of text simplification corpora. In *Proc. of LREC 2018*, pages 3895–3903.

Tonelli, S., Aprosio, A. P., and Saltori, F. (2016). SIMPITIKI: a Simplification corpus for Italian. In *CLiC-it/EVALITA*, pages 4333–4338.

Van den Bercken, L., Sips, R.-J., and Lofi, C. (2019). Evaluating neural text simplification in the medical domain. In *Proc. of the World Wide Web Conference*, pages 3286–3292.

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017). Multilingual and cross-lingual complex word identification. In *Proc. of the Int. Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 813–822.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proc. of the 23rd Intern. Conference on Computational Linguistics (COLING 2010)*, pages 1353–1361, Beijing, China.

IberLEF 2022: Resúmenes de las tareas de evaluación

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 199-205 recibido 05-07-2022 revisado 22-07-2022 aceptado 25-07-2022

ABSAPT 2022 at IberLEF: Overview of the Task on Aspect-Based Sentiment Analysis in Portuguese

Resumen de la Tarea de Análisis de Sentimientos Basado en Aspectos en Portugués (ABSAPT) en IberLEF 2022

Felix L. V. da Silva[1], Guilherme da S. Xavier[1], Heliks M. Mensenburg[1], Rodrigo F. Rodrigues[1], Leonardo P. dos Santos[3], Ricardo M. Araújo[1 2], Ulisses B. Corrêa[1 2], Larissa A. de Freitas[1 2]

[1] Center for Technological Development (CDTec),
Federal University of Pelotas (UFPel), Pelotas, RS, Brazil
[2] Artificial Intelligence Innovation Hub (H2IA),
Federal University of Pelotas, Pelotas, RS, Brazil
[3] University of São Paulo, São Paulo, SP, Brazil
{flvdsilva, gdsxavier, hmmersenburg, rfrodrigues, ricardo, ub.correa, larissa}@inf.ufpel.edu.br,
leonardosantsper@gmail.com

Resumen: Este artículo presenta la Tarea sobre Análisis de Sentimientos basado en Aspectos en Portugués (ABSAPT), realizada en el IberLEF 2022. Les pedimos a los participantes que desarrollaran sistemas capaces de identificar aspectos (AE) y extraer la polaridad (ASC) en textos escritos en portugués. Doce equipos se inscribieron en la tarea, entre los cuales cinco presentaron predicciones e informes técnicos. El sistema con mejor rendimiento logró un valor de precisión (Acc) de 0,67 para la subtarea de AE (Equipo Deep Learning Brasil) y un valor de precisión equilibrada (Bacc) de 0,82 para la subtarea de ASC (Equipo Deep Learning Brasil).
Palabras clave: Análisis de Sentimiento basado en Aspectos, Portugués, Reseñas de Hoteles.

Abstract: This paper presents the task on Aspect-Based Sentiment Analysis in Portuguese (ABSAPT), held within Iberian Languages Evaluation Forum (IberLEF 2022). We asked the participants to develop systems capable of extracting aspects (AE) and classifying sentiment of aspects (ASC) in texts. We created one *corpora* containing reviews about hotels. Twelve teams registered to the task, among which five submitted predictions and technical reports. The best performing system achieved an Accuracy (Acc) value of 0.67 in AE sub-task (Team Deep Learning Brasil) and a Balanced Accuracy (Bacc) value of 0.82 in ASC sub-task (Team Deep Learning Brasil).
Keywords: Aspect-Based Sentiment Analysis, Portuguese, Hotel Reviews.

1 Introduction

Sentiment Analysis (SA) is the field of Natural Language Processing (NLP) that automatically analyzes people's sentiments or opinions towards some entity. These sentiments can be valuable sources of information about the consumer's feelings about a particular product or idea, which can help in decisions by companies or governments (Liu, 2015).

SA can be done on different levels, focusing mainly on three possible granularity levels: document level, sentence level, and aspect level (de Freitas, 2015). At the aspect level, it is possible to analyze different opinions held towards different aspects of some entity or different entities in the same document or sentence.

The ABSAPT 2022 task aims to challenge different teams to propose techniques capa-

ISSN 1135-5948. DOI 10.26342/2022-69-17

ble of extracting aspects and classifying sentiment of aspects in the hotel reviews. The present paper presents an overview of the task. First, we briefly present some theoretical reflections Aspect-Based Sentiment Analysis (ABSA) (Section 2) and describe the proposal of our task (Section 3). Section 4 presents the *corpora* description and the annotation process. In Section 5, we describe the evaluation measures. Participant systems and the results are discussed in Section 6. Finally, the final remarks are done in Section 7.

2 Aspect-Based Sentiment Analysis

On this granularity level, all opinions expressed towards any aspect of the entity are analyzed individually. This level allows a better understanding of the opinions and entities in the text. To accomplish the analysis on this level, the task used to be broken into two sub-tasks: Aspect Extraction (AE) and Aspect Sentiment Classification (ASC).

2.1 Aspect Extraction

This sub-task determines which aspects of a given entity are considered in a text.

For example, in the sentence "Hotel com boa **localização**" ["Hotel with good **localization**"], the goal of AE would be to identify the aspect '**localização**'.

2.2 Aspect Sentiment Classification

This sub-task consists of the classification of the polarity for each aspect that has been identified in the text.

For example, in the sentence "Hotel com boa **localização**" ["Hotel with good **localization**"], the goal of ASC would be to classify the aspect '**localização**' as positive.

3 Task Description

People's opinions are a great source of information for other people and organizations, public or private. Typically works focused on Portuguese perform document level SA. It is hard to find ABSA approaches or datasets available for Portuguese.

We propose to create an ABSA for TripAdvisor reviews written in Portuguese. Two sub-tasks will be available: AE and ASC. The first sub-task comprehends the identification of aspects in the reviews, and the second subtask proposes to extract the sentiment orientation (polarity) of the review about a single aspect mentioned in it.

The availability of corpora written in Portuguese is scarce, which limits the amount of research done for this language.

This task will contribute to the progress of Portuguese NLP, as there is a demand for developing new methods and tools.

Previous ABSA competitions, such as SemEval 2014 (Pontiki et al., 2014), SemEval 2015 (Pontiki et al., 2015), SemEval 2016 (Pontiki et al., 2016) and EVALITA (Mattei et al., 2020) inspired us to develop a specific task for Portuguese.

4 Corpora Description and Annotation Process

This section describes the *corpora* proposed for evaluation and the annotation process (annotation guidelines and inter-annotator agreement).

4.1 *Corpora* Description

The corpora contain travellersŕeviews about accommodation services companies, written in Portuguese. In ABSAPT 2022, we used corpora developed previously by Freitas (de Freitas, 2015) and Corrêa (Corrêa, 2021). Freitasćorpus is publicly available, so it will be used only in the training dataset (3111 samples from 847 reviews). Corrêa's corpus is private and will be split into training and test dataset (257 samples to AE and 686 samples to ASC). The full dataset will be available after the event on the website http://absapt2022.tk/.

Both datasets were annotated following the same annotation guidelines (de Freitas, 2015). The concepts on the Accommodation Services Domain Ontology, HOntology (Chaves, de Freitas, and Vieira, 2012) were aspects annotated. HOntology contains 282 concepts categorized into 16 top-level concepts. The concept hierarchy has a maximum depth of 5.

In Figure 1, one can see an overview of the training dataset. The full training dataset contains 77 aspects. Due to space limitation, we present the 40 most frequent aspects and the polarities distribution. In Figure 2, we present an overview of the test dataset.

4.2 Annotation Process

The manual annotation of Freitas' corpus (training dataset) was conducted by two an-

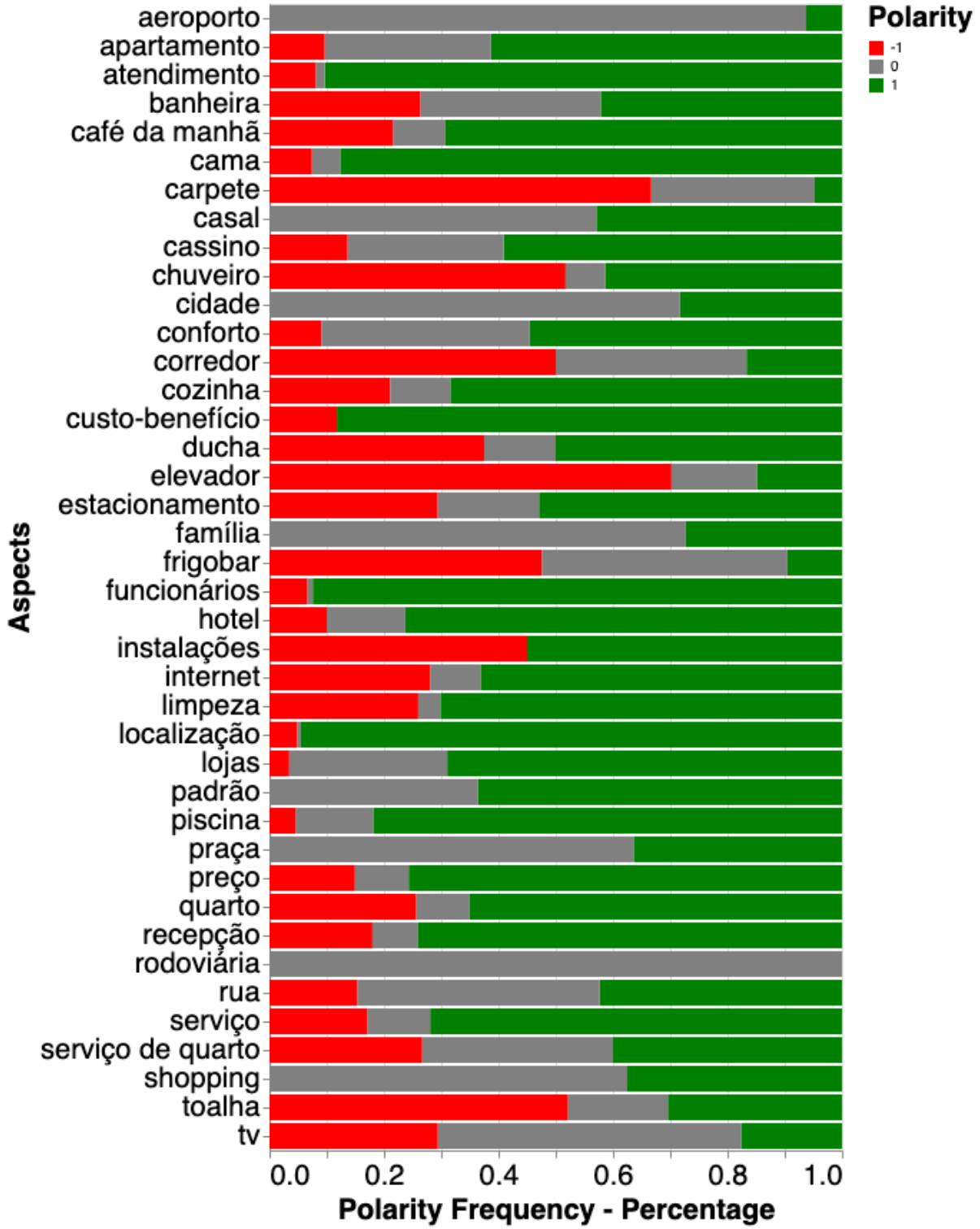


Figure 1: Training dataset: polarity frequency for aspects with at least 10 samples.

notators, both native speakers of Portuguese, one linguist, and one computer scientist. And the manual annotation of Corrêa's corpus (training and test dataset) was conducted by twelve students and professors of computer science and engineering annotators. Both used the tool developed in the context of (de Freitas, 2015).

The agreement between annotators (Freitas' corpus) was measured with Kappa Statistics (Landis and Koch, 1977). The annotators agreement about ABSA from do-

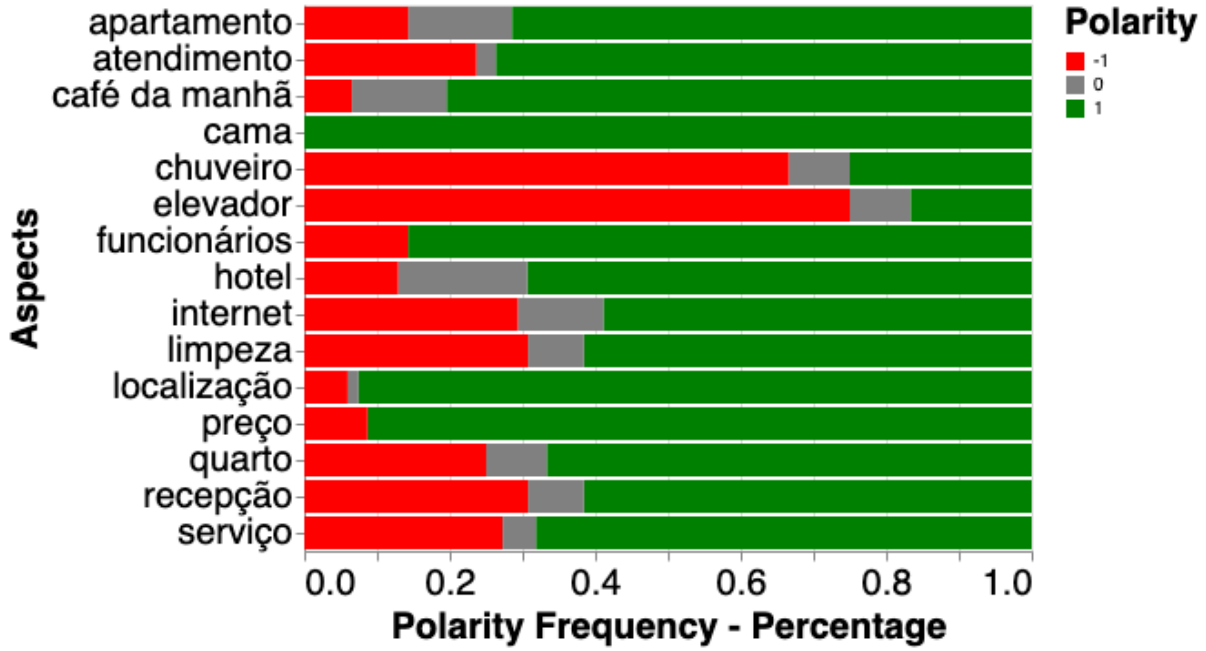


Figure 2: Test dataset: polarity frequency for aspects with at least 10 samples.

main ontology using Kappa was 0.58 for explicit aspects, which is considered a moderate agreement. We believe that the annotation has an acceptable Kappa value. It is also important to note that only in a few cases the annotators disagreed between negative and positive polarities, the majority of disagreements was about positive and neutral polarities, or negative and neutral polarities.

The agreement between annotators (Corrêa's corpus) was measured with Fleiss Kappa (Fleiss, 1971) suitable for more than two annotators. The majority annotator group share k $>$ 0.4 (moderate agreement).

5 Evaluation Measures

The training set was released on April 08th, and participants had sixteen days to train their systems. The test set was released on April 24th, and each participant had twelve days to submit one run.

Participating teams will receive training and test datasets. The latter was sent without the label of the samples.

We evaluated the predictions sent by the participants using several metrics: Acc (Eq. 1), Precision (Eq. 2), Recall (Eq. 3), F1 (Eq. 4), and Bacc (Eq. 6). Bacc to rank competitors. The submissions will be ranked according to Acc in AE sub-task and Bacc in ASC sub-task.

$$Acc = \frac{True\ Positives + True\ Negatives}{Total\ Number\ of\ Instances} \quad (1)$$

$$Precision = \frac{True\ Positives}{True\ Positives + True\ Negatives} \quad (2)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

$$F1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives} \quad (5)$$

$$Bacc = \frac{(Recall + Specificity)}{2} \quad (6)$$

6 Participants Systems and Discussion of the Results

Twelve teams registered for the task, among which five submitted predictions and technical reports. Participants are from universities

and institutes in Brazil (UFG, UFPI, IFPI, UFSCAR, USP, UFPR and UESC).

Participants used rules and lexicon (Team UFSCAR), traditional machine learning approaches as CRF (Team NILC and Team UFPR), and deep learning methods as Transformers (Team Deep Learning Brasil, Team PiLN, and Team UFPR).

Tables 1 and 2 present participants' results for each sub-task submitted run. The results are ranked according to the Acc in AE and Bacc in ASC.

Accuracy	Team
0.67	**TeamDeepLearningBrasil**
0.65	TeamPiLN
0.59	TeamUFSCAR
0.22	TeamNILC
0.17	TeamUFPR

Table 1: Participants results ranked in terms of Acc in AE sub-task.

For each system, best run is highlighted in bold. Team Deep Learning Brasil, used transformers to achieve an Acc of 0.67 in AE and a Bacc of 0.82 in ASC.

Below we summarize the proposed approach of each team:

- **Team Deep Leaning Brasil:** The authors proposed different methodologies for both sub-tasks of ABSA. The AE used a single sentence tagging approach, and the ASC tested with two different strategies, one as a Sentence Pair Classification and the other as a Conditional Text Generation. In addition, also were used other ABSA datasets such as Evalita, MAMs, Semeval 2014, 2015, and 2016 competition, and Bidirectional Encoder Representations from Transformers (BERT) pre-trained models on the Portuguese language and multilingual. The proposed approach reached the best-performing systems, achieving new state-of-the-art results on both subtasks (Gomes et al., 2022).
- **Team PiLN:** The authors proposed simple and well-known approaches to the sub-tasks of ABSA. The AE used a string-match strategy using a multilingual ontology for the accommodation sector named HOntology. The ASC used a BERTimbau (BERT pre-trained model on the Portuguese language), approaching the reviews as a Sentence Pair Classification. The proposed approach reached the second best-performing system, achieving the following results: Acc of 0.65 in AE, Bacc of 0.78, F1 of 0.77, Precision of 0.76, Recall of 0.78 in ASC (Neto et al., 2022).
- **Team UFSCAR:** The AE sub-task includes preprocessing, tokenization, feature extraction, a lexicon, and rule-based aspect identification. The ASC sub-task has two main steps: meaningful surroundings extraction and sentiment extraction using GoEmotions (Demszky et al., 2020), followed by polarity extraction. The proposed approach reached the third best-performing system, achieving the following results: Acc of 0.59 in AE, Bacc of 0.62, F1 of 0.61, Precision of 0.65, Recall of 0.62 in ASC (Assi et al., 2022).
- **Team NILC:** The authors participated only in AE sub-task. Its approach is based on the Conditional Random Fields (CRF) machine learning algorithm combined with a post-processing step. After applying the method, the authors performed an error analysis of detected and non-detected aspects (Machado and Pardo, 2022).
- **Team UFPR:** In the AE used CRF, the dataset was adapted with tokenization and the POS Tagging technique, and a pre-trained model for Portuguese was used for the POS Tagging process. If the POS Tagging process has a better adaptation for Portuguese, there will be a gain in results for the CRF performance. In the ASC used BERTimbau (Heinrich and Marchi, 2022). The proposed approach reached the worst run, achieving the Acc of 0.17 in AE.

7 Final Remarks

Motivated by the necessity of improvements in the ABSA task focused on Portuguese, we proposed a task within the IberLEF 2022. This paper overviews the first task on ABSA in Portuguese to identify aspects and extract the polarity in hotel reviews.

The datasets (training and test) have been manually annotated. The inter-annotator

Bacc	F1	Precision	Recall	Team
0.82	**0.81**	**0.81**	**0.82**	**TeamDeepLearningBrasil**
0.78	0.77	0.76	0.78	TeamPiLN
0.62	0.61	0.65	0.62	TeamUFSCAR
0.62	0.61	0.65	0.62	TeamUFPR

Table 2: Participants results ranked in terms of Bacc in ASC sub-task.

agreement for training and test dataset is considered moderate.

The deep learning methods based on Transformers performed better than approaches based on rules and lexicons. The Deep Learning Brasil Team achieved a Bacc of 0.67 for the AE and Acc of 0.82 for the ASC, while the UFSCAR Team achieved a Bacc of 0.59 for the AE and Acc of 0.62 for the ASC.

Acknowledgments

Thank you to all annotators for their essential work. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. This work was financed in part by the following Brazilian research agencies: CAPES and CNPq.

References

Assi, F. M., G. B. Candido, L. N. dos Santos Silva, D. F. Silva, and H. de Medeiros Caseli. 2022. Ufscar's team at absapt 2022: Using syntax, semantics and context for solving the tasks. In *Proceedings of the Iberian Languages Evaluation Fórum (IberLEF 2022), co-located with the 38th Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), Online. CEUR.org.*

Chaves, M. S., L. A. de Freitas, and R. Vieira. 2012. Hontology: A multilingual ontology for the accommodation sector in the tourism industry. In J. Filipe and J. L. G. Dietz, editors, *KEOD*, pages 149–154. SciTePress.

Corrêa, U. B. 2021. *Análise de sentimento baseada em aspectos usando aprendizado profundo: uma proposta aplicada à língua portuguesa.* Ph.D. thesis, Universidade Federal de Pelotas, Pelotas.

de Freitas, L. A. 2015. *Feature-level sentiment analysis applied to Brazilian Portuguese reviews.* Ph.D. thesis, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

Demszky, D., D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July. Association for Computational Linguistics.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gomes, J. R. S., R. C. Rodrigues, E. A. S. Garcia, A. F. B. Junior, D. F. C. Silva, and D. F. Maia. 2022. Deep learning brasil at absapt 2022: Portuguese transformer ensemble approaches. In *Proceedings of the Iberian Languages Evaluation Fórum (IberLEF 2022), co-located with the 38th Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), Online. CEUR.org.*

Heinrich, T. and F. Marchi. 2022. Teamufpr at absapt 2022: Aspect extraction with crf and bert. In *Proceedings of the Iberian Languages Evaluation Fórum (IberLEF 2022), co-located with the 38th Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), Online. CEUR.org.*

Landis, J. R. and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Liu, B. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* Cambridge University Press.

Machado, M. T. and T. A. S. Pardo. 2022. Nilc at absapt 2022: Aspect extraction for portuguese. In *Proceedings of the Iberian Languages Evaluation Fórum (IberLEF 2022), co-located with the 38th Conference*

of the Spanish Society for Natural Language Processing (SEPLN 2022), Online. CEUR.org.

Mattei, L. D., G. D. Martino, A. Iovine, A. Miaschi, M. Polignano, and G. Rambelli. 2020. Ate absita@ evalita2020: Overview of the aspect term extraction and aspect-based sentiment analysis task. *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), Online. CEUR.org.*

Neto, F. A. R., R. F. de Sousa, R. L. de S. Santos, R. T. Anchiêta, and R. S. Moura. 2022. Piln at absapt 2022: Lexical and bert strategies for aspect-based sentiment analysis in portuguese. In *Proceedings of the Iberian Languages Evaluation Fórum (IberLEF 2022), co-located with the 38th Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), Online. CEUR.org.*

Pontiki, M., D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. D. Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S. M. Jiménez-Zafra, and G. Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California, June. Association for Computational Linguistics.

Pontiki, M., D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado, June. Association for Computational Linguistics.

Pontiki, M., D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 207-215 recibido 05-07-2022 revisado 22-07-2022 aceptado 25-07-2022

Overview of DA-VINCIS at IberLEF 2022: Detection of Aggressive and Violent Incidents from Social Media in Spanish

Resumen de la Tarea DA-VINCIS en IberLEF 2022: Detección de Incidentes Violentos en Redes Sociales en Español

Luis Joaquín Arellano[1], **Hugo Jair Escalante**[1], **Luis Villaseñor-Pineda**[1,2], **Manuel Montes-y-Gómez**[1], **Fernando Sanchez-Vega**[3,4,5]
[1]Laboratorio de Tecnologías del Lenguaje (INAOE), Mexico.
[2]Centre de Recherche GRAMMATICA (EA 4521), Université d'Artois, France
[3]Mathematics Research Center (CIMAT), Guanajuato, Mexico.
[4]El Colegio de México (COLMEX), Mexico
[5]Consejo Nacional de Ciencia y Tecnología (CONACYT), Mexico.
{arellano.luis, hugojair, villasen, mmontesg}@inaoep.mx
fernando.sanchez@cimat.mx

Abstract: This paper presents the overview of the DA-VINCIS 2022 task, organized at IberLEF 2023 and co-located with the 38th International Conference of the Spanish Society for Natural Language Processing (SEPLN 2022). DA-VINCIS challenged participants to develop automated solutions for the detection of violent events mentioned in social networks. We released a novel corpus collected from Twitter and manually labeled with 4 categories of violent incidents (plus the no-incident label). The shared task focused on the Mexican variant of Spanish and it was divided into two tracks: (1) a binary classification task in which users had to determine whether tweets were associated to a violent incident or not; and (2) a multi-label classification task in which the category of the violent incident should be spotted. More than 40 teams registered for the task and 12 participants submitted predictions for the final phase. Very competitive results were reported in both sub tasks, where transformer-based solutions obtained the best results. Corpora and results are available at the shared task website at https://codalab.lisn.upsaclay.fr/competitions/2638.
Keywords: DA-VINCIS, violent event detection, text classification.

Resumen: Se presenta el resumen de la tarea DA-VINCIS 2022, organizada en IberLEF 2022 junto a la 38ª Conferencia Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2022). DA-VINCIS plantea el reto de detectar automáticamente piezas de información en redes sociales que estén asociadas a eventos violentos. Se liberó un nuevo corpus para el Español Mexicano que fue etiquetado manualmente con 4 categorías de eventos violentos (además de la categoría no-violento). Se propusieron dos subtareas: (1) una tarea de clasificación binaria donde se buscaba distinguir tuits asociados a eventos violentos del resto; y otra (2) donde se buscaba identificar la categoria del evento violento. Más de 40 participantes se registraron en el portal y 12 enviaron resultados para la fase final. Los resultados obtenidos fueron muy competitivos para ambas tareas; las soluciones que obtuvieron los mejores resultados se basaron en modelos tipo *transformer* para el español. El corpus y los resultados detallados pueden consultarse en el sitio web de la tarea: https://codalab.lisn.upsaclay.fr/competitions/2638.
Palabras clave: DA-VINCIS, Detección de eventos violentos, Clasificación de textos.

ISSN 1135-5948. DOI 10.26342/2022-69-18

1 Introduction

Violence has obvious negative effects on those who witness or experience it, including a higher incidence of depression, anxiety, post-traumatic stress disorder, among others. In addition, violence events have a high impact for governments, as they are in charge of guaranteeing security to their population. Therefore, the detection and tracking of violence related events is critical. In this context, social networks comprise a valuable information source for the detection and monitoring of violent events, as people very often post publications notifying the occurrence of violent events in real time. This represents an important opportunity for IT researchers that can provide solutions based on natural language processing for the timely detection of violent incidents in social networks. Solutions of this kind could be used by authorities to respond more efficiently to events happening in real time, and to develop crime prevention policies according to geographical zones and types of events. Likewise, such solutions would be very helpful to the population, as one could know what violent events are happening in which zones in real time.

We organized a shared task collocated with IberLEF2022 called DA-VINCIS. This task focused on the detection of violent incidents on Twitter. It challenged participants to develop methods able to classify tweets as reporting a violent event or not. For this first edition, the shared task targeted Spanish in its Mexican variant. This is motivated by the lack of resources in Spanish for approaching the task, and the fact that Mexican Spanish is the most spoken variant of this language[1]. We released a novel corpus carefully labeled according to violent event categories. The shared task comprised two tracks: *violent event identification* and *violent event category recognition*. Labeled data was provided to participants for both tracks for the development of their solutions, and unlabeled data was used for the final evaluation of the corresponding tracks.

As far as we know, this was the first shared-task aiming at detecting violent events from social media. This is an issue that has received little attention from the community, despite its enormous potential impact. Therefore, the aim of the challenge was to motivate research on a topic little explored in Spanish, but with great potential impact for the whole population and authorities. In addition, an implicit goal was to raise awareness of the relevance of this problem.

The task posed several challenges to the community, including: dealing with Mexican Spanish, the ambiguous language inherent to Twitter, the high class imbalance ratios present in our datasets, among others. We are confident that the shared task will give rise to novel solutions that could be used in the near future for applications of societal impact, for example, generating real-time occurrence crime maps. Last but not least, we plan to release the associated corpus in the near future so that the community can keep working on it even at the end of the shared task.

The remainder of this paper is organized as follows. Section 2 describes the shared task in detail. Then, Section 3 introduces the DA-VINCIS corpus. Section 4 presents the results obtained and a summary of participants' solutions. Finally, Section 5 outlines conclusions and future work directions.

2 Task description

As previously mentioned, the DA-VINCIS shared task comprised two tracks: a binary classification subtask that aimed at distinguishing tweets associated to violent incidents from those that are not; and (2) a task that challenged participants to identify the type of violent incident (if any) being reported in tweets. The categories considered for the latter task are described in Table 2.

The DA-VINCIS corpus, described in detail in the next section, was used for the evaluation of both subtasks. The challenge was run in the CodaLab platform (Pavao et al., 2022). The shared task was divided into two stages as follows:

- **Development phase.** Participants were provided with labeled training data and unlabeled validation data. During this phase, which lasted about two months, participants were able to submit predictions for the validation set and receive immediate feedback in the CodaLab site.
- **Final phase.** Participants were provided with unlabeled test data. They were able to upload up to five submis-

[1] In terms of the number of native speakers.

Reference	Considered categories
(Mata Rivera et al., 2016)	Theft, Crime, Theft with violence, Theft walking, Theftin car, Theft without violence
(Sandagiri, Kumara, and Kuhaneswaran, 2020b)	Assault, Burglary, Drugs Violations, Homicide, Sex Offences, Suicide
(Sandagiri, Kumara, and Kuhaneswaran, 2020a)	Assault, Burglary, Drugs Violations, Homicide, Sex Offences
(Piña-García and Ramírez-Ramírez, 2019)	Robbery passerby, Theft of motor vehicle, Robbery of business property, Card fraud, Homicide, Domestic burglary, Robbery on public transportation, Rape, Firearm injuries, Robbery in subway, Robbery on taxi, Robbery to Carrier, Robbery to deliver person
DA-VINCIS	Accident, Homicide, Theft, Kidnapping, Non-incident

Table 1: Violent incidents considered in previous work.

sions during the competition. Performance on the test set was used to rank participants.

For subtask 1, recall, precision and f_1 score with respect to the *violent-incident* class were considered as evaluation measures. For subtask 2, macro average recall, precision and f_1 score were considred. In both cases, the leading evaluation measure was that of f_1 score.

3 DA-VINCIS corpus

The DA-VINCIS corpus is a collection of tweets[2] associated to reports of violent incidents in Mexican Spanish. The aim of this novel corpus is to boost research in the automated detection and monitoring of violent incidents in social networks. Summarizing, a large number of tweets was retrieved using queries associated to predefined categories. Then, the tweets were filtered, and a subset of these was manually labeled. In the remainder of this section we provide details on the construction of this corpus.

A set of categories of violent incidents was defined after a careful analysis of relevant literature, see Table 1. The categories considered in each study differ according to the legal, psycho-social or geographical context, and commonly they are finally filtered by the criteria of the research group involved.

[2]Please note that the corpus is formed by both, the text in tweets and their associated images, if any, for this shared task only textual information was considered.

The categories considered in the DA-VINCIS corpus are shown in the last row of Table 1. The criteria for selecting such categories involved: categories that appear in most of previous studies (e.g., *Homicide* and *Theft*), generic categories (e.g., we considered a single Theft category) and categories that appeared most frequently among in Twitter accounts associated to local news in Mexico (e.g., *Kidnapping*). Finally, our choice for these categories relied on the number of tweets that we retrieved per each category.

It is important to mention that since the long term goal of this project is the real-time monitoring of violent incidents, the *Accident* category was taken into account. As on a daily basis authorities use the same communication channels to deal with this kind of problem. Categories such as *Sexual offences* and *Drug violations* were initially considered because of their relevance and the urgent need to prevent them. However the study of these categories is particularly complicated, because although there are reports or complaints on the internet, these are not frequent, in addition to the fact that these are common topics of conversation and discussion, therefore it makes it extremely difficult to find the reports among the large number of opinions. Definitions for the categories considered in the DA-VINCIS corpus are shown in Table 2.

To obtain keywords for the retrieval of violent incidents tweets, a research work was carried out where 30,000 tweets published in news accounts in Spanish were recovered, 5,000 tweets were manually tagged to identify if the news was violent (i.e. binary labeling) once established, an ML model was applied to label the rest of the corpus, the pseudo labels obtained were used to study the tweets and the unigrams, bigrams and trigrams that provided the most information for the classification, the most significant words from the top-100 were filtered, these were the keywords used to search for tweets of violent incidents.

Once the keywords were obtained, a tweet retrieval was performed using each of the selected keywords, where it was required that (1) tweets had an associated image, (2) language was Spanish and, (2) the tweet was geolocated in approximation to Latin America. The result of this process were 8000 tweets that were further filtered by eliminating those

Category	Definition
Accident	Eventual event or action that results in involuntary damage to people or things.
Homicide	Deprivation of life.
Theft	Seizure or willful destruction of someone else's property without the right and without the consent of the person who can legally dispose of them.
Kidnapping	Deprivation of liberty.
Non-incident	Selected when there is no crime reported.

Table 2: Definition of the categories considered in the DA-VINCIS corpus.

that could no longer recover any of their elements, that were written in a language other than Spanish but that were filtered in the search and the empty elements or that only consisted of a series of hashtags.

The filtered dataset was formed by 5000 tweets. Each tweet in the dataset was labeled by at least two annotators. Labels assigned by annotators took into account the context provided by the text of the tweet and the associated images, if any. However, even when having all the context available, the labeling process was not straightforward in some cases. Sometimes the images were conducive to confusion or vice versa. For example, confusing a traffic accident with a homicide with a car (cyclist hit by a car).

Randomly selected tweets from each category are shown in Table 3. Despite this samples were assigned the correct label, there were some samples that could be considered noisy, see Section 4. Table 4 shows the proportion of samples per class in the dataset. Please note that since categories are not necessarily disjoint (except the no-incident one). More than one label can be assigned to a single tweet, that is why the total number of labels is different from 5000.

To analyze the difficulty of the task, the agreement between the annotators was calculated. Table 5 presents the results of the Kappa coefficient, by the number of judgments collected. The coefficient values indicate moderate agreement, see (McHugh, 2012). However, we found some samples with noisy annotations, see Section 4, evidencing the need for a detailed curation for the DA-VINCIS corpus.

4 Participants approaches and results

In the following subsections we describe the main ideas addressed by the different participants, and present a general analysis of their results.

4.1 Systems' descriptions

A total of 12 teams participated in the DA-VINCIS shared-task; the majority tackled both subtasks, however, four teams only addressed the subtask 1 and one team only presented a solution for subtask 2. Something interesting to highlight is that the teams with the best performance in each of the two subtasks presented a proposal exclusively focused on the particular subtask. This indicates that both subtasks have their own specific challenges and, therefore, that it is not always convenient to approach them using the same strategy.

From the different solutions presented at the DA-VINCIS shared task, we found several coincidences, which indeed align with some general trends in Natural Language Processing. The main shared aspects correspond to the use of:

- **Pretrained Transformers**: All participant approaches used some pretrained transformer to take advantage of all the knowledge encoded in their pretraining. Some applied the traditional fine-tuning (Ta et al., 2022b), while others proposed some interesting modifications (Vallejo-Aldana, López-Monroy, and Villatoro-Tello, 2022; Turón et al., 2022; Montañés-Salas, del Hoyo-Alonso, and Peña-Larena, 2022; García-Díaz et al., 2022; Ta et al., 2022b). On the other hand, some approaches used the pretrained transformer as a frozen source of knowledge, only using the contextual embedding encodings (García-Díaz et al., 2022), or extracting relationships from the instances and task description using a Prompt-based framework (Qin et al., 2022).
- **Ensembles**: Multiple approaches employed ensembles to take advantage of variations of their base solution models. For example, the majority voting scheme was successfully used for subtask 1 (Vallejo-Aldana, López-Monroy, and Villatoro-Tello, 2022; Turón et al., 2022;

Categories	Original text	Translation
Accident	*#Ahora Reportan accidente de tránsito en el ingreso al municipio de Salcajá. Dos vehículos tipo picop involucrados en el percances. Precaución al conducir por el sector. Ampliaremos la información. #Stereo100Noticias*	#Now Car accident is being reported at the entrance of the Salcajá municipality. Two pickup vehicles involved. Caution when driving nearby. We will extend the information #Stereo100Noticias
Homicide	*La violencia y las ejecuciones continúan cada día en la CDMX un hombre fue ejecutado a 2 calles de la alcaldía de Cuahutémoc en la calle de Pedro Moreno*	Violence and killings continue everyday in CDMX a mean was killed two blocks from Cuahutémoc town hall in Pedro Moreno street
Theft	Imágenes en las que un sujeto que ingresó a robar a un local ubicado en Av.Tonalá y Madero en la Cabecera Municipal. El hombre iba armado y después del robo huyó en un auto Kia color gris que lo esperaba afuera del local.	footage in which a subject that entered to steal a facility in Av. Tonalá and Madero in the municipality. The man was armed and after the robbery escaped in a grey Kia that was waiting outside the facility.
Kidnapping	Secuestraron a sujeto frente al palacio municipal de Coatzacoalcos A plena luz del día realizan acto delictivo; los detienen y desarticula UECS banda de plagiarios recién formada; se quedan en el Cereso Duport Ostión	A man was kidnapped in front of Coatzacoalcos' town all. The criminal act was perfored during daylight; they were arrested and the UECS dismantled a band of kidnappers just formed; they are staying in the Duport Ostión prision.

Table 3: Samples from the DA-VINCIS corpus for the violent incident categories.

Montañés-Salas, del Hoyo-Alonso, and Peña-Larena, 2022). More sophisticated ensemble techniques were also applied, such as intermediate fusion of NNs using Knowledge Integration (KI), ensemble learning (García-Díaz et al., 2022), and a kind of multilevel fusion that incorporates information from multiple sources (Qin et al., 2022).

- **Multi-Task Learning (MTL)**: Several proposals took advantage of the pairing of subtasks 1 and 2 to carry out some kind of multi-task learning. For example, (Vallejo-Aldana, López-Monroy, and Villatoro-Tello, 2022) performed MTL for subtask 1 through a binary transformation of subtask 2 that is jointly learned in order to incorporate additional information for subtask 1. In contrast, in (Ta et al., 2022b) the prediction of each class of subtask 2 was transformed into a binary problem that is performed with MTL on the complete set of binary problems. Finally, (Ta et al., 2022a) proposed an interesting MTL approach where subtask 2 was carried out while jointly learning to distinguish real instances from instances generated by a GAN.

Categories	# Examples	% Total
Accident	1800	33.45
Homicide	417	7.75
Theft	286	5.31
Kidnapping	72	1.33
Non-violent	2878	53.49

Table 4: Proportion of samples from each class

Judgments	Tweets	Coefficient
2	1349	0.5758
3	1643	0.5767
4	1024	0.5979
5	350	0.5829

Table 5: Kappa coefficients by number of judgments (only results for 2 to 5 judgments are shown).

- **Data Augmentation (DA)**: This technique was also widely used by the participant teams. The most used method was back-translation (Turón et al., 2022; Montañés-Salas, del Hoyo-Alonso, and Peña-Larena, 2022; Ta et al., 2022b; Ta et al., 2022a), however, some approaches also integrated the examples in the intermediate languages to the augmented data, and in consequence used multilingual models in their training phase (Tonja et al., 2022).

- **Preprocessing**: Most teams performed standard preprocessing operations to allow the transformers-based language

models to handle the input texts. For example, they removed URLs, hashtag symbols, non-alphanumeric symbols, and adjusted user mentions (strings with @). Additionally, in (Vallejo-Aldana, López-Monroy, and Villatoro-Tello, 2022; Turón et al., 2022; Montañés-Salas, del Hoyo-Alonso, and Peña-Larena, 2022) emojis were replaced by their descriptive words, and acronyms and abbreviations were expanded in (García-Díaz et al., 2022).

Three approaches show some interesting features that do not fit the generalities described above; these are:

- **Noise Reduction**: *VICOMTECH* (Turón et al., 2022) carried out a relabelling process of the training data considering the votes of 5 systems learned from the original noisy data set. They "corrected" the instance labels if at least 4 of the 5 systems agreed to do so; using this approach they modified around 5% of the training set labels.
- **Use of Advanced Linguistic Features**: *UM-UJ-URJC* (García-Díaz et al., 2022) considered the use of a variety of features with the purpose of taking into account multiple aspects of the writing and communication style of tweets.
- **Use of Prompt Learning:** GDUT (Qin et al., 2022) employed a prompt learning module to inject information from a pre-trained language model into the violent event category recognition task. This approach incorporates the text provided by the prompt module into the tweet representation.

4.2 Evaluation campaign results

Table 6 presents the results obtained by the participant teams in subtask 1, the binary identification of violent incidentes. The teams are sorted by their F1-score over the positive class (i.e., the violent incident class); Precision and Recall are also reported to allow a better interpretation of these results. At the bottom it is included our baseline[3]

[3]Please note that during the final phase of the shared task we uploaded a single run of the baseline that obtained better results. However, in this paper we report the average over 10 runs of the performance of the baseline, which is a more reliable estimate of its performance.

Subtask 1: Binary violent event identification			
Team	**Precision**	**Recall**	**F1-Score**
CIMAT-UG-UAM-IDIAP	0.803	0.750	**0.775**
VICOMTECH	**0.812**	0.737	0.773
ITAINNOVA	0.779	0.751	0.765
UM-UJ-URJC	0.774	0.753	0.764
Sdamian	0.761	0.750	0.756
Bernardo	0.780	0.730	0.754
IPN-DLU-UNOMAHA-1	0.755	0.740	0.748
CIC-IPN	0.761	0.730	0.745
IPN-DLU-UNOMAHA-2	0.740	0.747	0.744
JuanCalderon	0.723	**0.763**	0.742
Sustaitangel	0.710	0.742	0.726
Baseline	0.763	0.780	0.750

Table 6: Results of the participant teams in Subtasks 1. They correspond to the Precision, Recall and F1 score in the positive class.

result, which corresponds to the direct use of a traditional fine-tuning (i.e., using a single linear layer and the use of softmax for classification) of BETO (Cañete et al., 2020), a well-known pre-trained language model in Spanish.

The best performance in Subtask 1 was obtained by the CIMAT team (Vallejo-Aldana, López-Monroy, and Villatoro-Tello, 2022) followed by VICOMTECH (Turón et al., 2022) These two approaches have in common that they took advantage of the parallelism of both subtasks, particularly, they included in their model for subtask 1 some information from subtask 2. The third best approach is ITAINNOVA (Montañés-Salas, del Hoyo-Alonso, and Peña-Larena, 2022), which, similarly to the CIMAT team, used an ensemble of multiple transformer-based models. On the one hand, the CIMAT approach combined the output of three BERT-based models fine-tuned to perform MTL. In this case, MTL is used to simultaneously learn the subtask 1 and a binary version of a violent event subcategory classification (i.e., each BERT model is different in the specific event subcategory chosen). On the other hand, the ITAINNOVA approach uses different pretrained models (such as BETO, Twitter-XLM-Roberta and BSC-Roberta), thus obtaining its diversity from the models and not from the data.

It should be noted that the different approaches obtained very close results; the best performance is only 6.8% greater than the lowest, and the standard deviation of the set of F1-scores is only 0.015.

The results obtained by the teams in subtask 2, the violent event category recognition, are shown in Table 7. The best performance in this subtask corresponds to the GDUT

Subtask 2: Violent event category recognition			
Team	**Precision**	**Recall**	**F1-Score**
GDUT	0.550	**0.564**	**0.554**
VICOMTECH	0.517	0.545	0.528
ITAINNOVA	0.509	0.503	0.504
CIC-IPN	0.467	0.520	0.490
CIMAT-UG-UAM-IDIAP	**0.655**	0.421	0.473
UM-UJ-URJC	0.442	0.549	0.469
Sustaitangel	0.459	0.424	0.433
CIC-IPN-DLU-UNOMAHA-1	0.377	0.438	0.392
Baseline	0.498	0.460	0.570

Table 7: Results of the participant teams in the Subtask 2. They correspond to the macro average values of Precision, Recall and F1 score.

team. Their solution is mainly characterized by the incorporation of semantic relations between the text instances and the name of the categories through the application of a Prompt learning module. The second and third best performances were obtained, as in subtask 1, by the VICOMTECH (Turón et al., 2022) and ITAINNOVA (Montañés-Salas, del Hoyo-Alonso, and Peña-Larena, 2022) teams, respectively. Their adequate performance in both tracks suggests the relevance and robustness of these two approaches for the task addressed.

From the results, it is notorious that subtask 2 is much more challenging than subtask 1; something expected due to the high imbalance in some of the categories. This is reflected in a greater standard deviation (0.051) in the reported F1-scores, and also in the larger difference between the best and worst reported results (in this case, the former is 41% greater than the later).

4.3 Analysis

To provide insights on the complimentary and redundancy of solutions, the Intra-ensemble Coincident–Failure Diversity (CFD) was calculated for the 11 submissions ranked in Table 6. This index indicates how diverse the errors of each model are with respect to each other if one would build an ensemble with them. The resulting CFD was 0.5590, indicating regular diversity (the range of values of CFD is [0,1]), that could be exploited for building a more robust model. On the other hand the maximum possible accuracy (a tweet is counted as well classified, if any of the models classified it correctly) was 0.9181. Further evidencing the potential benefits of building an ensemble with the 11 evaluate solutions.

In order to illustrate the inherent difficulties of the shared task, Table 8 shows examples of tweets that were missclassified by most participants when approaching subtask 1. Several interesting aspects can be discussed around these examples. First, there are tweets that were wrongly labeled by the annotators, for instance sample 3. This was a problem highlighted by participants during the challenge (Turón et al., 2022). Secondly, there are tweets for which the assigned category is debatable. For instance, tweet 4 refers to an accident happening in the context of an F1 race, it is an accident, but not really relevant for the purpose of the project. Also, tweet 1 refers to a report associated to several violent events happening in different places (we hypothesize this is why it was labeled as Non-violent). Summarizing, a large portion of samples missclassified by the systems could be due to subjective labeling. Therefore, we conclude the dataset needs of further manual curation. Still, we think the DA-VINCIS corpus is a valuable resource that will boost research in this relevant task.

5 Conclusions

The DA-VINCIS shared task at IberLEF promotes research into the identification of violent incidents on social networks, a task with a high social impact. A new dataset for the task of identification of violent incidents as well as their subcategorization is presented. This evaluation campaign made it possible to evaluate an important diversity of approaches and contrast their effectiveness. Different models, characteristics and techniques of the proposed approaches were presented, contributing to the progress of the identification of violent incidents in Spanish language.

The results indicate, as might be expected, that the fine-grained subtask 2 was more challenging. A strong presence of approaches based on transformers was found, but also there was a vitalizing variety of proposals with important novelties such as the application of GANs, the automatic correction of instances, and the use of non-learning tools to act as a kind of oracle, all of them introduced to improve the methods' performance as well as to to deal the specific challenges of the task at hand.

It was found that having some information on the subcategory of the general class of interest seems to help to make a better iden-

ID	Translation	Text	Category
1	Intense police activity in Coacoatzintla's municipality, in response to the supposed kidnap of a young male. A family member of the kidnapped person was killed when trying to impede this crime. In Jilotepec was found the vehicle where the person was abducted SP_Veracruz	*Una fuerte movilización policiaca se registró en el municipio de Coacoatzintla ante el presunto secuestro de un joven. Al tratar de impedir el hecho, un familiar fue asesinado. En Jilotepec fue hallado el vehículo en el que se cometió el ilícito SP_Veracruz*	Non-violent
2	30 years now from the Cimitarra massacre, a violence act that left more than 250 thousand deaths turning Colombia into a a huge common grave.	*A 30 años de la masacre de Cimitarra, una violencia que dejo más de 250 mil muertos convirtiendo a Colombia en una gran fosa común.*	Violent
3	Homicide - In a clinic at #Cartago Bibiana Liseth Guzmán Ordóñez, 31 years old and official of the @ipscartago, died, after she was shoot with a firearm. In the same incident a 26 years old man was hurt. The women left a daughter.	*Homicidio - En una clínica de #Cartago falleció Bibiana Liseth Guzmán Ordóñez de 31 años de edad, funcionaria de la @ipscartago luego de que le propinaran varios impactos con arma de fuego. En este mismo hecho resultó lesionado un hombre de 26 años. La mujer deja una hija.*	Non-violent
4	"The accident could have been avoided if they would leave me enough space to take the curve. You need of two persons for this to work, and I felt they throw me away. When we challenge to each other in a race this things can happen, unfortunately."	"El accidente se pudo haber evitado si me hubieran dejado espacio suficiente para tomar la curva. Necesitas 2 personas para que esto funcione y yo sentí que sacaban. Cuando nos retamos mutuamente en una carrera estas cosas pueden pasar, desafortunadamente."	Violent

Table 8: Examples of tweets incorrectly classified by all of the participant teams.

tification. Multitask Learning is strongly positioned as a good alternative that improves performance and takes advantage of the parallelism between subtasks 1 and 2. These findings open the possibility that other future approaches could use virtual subtasks with different fine grain levels in order to take advantage of this type of scheme.

Likewise, an in depth analysis of the corpus revealed that there is room for improvement in terms of the quality of annotations. On the one hand, a curation process trying to identify noisy annotations should be performed. On the other hand, the definition of categories should be further tuned, so that annotation guidelines result in objective annotations. This is work in progress.

As previously mentioned, the DA-VINCIS corpus also comprises visual information, therefore another venue of current work is studying the potential added value of using images associated to tweets when detecting violent incidents. The corpus will allow us to study the performance of solutions that consider multimodal information.

Acknowledgements

This work was supported by CONACyT under grant CB-S-26314, *Integración de Lenguaje y Visión mediante Representaciones Multimodales Aprendidas para Clasificación y Recuperación de Imágenes.* We also would like to thank CONACyT for partially supporting this work under grant CB-2015-01-257383. Additionally, the authors thank CONACYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies.

References

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

García-Díaz, J. A., S. M. Jiménez-Zafra, M. Rodríguez-García, and R. Valencia-García. 2022. UMUTeam at DA-VINCIS 2022: Aggressive and Violent classification using Knowledge Integration and Ensemble Learning. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

Mata Rivera, M., M. Torres-Ruiz, G. Guzmán, R. Quintero, R. Zagal-Flores, M. Moreno, and E. Loza. 2016. A Mobile Information System Based on Crowd-Sensed and Official Crime Data for Finding Safe Routes: A Case Study of Mexico City. *Mobile Information Systems*, 2016:1–11, 03.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic.

Montañés-Salas, R. M., R. del Hoyo-Alonso, and P. Peña-Larena. 2022. ITAINNOVA@DA-VINCIS: A Tale of Transformers and Simple Optimization Techniques. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

Pavao, A., I. Guyon, A.-C. Letournel, X. Baró, H. Escalante, S. Escalera, T. Thomas, and Z. Xu. 2022. CodaLab Competitions: An open source platform to organize scientific challenges. Technical report, Université Paris-Saclay, FRA., April.

Piña-García, C. and L. Ramírez-Ramírez. 2019. Exploring crime patterns in Mexico City. *Journal of Big Data*, 6, 07.

Qin, G., J. He, Q. Bai, N. Lin, J. Wang, K. Zhou, D. Zhou, and A. Yang. 2022. Prompt Based Framework for Violent Event Recognition in Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

Sandagiri, C., B. Kumara, and B. Kuhaneswaran. 2020a. Detecting Crime Related Twitter Posts using Artificial Neural Networks based Approach. pages 5–10, 11.

Sandagiri, S., B. Kumara, and B. Kuhaneswaran. 2020b. Deep Neural Network-Based Approach to Identify the Crime Related Twitter Posts. *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pages 1000–1004.

Ta, H. T., A. B. S. Rahman, L. Najjar, and A. Gelbukh. 2022a. GAN-BERT: Adversarial Learning for Detection of Aggressive and Violent Incidents from Social Media. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

Ta, H. T., A. B. S. Rahman, L. Najjar, and A. Gelbukh. 2022b. Multi-Task Learning for Detection of Aggressive and Violent Incidents from Social Media. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

Tonja, A. L., M. Arif, O. Kolesnikova, A. Gelbukh, and G. Sidorov. 2022. Detection of Aggressive and Violent Incidents from Social Media in Spanish using Pretrained Language Model. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

Turón, P., N. Perez, A. García-Pablos, E. Zotova, and M. Cuadros. 2022. Vicomtech at DA-VINCIS: Detection of Aggressive and Violent Incidents from Social Media in Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

Vallejo-Aldana, D., A. P. López-Monroy, and E. Villatoro-Tello. 2022. Leveraging Events Sub-Categories for Violent-Events Detection in Social Media. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR Workshop Proceedings. CEUR-WS.org.*

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 217-228 recibido 05-07-2022 revisado 28-07-2022 aceptado 31-07-2022

Overview of DETESTS at IberLEF 2022: DETEction and classification of racial STereotypes in Spanish

Resumen de la tarea de DETESTS en IberLEF 2022: DETEcción y clasificación de eSTereotipos raciales en eSpañol

Alejandro Ariza-Casabona[1,*], Wolfgang S. Schmeisser-Nieto[1,*], Montserrat Nofre[1], Mariona Taulé[1], Enrique Amigó[2], Berta Chulvi[3,4], Paolo Rosso[3]
[1]CLiC, UBICS, Universitat de Barcelona, Spain
[2]Research Group in NLP and IR, Universidad Nacional de Educación a Distancia, Spain
[3]PRHLT Research Center, Universitat Politècnica de València, Spain
[4]Universitat de València, Spain
{alejandro.ariza14, wolfgang.schmeisser, montsenofre, mtaule}@ub.edu,
enrique@lsi.uned.es, berta.chulvi@upv.es, prosso@dsic.upv.es

Abstract: This paper presents an overview of the DETESTS shared task as part of the IberLEF 2022 Workshop on Iberian Languages Evaluation Forum, within the framework of the SEPLN 2022 conference. We proposed two hierarchical subtasks: For subtask 1, participants had to determine the presence of stereotypes in sentences. For subtask 2, participants had to classify the sentences labeled with stereotypes into ten categories. The DETESTS dataset contains 5,629 sentences in comments in response to newspaper articles related to immigration in Spanish. 51 teams signed up to participate, of which 39 sent runs, and 5 of them sent their working notes. In this paper, we provide information about the training and test datasets, the systems used by the participants, the evaluation metrics of the systems and their results.
Keywords: Stereotype detection and classification, DETESTS dataset, evaluation metrics.

Resumen: Este artículo presenta un resumen de la tarea DETESTS como parte del workshop IberLEF 2022, dentro de la conferencia SEPLN 2022. Proponemos dos subtareas jerárquicas: En la subtarea 1, los participantes tuvieron que determinar la presencia de estereotipos raciales en oraciones. En la subtarea 2, de las oraciones etiquetadas con estereotipo, los participantes tuvieron que clasificarlas en una o más de diez categorías. El dataset DETESTS contiene 5.629 oraciones de comentarios que responden a artículos de periódicos sobre inmigración en español. 51 equipos se registraron para participar, de los cuales 39 enviaron predicciones de sistemas y 5 de ellos enviaron artículos. En este artículo presentamos información sobre los datasets de entrenamiento y de prueba, los sistemas utilizados por los participantes, las métricas de evaluación y sus resultados.
Palabras clave: Detección y clasificación de estereotipos, dataset DETESTS, métricas de evaluación.

1 Introduction

The DETESTS (DETEction and classification of racial STereotypes in Spanish) task, held at IberLEF 2022, focuses on the detection and classification of stereotypes related to immigration in sentences taken from comments posted in Spanish in response to different online news articles. The present task is proposed to participants interested in racial, national, or ethnic stereotype detection and classification tasks, which is a relevant and relatively novel area of research due to its impact on modern society. Furthermore, the annotated dataset is a valuable resource for exploratory linguistic analysis, as well as for comparing the application of deep learning and classical machine learning models to Spanish stereotyped expres-

[*] These authors contributed equally to this work.

ISSN 1135-5948. DOI 10.26342/2022-69-19

sions under the recently introduced learning with disagreements paradigm (Basile et al., 2021; Uma et al., 2021).

The following sections of this paper describe the key aspects of this task. Section 2 offers a background on what is understood as stereotypes and the related work on Natural Language Processing (NLP). Section 3 presents both proposed subtasks. Section 4 describes the DETESTS corpus, its training and test datasets and the annotation process. Section 5 presents the systems used by the participants, the evaluation metrics and the results. Finally, Section 6 corresponds to conclusions and draws some lines for future work.

2 Background

One of the components that reinforces toxic and hateful speech is stereotypes. Understanding how they emerge and spread is crucial to tackling this issue, since stereotypes are not always expressed explicitly. The presence of stereotypes on social media and the need to identify and mitigate them is driving the development of systems for their automatic detection, especially in news comments. Therefore, this is a new task that is attracting growing interest from the NLP community.

A stereotype is defined in social psychology as a set of beliefs about others who are perceived as belonging to a different social category. The stereotype oversimplifies the group and generalizes a characteristic, applying it to all its members (Allport, Clark, and Pettigrew, 1954). The common assumption in social psychology literature is that some of the behavior toward others is driven by stereotypes (cognitive component) and prejudices (emotional component). One way of manifesting stereotypes is through language in different degrees ranging from explicit to implicit, thereby becoming a complex concept when they must be operationalized for natural language processing. In order to narrow down this concept, we considered some criteria for deciding whether a message contains a stereotype. Since not every linguistic expression about immigration carries a racial, national or ethnic stereotype, the first criterion to observe is whether there is a homogenization of the target group in the comment. Homogenization involves a process of the generalization of a feature to the status of a social category, which negates individual diversity (Tajfel, Sheikh, and Gardner, 1964; Tajfel, 1984). In a second criterion, stereotypes are expressed in language through several communication acts, which can be explicit, that is, transparent and manifest, or implicit, which means that a process of inference is necessary for the stereotype to be perceived (Schmeisser-Nieto, Nofre, and Taulé, 2022).

Several works on stereotype detection and classification have been carried out, in which specific social groups, e.g., women and immigrants, have been the focus of research, since they are usually the target of such messages. For instance, Automatic Misogyny Identification (Fersini, Rosso, and Anzovino, 2018) presents a classification subtask in which one of the categories of misogyny is Stereotype and Objectification understood as a fixed and oversimplified image or idea of a woman. Last year's IberLEF 2021 edition task EXIST (Rodríguez-Sánchez et al., 2021) tackled the topic of sexism in social networks. Moreover, studies on the detection of gender stereotypes have also been addressed in (Cryan et al., 2020; Chiril, Benamara, and Moriceau, 2021). Among the perspectives on identifying stereotypes within narratives, there are studies of microportraits in Muslim stereotyping in which a description of the target group is provided in a single text (Fokkens et al., 2019). Sap et al. (2020) approach the problem of stereotypes for several target groups in the Social Bias Frame, a new conceptual formalism that aims to model the pragmatic frames in which people project social bias and stereotypes onto others. Evalita 2020's HaSpeeDe 2 task includes a subtask on the identification of immigrants, Muslims and Roma (Sanguinetti et al., 2020). Narrowing down on the topic of immigration, Sánchez-Junquera et al. (2021) put forward a classification of such stereotypes as manifested in political debates. The stereotype classification applied in this task is based on the latter work but uses a corpus extracted from comments authored by web users on Spanish news articles related to immigration. In general, in these comments, a racial stereotype based on origin, ethnicity, race and religion is associated with a target group.

3 Task Description

The aim of the DETESTS task is to detect and classify stereotypes in sentences from

comments posted in Spanish in response to different online news articles related to immigration. A sentence can contain one or more stereotypes belonging to different categories and, therefore, it may have multiple labels that need to be accurately detected. This scenario is known in the literature as a multi-label classification problem. However, to adapt the problem to a variety of participants' interests, the task is designed in a hierarchical fashion by chaining two subtasks and allowing participants to either model the simple binary scenario or complete the entire pipeline by modeling the complex multi-label classification problem.

Subtask 1: Detection of Stereotypes Participants that tackled this problem had to determine whether the sentences in a comment contain at least one stereotype (positive example) or none (negative example) considering the full distribution of labels provided by the annotators. The gold standard of this subtask is left as a proxy to determine the subset of sentences that will be evaluated in the posterior subtask. For this subtask, we also invited participants to consider a learning with disagreements approach, proposed in SemEval 2021 Task 12 (Uma et al., 2021), in which the authors state that there does not necessarily exist a single gold standard for every sample in the dataset.

Subtask 2: Classification of stereotypes This subtask consists of determining whether a sentence contains at least one stereotype or none and assigning those sentences previously marked as positive (with stereotypes) to at least one of the ten categories that present immigrants as: 1) 'victims of xenophobia', 2) 'suffering victims', 3) 'economic resources', 4) a problem of 'migration control', 5) people with 'cultural and religious differences', 6) people that take advantage of welfare 'benefits', 7) a problem for 'public health', 8) a threat to 'security', 9) 'dehumanization' and 10) 'other' types of stereotypes. Since a sentence can contain multiple stereotypes belonging to different categories, this subtask is presented as a multi-label hierarchical classification problem.

Teams were allowed (and encouraged) to submit multiple runs (max. 5). Subtask 2 was optional.

4 Dataset

The DETESTS dataset consists of 5,629 sentences, with an average of 24% of them containing stereotypes. It is made up of two parts -one from the NewsCom-TOX corpus (Taulé et al., 2021) (3,306 sentences) and the other from the StereoCom corpus (2,323 sentences), which was created especially for this task. Both corpora consist of comments published in response to different articles extracted from Spanish online newspapers (ABC, elDiario.es, El Mundo, NIUS, etc.) and discussion forums (such as Menéame[1]). In the case of NewsCom-TOX, the dates of the articles range from August 2017 to August 2020, while in StereoCom they range from June 2020 to November 2021.

To collect the NewsCom-TOX corpus, a keyword-based approach was used to search for articles related mainly to racism and xenophobia. Then, the articles were manually selected based on their controversial subject matter, potential toxicity and the number of published comments (minimum 50 comments per article). Since the NewsCom-TOX corpus was designed primarily to study toxicity and not stereotypes, we used only the part of the corpus with the highest percentage of stereotypes, which had been annotated previously. In order to obtain a sufficient and balanced data volume in terms of the presence or absence of stereotypes, the same content was also collected for the StereoCom corpus, i.e., comments in response to immigration-related news items in Spanish digital media, selected by subject matter on the basis of a keyword search.

The comments were presented in the same order in which they appeared in the temporal web thread, along with the conversational thread. Each comment was segmented into sentences, and the comment to which every sentence belongs and its position within the comment are indicated.

The default dataset includes the gold standard annotation. If the participants wish to apply methods of learning with disagreements, we will provide, upon request, the pre-aggregated annotation, that is, the annotation of each annotator.

[1]https://www.meneame.net

4.1 Annotation Scheme

To accomplish the classification tasks, we annotated the dataset with the main labels to indicate the presence or absence of stereotypes and the category/ies of the stereotype to which they belong. Moreover, we annotated extra features that could help the participants to train their systems. Since more than one stereotype corresponding to different categories can appear in one sentence, this is a multi-label task. We based our stereotype categories on the work proposed by Sánchez-Junquera et al. (2021). All the labels are annotated with binary values (0=absence of the feature and 1=presence of the feature).

4.1.1 Main labels

For each sentence, annotators had to decide whether there was at least one stereotype related to a target group.

Stereotype: There is a process of homogenization of one characteristic of an individual or part of a group that is applied to the entire group based on their place of origin, ethnicity or religion. Stereotypes can be expressed explicitly or implicitly.

All sentences annotated with stereotypes are also annotated with at least one of the categories listed below (see examples on the task's website[2]):

Xenophobia Victims: The members of the target group are perceived as victims of xenophobia and discrimination.

Suffering Victims: The members of the target group are portrayed as victims of poverty and violence in their places of origin, and as having to face difficult situations in their host countries.

Economic Resource: The members of the target group are seen as an economic resource. They do the jobs that locals do not want to do, pay taxes, and solve the problems arising from low population growth.

Migration Control: Immigration presents a threat due to massive influxes and a lack of control at the borders. Immigrants are illegal and they should be expelled. It is seen as an invasion.

Cultural and Religious Differences: The major threat consists of the loss of the ingroup's values and traditions, and the replacement of the target group's customs and religions. Immigrants are also seen as uneducated and should adapt to their host country.

Benefits: The target group competes with the ingroup for resources such as public subsidies, school places, jobs, health care and pensions. There is a perception of the target group being priveliged over the ingroup.

Public Health: Immigrants are thought to be carriers of infections and diseases such as COVID-19, Ebola and HIV.

Security: Immigration brings security issues. Due to immigration, there is an increase in crime, domestic violence, robbery, drug use, sexual assault, murder, terrorist attacks and public disorders.

Dehumanization: The members of the target group are seen as inferior beings and are compared with animals, parasites or scum. Their lives have less value than those of the ingroup.

Others: Any other racial stereotype that is not covered in the previous categories.

4.1.2 Additional labels

The DETESTS dataset has also been annotated with three other labels that may provide extra features at the disposal of the participants to use optionally to train their systems. These additional labels are:

Racial target: The target group is defined by place of origin, ethnicity or religion.

Other target: The target group corresponds to other minorities or oppressed groups based on gender, sexual orientation, physical or mental health conditions or age, among others.

Implicitness: This category refers to whether the stereotype in the sentence is expressed implicitly or explicitly.

4.2 Annotation Process

Once we had defined what we understand by stereotypes, which categories we can observe in our data, and in which ways they can be manifested in texts, we drew up annotation guidelines for the annotators.

The annotation process consisted of two stages. In the first stage, the annotation of the categories 'stereotype', 'racial_target', 'other_target' and 'implicitness' was carried out. The second stage consisted of the

[2]https://detestsiberlef.wixsite.com/detests/tasks

annotation of the categories of the stereotypes. Then, disagreements were discussed by the annotators and a senior annotator until agreement was reached. The team of annotators involved in the task consisted of two expert linguists and two trained annotators who are students of linguistics.

Each sentence was annotated in parallel by three annotators and an inter-annotator agreement (IAA) test was performed once all the sentences had been annotated. As shown in Table 1, overall, the IAA test gave high results, excluding the feature 'other_target', which had a Fleiss' Kappa of 0.139. This may be due to the scarcity of data corresponding to that feature, since the average pairwise % agreement is still one of the highest. A similar case, although with higher results, can be observed for the category 'others'. It is worth noticing as well that the categories of stereotypes with less IAA correlate with the categories with the highest distribution among the sentences (see Table 2). These categories are 'migration control', 'security', 'benefits' and 'culture'. Moreover, these categories also co-occur together with a higher frequency than other categories (see Figure 5 in Appendix A).

Label	Av. pairwise % Agreement.	Fleiss' Kappa
stereotype	84.36%	0.573
xenophobia	97.65%	0.348
suffering	94.55%	0.523
economic	96.86%	0.593
migration control	83.86%	0.669
culture	90.68%	0.65
benefits	91.61%	0.764
health	98.92%	0.744
security	89.85%	0.735
dehumanization	93.43%	0.488
others	92.74%	0.372
racial_target	84.05%	0.619
other_target	98.61%	0.139
implicitness	81.66%	0.412

Table 1: Inter-annotator agreement test.

4.3 Training and Test Datasets

Participants were provided with 70% of the corpus to train and validate their models on (3,817 comments) and the remaining 30% of the corpus (1,812 comments) was used as a test set to evaluate their performance against unseen sentences[3]. In order to avoid data leakage from the NewsCom-TOX corpus released in the DETOXIS shared task, all test sentences were extracted from the newly added StereoCom corpus in a stratified manner to keep a similar label distribution to the one found in the training set. Note that, despite the fact that the training dataset contains all gold standard categories (see Section 3) together with three additional features – 'racial_target', 'other_target', 'implicitness' – none of this information is provided in the test set, which merely includes comment and sentence identifiers of each instance – the identifier of the sentence it replies to (if any), and the sentence text.

Category	Comments	Percentage
xenophobia	21	1.55%
suffering	113	8.31%
economic	62	4.56%
migration	553	40.69%
culture	265	19.50%
benefits	315	23.18%
health	37	2.72%
security	376	27.67%
dehumanization	100	7.36%
others	90	6.62%

Table 2: Category distribution of sentences that contain at least one type of stereotype.

Table 2 shows the category distribution for the subset of examples that are annotated as containing at least one stereotype. This subset contains 1,359 sentences (out of 5,629), that is, 24.14% of the whole corpus. It is important to mention that, given the multi-label nature of the task, some sentences may contain stereotypes belonging to multiple categories and the amount of overlapping among categories can be noticed in the histogram provided in Figure 1.

5 Systems and Results

This section contains a brief description of the proposed baselines, as well as an overview of the systems submitted by the participants, a brief comparison of such models regarding the selected evaluation metrics for each sub-

[3]To avoid any conflict with the sources of the comments regarding their intellectual property rights (IPR), a password to access the data was sent privately to each participant who was interested in the task after filling in a registration form. This dataset will only be made available for research purposes.

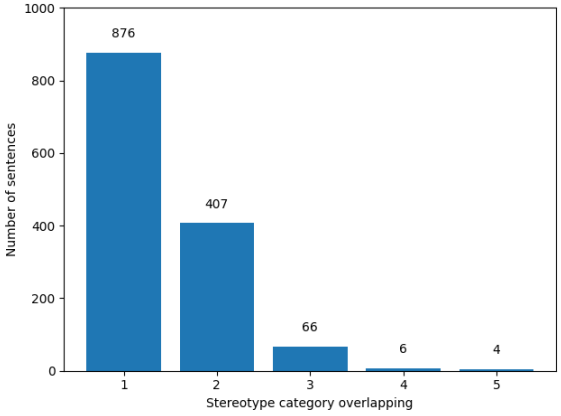


Figure 1: Multi-label distribution.

task, and a short analysis of their multi-label capabilities. A Github repository is publicly available with the implementation of the official metrics, the baselines, the systems evaluation, and an overview analysis[4].

5.1 Baselines

In order to analyze certain performance boundaries in both subtasks, five different baselines have been considered as reference models to be compared with the participant's systems: AllOnes, AllZeros, RandomClassifier, TFIDF+SVC and FastText+SVC. Due to the fact that the second subtask consists of a hierarchical multi-label classification task, we have extended these baselines in a hierarchical fashion by first determining whether the sentence contains at least one stereotype and a set of new baseline classifiers is trained upon those positive cases to predict each of the stereotype categories (to tackle the multi-label classification problem).

Each baseline is briefly introduced below:

AllOnes: This baseline maps all instances to the positive class it is trying to classify.

AllZeros: Analogously to AllOnes, this baseline maps all instances to the negative class. Therefore, this baseline is only considered in subtask 2 in which the negative class is actually accounted for by the evaluation metrics.

RandomClassifier: A weighted random classifier picks a random class with probabilities based on the label distribution learnt from the training set.

TFIDF+SVC: A TF-IDF vectorizer is used to extract sentence-level features based on the learnt 10,000, unicode, lowercased vocabulary of n-grams with sizes 1 to 3. The classifier selected to classify instances based on the extracted features is a Support Vector Classifier (SVC) with a linear kernel.

FastText+SVC: This baseline replaces the classical TF-IDF vectorizer with a word vector extractor based on the FastText algorithm followed by a mean pooling operation for sentence-level representation. A SVC classifier with a linear kernel is also used as a component of this baseline.

All baselines have been implemented using Python language, together with the following libraries: Numpy[5], Pandas[6], Scikit-learn[7], and SpaCy[8].

5.2 Systems Overview

The DETESTS shared task received submissions from 39 teams for subtask 1, although only five of these teams decided to tackle subtask 2 as well. Participants were allowed to provide up to five submissions per subtask. Among the top-performing systems, we observe an extended use of pre-trained language models for the Spanish language including both BERT and RoBERTa. The main differences that lead to the leaderboard ranking presented in Tables 3 and 4 depended on how they approached problems such as data unbalance, the multi-label problem or contextual information (for the ranking including the total of participants, visit task's website[9]). Despite their lower performance, more classical machine learning and NLP techniques were considered either as baselines or submission systems by multiple participants. These participants provided ensemble architectures and bagging strategies with Bag-of-Words representations and models such as SVC, Random Forest Classifier and/or Logistic Regression. It is worth noting that both DETESTS subtasks are really challenging, especially for those classical machine learning models whose representational capabilities depend mainly on the quality of the input features. Another main problem that participants had to face in this competition was the fact that the variety of pre-

[4] https://github.com/alarca94/detests

[5] https://numpy.org/doc/stable/index.html

[6] https://pandas.pydata.org/

[7] https://scikit-learn.org/stable/

[8] https://spacy.io/

[9] https://detestsiberlef.wixsite.com/detests/evaluation-results

trained and fine-tuned language models for Spanish, although continuously increasing, is still very limited. The most interesting approaches in the competition are summed up below.

First, the top scoring team **I2C_III** (Vázquez et al., 2022) opted for two merging multiple strategies that tackled the problems of unbalanced data and semantic textual representation. On the one hand, they tried to balance the dataset with both undersampling and Bagging of the majority class, and oversampling of the minority class with a double translation from Spanish to English and back. Moreover, I2C_III implemented an ensemble architecture combining not only balancing techniques but different pretrained language models to increase the semantic representation capabilities of the system.

Second, **UMUTeam** (García-Díaz, Jiménez-Zafra, and Valencia-García, 2022) made use of their own UMUTextStats tool to extract a set of 389 linguistic feature sets that were combined together with some negation features, non-contextual word vector representations (FastText) and contextual pre-trained language modelling using both BETO and RoBERTa. In the end, their model combined these representations via either knowledge integration or ensemble learning, thereby proving the importance of good feature selection. It is important to note that negation features only boosted their model for subtask 2, which may indicate a bigger impact on the discriminative power of the models for stereotype category classification, as opposed to their influence on simpler stereotype binary detection.

An important point regarding the submitted models is that none of them tries to enrich the contextual information by extracting representations from other sentences in the same comment. However, the **Lak_NLP** team (Laknani and García-Martinez, 2022) benefits from the additional features ('implicitness' and 'racial_target') included in the training set that participants were provided with. Given the fact that these features were not part of the test dataset, Lak_NLP develop a meta-classifier to learn this additional feature distribution and included its prediction as auxiliary input to the pre-trained BETO model leading to an overall good performance in both subtasks.

Furthermore, the **DaMinCi** team (Cabestany, Adsuar, and López, 2022) tried to distinguish itself from the rest of the participants by incorporating Adapters to the fine-tuning strategy of the pre-trained language models. This adapter-based model consists of incorporating bottleneck layers between the existing hidden layers of the selected model (RoBERTa in their case) and freezing pre-existing model weights during fine-tuning. According to their own validation and their final score on subtask 1, this approach outperforms other interesting alternatives such as a fine-tuned RoBERTa model on auxiliary tasks that leverage knowledge learnt from related domains.

Last but not least, the **MALNIS** team (Ramirez-Orta1 et al., 2022) approached the DETESTS shared task as a Multi-Task Learning problem in which a final classification head per stereotype category is stacked on top of a pre-trained RoBERTa model and fine-tuned using a point-wise Cross-Entropy loss function. Their system showed the importance of jointly modelling the distribution of all stereotype categories in the overall model performance for both subtasks by ranking first in subtask 2. Although not all participants mentioned their preprocessing strategies in their respective working notes, pre-processing may play an important role in the behavior of the models, especially if we are considering classical machine learning models built from scratch. Some of the steps that have been implemented by several participants range from common tokenization, stopwords removal, lowercasing, numbers removal, URL and user tags masking, as well as spell correction.

5.3 Metrics

Subtasks 1 and 2 have been evaluated with different metrics. Subtask 1 is a binary classification problem and the F-measure combining Precision and Recall on the positive class (stereotype) was applied. In addition, subtask 2 was interpreted as a two-level multiclass hierarchical classification problem. The first level corresponds to the binary classification of the previous task (stereotype or non-stereotype). On a second level, the positive class is decomposed into the ten subcategories described in Section 4.1. The multiclass classification metrics can be label or instance-based. Label-based metrics evaluate systems independently for each class. We

have discarded this type of metrics as they do not consider the specificity and relative weight of the classes. In contrast, instance-based metrics evaluate label sets item by item. Within this family we have considered the following three metrics. The first is label propensity applied over precision and recall for single items. Each accurate class in the intersection is weighted according to the class *propensity* p_c (Jain, Prabhu, and Varma, 2016). In particular, we have considered the variant proposed by Amigó and Delgado (2022), with $s(i)$ and $g(i)$ being the set of classes assigned to item i in the system output and gold standard respectively.

$$\text{Prop}_P(i) = \frac{\sum_{c \in s'(i) \cap g'(i)} \frac{1}{p_c}}{\sum_{c \in s'(i)} \frac{1}{p_c}}$$

$$\text{Prop}_R(i) = \frac{\sum_{c \in s'(i) \cap g'(i)} \frac{1}{p_c}}{\sum_{c \in g'(i)} \frac{1}{p_c}}$$

where $s'(i) = s(i) \cup \{c_\emptyset\}$ and $g'(i) = g(i) \cup \{c_\emptyset\}$. The reason for adding the empty class $c_\emptyset$ is to capture the specificity of classes in mono-label items. The propensity factor p_c for each class is computed as: $p_c = \frac{1}{1+Ce^{-A \log_2(N_c+B)}}$ where N_c is the number of data points annotated with label c in the observed ground truth data set of size N and A, B are application specific parameters and $C = (logN - 1)(B + 1)^A$. In this evaluation campaign, we set the recommended parameter values $A = 0.55$ and $B = 1.5$. Propensity F-measure (PROP-F) is computed as the harmonic mean of these values.

The previous metric captures the specificity of classes appropriate in unbalanced data sets. However, it does not capture hierarchical relationships. For this, we also applied hierarchical-based metrics that consider the ancestor overlap (Kiritchenko, Matwin, and Famili, 2004; Costa et al., 2007). More concretely, hierarchical precision and recall are computed as the intersection of ancestor divided by the amount of ancestors of the system output category and of the gold standard respectively. In our evaluation, when computing the ancestor overlap we consider the common empty label (root class) in order to avoid undefined situations. Their combination is the Hierarchical F-measure (HF). Since these metrics are based on category set overlap, they can be applied as example based multi-label classification by joining ancestors and computing the F measure. Their drawback is that the specificity of categories is not strictly captured since they assume a correspondence between specificity and hierarchical deepness. However, this correspondence is not necessarily true. Categories in first levels can be infrequent whereas leaf categories can be very common in the data set.

In order to capture both aspects simultaneously, the official metric in this campaign is the *Information Contrast Model* (ICM) (Amigó and Delgado, 2022), which is a similarity measure that unifies measures based on both object feature sets and Information Theory (Amigó et al., 2020). Given two class sets $s(i)$ and $g(i)$, ICM is computed as:

$$\text{ICM}(A, B) = \alpha_1 I(s(i)) + \alpha_2 I(g(i)) - \beta I(s(i) \cup g(i))$$

where $I(X)$ represents the information content ($-log(P(X))$ of the class set X. The intuition is that the more unlikely the category sets are to occur simultaneously (large $I(s(i) \cup g(i))$), the less they are similar. Given a fixed joint IC, the more the category sets are specific ($I(s(i))$ and $I(g(i))$), the more they are similar. ICM is grounded on similarity axioms supported by the literature in both information access and cognitive sciences (Amigó et al., 2020). According to Amigó and Delgado (2022), the information content of a class set can be computed as:

$$I(\{c_1, c_2, .., c_n\}) =$$
$$I(c_1) + I\left(\bigcup_{i=2..n} \{c_i\}\right) - I\left(\bigcup_{i=2..n} \{\text{lso}(c_1, c_i)\}\right)$$

where $\text{lso}(c_i, c_j)$ represents the common ancestor of the classes c_i and c_j.

5.4 Subtask 1

Table 3 shows the ranking of participating systems for subtask 1 according to the F-measure on the positive class. The table includes the best run per team that sent working notes. All the systems show better results than the baselines. The random classifier is the worst baseline and labeling all items as positive achieves an F-score of 0.42.

Figure 2 plots the precision and recall scores for every run. As the figure shows, some systems manage to distinguish themselves from the rest in both precision and recall by following the diagonal in the di-

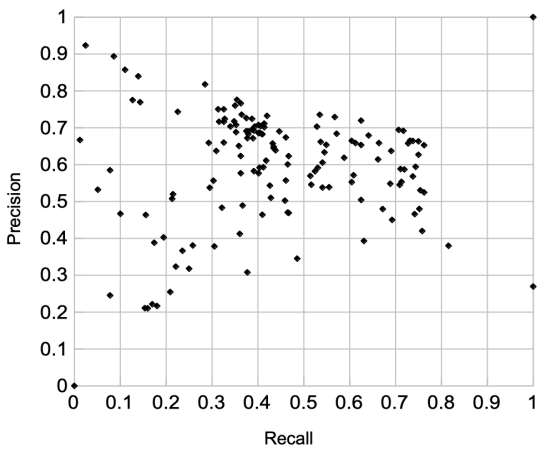


Figure 2: Precision vs. Recall in subtask 1.

Ranking	Team Name	F-Score
	Gold Standard	1.0000
1	I2C_III	0.7042
3	UMUTeam	0.6990
5	Lak_NLP	0.6627
6	DaMinCi	0.6596
9	MALNIS	0.6382
	FastText+SVC	0.4861
	TFIDF+SVC	0.4706
	AllOnes	0.4243
	RandomClassifier	0.2295

Table 3: Evaluation results in subtask 1.

rection of the (1,1) point of the gold standard. This distribution indicates that the standard F-measure weighting (precision and recall equally weighted) is appropriate for establishing the official ranking.

5.5 Subtask 2

Table 4 shows the ranking of systems according to the metrics ICM, hierarchical F-measure (HF) and Propensity F (Prop-F). Again, the baseline systems (AllZeros, RandomClassifier and AllOnes) obtain lower results than those obtained by the participating systems. In particular, assigning all possible labels to all items (AllOnes) is penalized by all metrics and especially by ICM since the system introduces a lot of missing information in relation to very specific classes. All three metrics agree that not assigning any class (AllZeros) is a better option than any other arbitrary baseline.

Figure 3 shows the relationship between HF and Prop-F scores. As the figure shows, these metrics correlate in this benchmark.

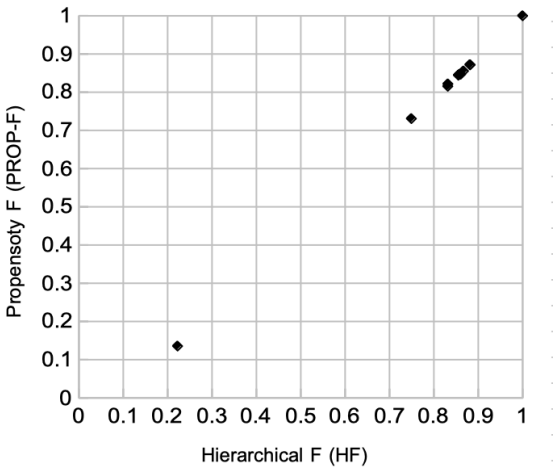


Figure 3: Hierarchical F-measure vs. Propensity F-measure in subtask 2.

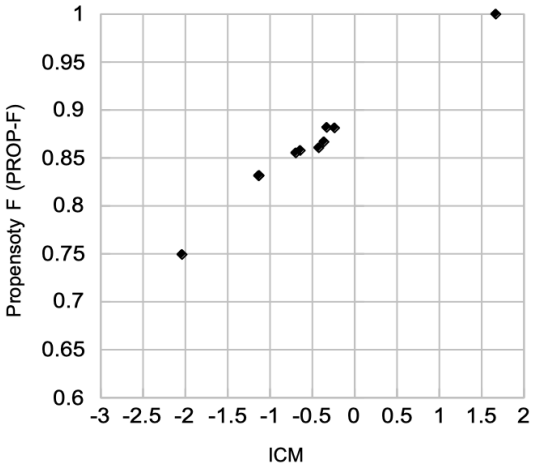


Figure 4: ICM vs. Propensity F-measure in subtask 2.

This suggests that both the hierarchical distance captured by HF and the class specificity captured by Prop-F are not deterministic aspects in this task. This is because the hierarchical structure is quite simple and the classes are relatively balanced in the data set.

However, as Figure 4 shows, there is a slight mismatch between ICM and the other two metrics. This is because both HF and PROP-F compare, for each item, the set of labels assigned by the system and the set of classes to which it belongs through the F-measure on Precision and Recall. Note that Precision and Coverage are ratio-based similarity criteria between intersection and one of the sets (system output in the case of Precision and gold standard in the case of

Ranking	Team Name	ICM	HF	Prop-F
	Gold Standard	1.6676	1.0000	1.0000
1	MALNIS	-0.2380	0.8813	0.8717
2	UMUTeam	-0.3298	0.8818	0.8718
4	Lak_NLP	-0.4242	0.8606	0.8470
	TFIDF+SVC	-0.6954	0.8552	0.8442
	AllZeros	-1.1280	0.8317	0.8215
	FastText+SVC	-1.1348	0.8314	0.8154
	RandomClassifier	-2.0403	0.7493	0.7308
	AllOnes	-36.3162	0.2224	0.1354

Table 4: Evaluation results in subtask 2.

Recall). In contrast, the similarity scheme used in ICM considers the individual sets and their union. In other words, for evaluation purposes, our results suggest that the multi-labeling and the way in which the label sets are compared has more effect than the hierarchical structure or the class balance.

6 Conclusions and Future Work

This paper has described the DETESTS challenge at IberLEF 2022 and summarized the participation of several teams in both subtasks, emphasizing the relevant differences that led to the final ranking. It is clear how important pre-trained language models are for complex natural language tasks such as stereotype classification and the fact that new model checkpoints for the Spanish language are increasingly being shared, allowing participants to achieve better results and come up with innovative solutions that couple well with state-of-the-art systems. Regarding the actual task, it has been designed as a hierarchical task that aims for stereotype detection and classification in Spanish sentences. Each sentence can contain up to ten different stereotype categories and three additional features are included to aid in the pattern representation of the models. Also, our dataset (by explicit request) also incorporates the labels of all annotators prior to their aggregation in case participants want to apply methods of learning with disagreements.

The winners of both subtasks tackled the major problems directly. On the one hand, for this first subtask, I2C_III noticed the negative effect of the unbalanced data and incorporated UnderBagging and Oversampling strategies to overcome it while employing powerful language models in an ensemble architecture. On the other hand, for the second subtask, MALNIS modeled the joint category distribution with a Multi-Task Learning strategy giving their system an important boost in terms of ICM, HF and Prop-F.

Unfortunately, the effect of data balancing was not explored for subtask 2 and, thus, remains open for future work. Other future research directions worth following that did not appear in any participant's model includes methods of learning with disagreements, adding more contextual information to the current sentences such as comment-level representation or topic modelling, among others. Finally, despite the fact that the DaMinCi team tried to use fine-tuned models on related tasks, it would be interesting to verify domain commonalities and try to transfer complementary information to these pre-trained architectures more efficiently.

Acknowledgements

This work is supported by the following projects: 'STERHEOTYPES: STudying European Racial Hoaxes and sterEOTYPES' funded by Fondazione Compagnia di San Paolo and grant 'XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics' (PLEC2021-007681) funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by the "European Union NextGenerationEU/PRTR". The work of Paolo Rosso was carried out within the framework of the research project PROMETEO/2019/121 (DeepPattern) by the Generalitat Valenciana.

References

Allport, G. W., K. Clark, and T. Pettigrew. 1954. *The nature of prejudice*. Addison-wesley Reading, MA.

Amigó, E. and A. D. Delgado. 2022. Evaluating extreme hierarchical multi-label classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5809–5819.

Amigó, E., F. Giner, J. Gonzalo, and F. Verdejo. 2020. On the foundations of similarity in information access. *Inf. Retr. J.*, 23(3):216–254.

Basile, V., M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, and A. Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online, August. Association for Computational Linguistics.

Cabestany, D., C. Adsuar, and M. López. 2022. DaMinCi at IberLEF-2022 DETESTS task: Detection and Classification of Racial Stereotypes in Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*. CEUR Workshop Proceedings, CEUR-WS.org.

Chiril, P., F. Benamara, and V. Moriceau. 2021. "Be Nice to your wife! The Restaurants are Closed": Can Gender Stereotype Detection Improve Sexism Classification? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2833–2844, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Costa, E. P., A. C. Lorena, A. C. Carvalho, and A. A. Freitas. 2007. A review of performance evaluation measures for hierarchical classifiers. *AAAI Workshop - Technical Report*, 01.

Cryan, J., S. Tang, X. Zhang, M. Metzger, H. Zheng, and B. Y. Zhao, 2020. *Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods*, page 1–11. Association for Computing Machinery, New York, NY, USA.

Fersini, E., P. Rosso, and M. E. Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@SEPLN*.

Fokkens, A., N. Ruigrok, C. Beukeboom, S. Gagestein, and W. Van Atteveldt. 2019. Studying muslim stereotyping through microportrait extraction. In H. Isahara, B. Maegaard, S. Piperidis, C. Cieri, T. Declerck, K. Hasida, H. Mazo, K. Choukri, S. Goggi, J. Mariani, A. Moreno, N. Calzolari, J. Odijk, and T. Tokunaga, editors, *Proceedings of the LREC 2018, Eleventh International Conference on Language Resources and Evaluation*, pages 3734–3741. European Language Resources Association (ELRA). Conference date: 07-05-2018 Through 12-05-2018.

García-Díaz, J. A., S. M. Jiménez-Zafra, and R. Valencia-García. 2022. UMUTeam at IberLEF-2022 DETESTS task: Feature Engineering for the Identification and Categorization of Racial Stereotypes in Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*. CEUR Workshop Proceedings, CEUR-WS.org.

Jain, H., Y. Prabhu, and M. Varma. 2016. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 935–944, New York, NY, USA. Association for Computing Machinery.

Kiritchenko, S., S. Matwin, and F. Famili. 2004. Hierarchical text categorization as a tool of associating genes with gene ontology codes. *Proceedings of the 2nd European Workshop on Data Mining and Text Mining in Bioinformatics*, 01.

Laknani, F. and M. García-Martinez. 2022. Lak_NLP at IberLEF-2022 DETESTS task: Automatic Classification of Stereotypes in Text. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*. CEUR Workshop Proceedings, CEUR-WS.org.

Ramirez-Orta1, J., M. V. Sabando, M. Maisonnave1, and E. Milios. 2022. MALNIS at IberLEF-2022 DETESTS Task: A Multi-Task Learning Approach for Low-Resource Detection of Racial Stereotypes in Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*. CEUR Workshop Proceedings, CEUR-WS.org.

Rodríguez-Sánchez, F., J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67(0):195–207.

Sanguinetti, M., G. Comandini, E. di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, and I. Russo. 2020. Haspeede 2 @ evalita2020: Overview of the evalita 2020 hate speech detection task. In V. Basile, D. Croce, M. Di Maro, and L. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, volume 2765. CEUR Workshop Proceedings (CEUR-WS.org). Conference date: 17-12-2020.

Sap, M., S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi. 2020. Social bias frames: Reasoning about Social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July. Association for Computational Linguistics.

Schmeisser-Nieto, W., M. Nofre, and M. Taulé. 2022. Criteria for the annotation of implicit stereotypes. In *Proceedings of the Language Resources and Evaluation Conference*, pages 753–762, Marseille, France, June. European Language Resources Association.

Sánchez-Junquera, J., B. Chulvi, P. Rosso, and S. P. Ponzetto. 2021. How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*, 11(8).

Tajfel, H. 1984. *Grupos humanos y categorias sociales*. Herder.

Tajfel, H., A. A. Sheikh, and R. C. Gardner. 1964. Content of stereotypes and the inference of similarity between members of stereotyped groups. *Acta Psychologica,*, 22(3):191–201.

Taulé, M., A. Ariza, M. Nofre, E. Amigó, and P. Rosso. 2021. Overview of DETOXIS at IberLEF 2021: DEtection of TOXicity in comments In Spanish. *Procesamiento del Lenguaje Natural*, 67(0):209–221.

Uma, A., T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, and M. Poesio. 2021. SemEval-2021 Task 12: Learning with Disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online, August. Association for Computational Linguistics.

Vázquez, J. M., V. P. Álvarez, C. T. Taybi, and P. P. Sánchez. 2022. I2C at IberLEF-2022 DETESTS task: Detection of Racist Stereotypes in Spanish Comments using UnderBagging and Transformers. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*. CEUR Workshop Proceedings, CEUR-WS.org.

A Appendix: Co-occurrence of Stereotype Categories within a sentence

This appendix provides a heatmap of the co-occurrence of stereotype categories within a sentence to visually spot those categories that are used together more often (see Figure 5).

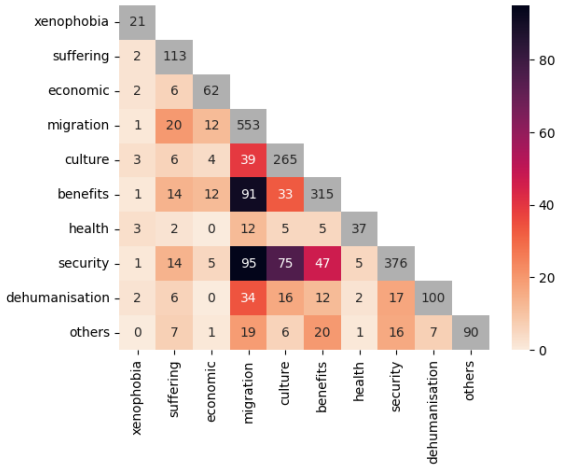


Figure 5: Heatmap representation of the sentence-level co-ocurrence of stereotype categories with the occurrence count of each category coloured in gray.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 229-240 recibido 05-07-2022 revisado 22-07-2022 aceptado 25-07-2022

Overview of EXIST 2022: sEXism Identification in Social neTworks

Overview de EXIST 2022: Identificación de Sexismo en Redes Sociales

Francisco Rodríguez-Sánchez[1], Jorge Carrillo-de-Albornoz[1], Laura Plaza[1], Adrián Mendieta-Aragón[1], Guillermo Marco-Remón[1], Maryna Makeienko[1], María Plaza[1], Julio Gonzalo[1], Damiano Spina[2], Paolo Rosso[3]
[1]Universidad Nacional de Educación a Distancia
[2]RMIT University, Australia
[3]Universitat Politècnica de València
frodriguez.sanchez@invi.uned.es,
{jcalbornoz, lplaza, gmarco,julio}@lsi.uned.es,
{mmakeienko,amendieta}@cee.uned, maria.plaza.morales95@gmail.com,
damiano.spina@rmit.edu.au, prosso@dsic.upv.es

Abstract: The paper describes the organization, goals, and results of the sEXism Identification in Social neTworks (EXIST)2022 challenge, a shared task proposed for the second year at IberLEF. EXIST 2022 consists of two challenges: sexism identification and sexism categorization of tweets and gabs, both in Spanish and English. We have received a total of 45 runs for the sexism identification task and 29 runs for the sexism categorization task, submitted by 19 different teams. In this paper, we present the dataset, the evaluation methodology, an overview of the proposed systems, and the results obtained. The final dataset consists of more than 12,000 annotated texts from two social networks (Twitter and Gab) labelled following two different procedures: external contributors and trained experts.
Keywords: Sexism Detection, Twitter, Gab, Spanish-English.

Resumen: El artículo describe la organización, objetivos y resultados de EXIST 2022 (sEXism Identification in Social neTworks), una competición que se celebra por segundo año consecutivo en el foro IberLEF. EXIST 2022 consta de dos tareas: detección de sexismo y categorización de sexismo de tweets y gabs, tanto en español como en inglés. Hemos recibido un total de 45 ejecuciones para la tarea de detección de sexismo y 29 ejecuciones para la tarea de categorización de sexismo, enviadas por 19 equipos diferentes. En el presente artículo, presentamos el conjunto de datos, la metodología de evaluación, una descripción general de los sistemas propuestos y los resultados obtenidos. El conjunto final de datos consta de más de 12.000 textos anotados de dos redes sociales (Twitter y Gab) etiquetados siguiendo dos procedimientos diferentes: colaboradores externos y expertos en el dominio.
Palabras clave: Detección de Sexismo, Twitter, Gab, Español-Inglés.

1 Introduction

The Oxford English Dictionary defines sexism as "prejudice, stereotyping or discrimination, typically against women, on the basis of sex". As stated in (Rodríguez-Sánchez, Carrillo-de Albornoz, and Plaza, 2020), sexism is frequently found in many forms in social networks, includes a wide range of behaviours (such as stereotyping, ideological issues, sexual violence, etc.) (Donoso-Vázquez and Rebollo-Catalán, 2018; Manne, 2017) and may be expressed in different forms (direct, indirect, descriptive, reported, etc.) (Mills, 2008; Chiril et al., 2020). Subtle forms of sexism are particularly dangerous as they can go unnoticed, and affect women in many facets of their lives (Swim et al., 2001; Berg, 2006).

However, research on sexism in online platforms has focused on detecting violent

ISSN 1135-5948. DOI 10.26342/2022-69-20

sexism and hate against women (Waseem, 2016; Waseem and Hovy, 2016; Frenda et al., 2019). The previous edition of EXIST (Rodríguez-Sánchez et al., 2021) was the first attempt to automatically detect and classify sexism in a broad sense, from explicit misogyny to other subtle expressions that involve implicit sexism behaviours. Therefore, the EXIST 2022 challenge is the second shared task on sexism detection in social networks whose aim is to identify and classify sexism in a broad sense. Like its first edition in 2021, EXIST 2022 has been proposed at IberLEF. During the first edition, we received a total of 70 runs for the sexism identification task and 61 for the sexism categorization challenge, submitted by 31 different teams from 11 countries, showing the great interest of the community around sexism detection in social networks.

The EXIST 2022 shared task has been focused on the same tasks as its first edition: sexism identification and categorization. Furthermore, we proposed a new test set labelled by six experts trained to perform the task. Thus, this new edition focuses on augmenting the quality of the labels and comparing the dataset labelled by crowdsourcing to expert annotators. Moreover, balance between the genders of the annotators was ensured in order to avoid gender bias in the labeling process. Annotators of different age groups were also considered.

In this second edition of EXIST, we have received a total of 45 runs for the sexism identification task and 29 runs for the sexism categorization task, submitted by 19 different teams. Results have improved with respect to the previous edition for task 1 (sexism identification) and have remained similar for task 2 (sexism categorization), which seems to indicate that classifying sexist expressions according to the facet of women they undermine is a difficult task that requires further research.

2 Tasks

2.1 Task Description

The EXIST 2022 shared task is defined as a multilingual classification task. In particular, the EXIST challenge is organized according to two main subtasks: (i) sexism identification (task 1), which aims to identify if a message or post contains sexist content; and (ii) sexism categorization (task 2), which aims to classify the type of sexism contained in a given sexist message or post. Participants were welcome to present systems that attempt both subtasks or one of them.

Task 1 is defined as a binary classification problem, where every system should determine whether a text or message is sexist or not. It includes any type of sexist expression or related phenomena, like descriptive or reported assertions where the sexist message is a report or a description of a sexist event. In particular, we consider two labels:

- **Sexist:** the tweet or gab expresses sexist behaviours or discourses.
- **Non-Sexist:** the tweet or gab does not express any sexist behaviour or discourse.

Once a message has been classified as sexist, task 2 aims to categorize the message according to the type of sexism it encloses. The categorization has been revised by two experts in gender issues, Trinidad Donoso and Miriam Comet from the University of Barcelona, and takes into account the different aspects of women that are undermined. This task is defined as a multi-class classification problem where each sexist tweet or gab must be categorized in one of the 5 following classes:

- **Ideological and inequality:** The text discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression.
- **Stereotyping and dominance:** The text expresses false ideas about women that suggest they are more suitable to fulfill certain roles (mother, wife, family caregiver, faithful, tender, loving, submissive, etc.), or inappropriate for certain tasks (driving, hardwork, etc), or claims that men are somehow superior to women.
- **Objectification:** The text presents women as objects apart from their dignity and personal aspects, or assumes or describes certain physical qualities that women must have in order to fulfill traditional gender roles (compliance with beauty standards, hypersexualization of female attributes, women's bodies at the disposal of men, etc.).

Text	Task 1	Task 2
Where are all the white women at?	non-sexist	non-sexist
Feminism is a war on men, but it's also a war on women. It's a war on female nature, a war on femininity.	sexist	ideological-inequality
Woman driving, be careful!	sexist	stereotyping-dominance
No offense but I've never seen an attractive african american hooker. Not a single one	sexist	objectification
I wanna touch your tits..you can't imagine what I can do on your body.	sexist	sexual-violence
I hate misogyny more than I hate women	sexist	misogyny-non-sexual-violence

Table 1: Examples of messages for each task.

- **Sexual violence:** Sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault) are made.
- **Misogyny and non-sexual violence:** The text expresses hatred and violence towards women.

Examples of each category for both tasks are reported in Table 1.

A substantial difference between EXIST 2022 and its first edition in 2021 is that, in 2022, the test set was labelled by 6 experts trained to perform the task, and therefore annotation quality is considerably higher. EXIST 2021, on the other hand, presented a bigger test set labelled by crowdsourcing annotators using the Amazon Mechanical Turk[1] (MTurk) platform. Moreover, EXIST 2022 takes into account women and men may differ in their perception of what is sexism, and therefore the annotation group is composed of three women and three men.

2.2 Evaluation Measures and Baselines

In order to evaluate the performance of the different approaches proposed by the participants, we will use the Evaluation Framework EvALL[2] (Amigó et al., 2017; Amigó, Spina, and Carrillo-de Albornoz, 2018; Amigó et al., 2020). Within this framework, we will evaluate the system outputs as classification tasks (binary and multi-class respectively) using standard evaluation metrics, including Accuracy, Precision, Recall, and macro-averaged F1-score.

In task 1, Sexism Identification, the results of participants will be ranked using Accuracy, as the distribution between sexist and non-sexist categories is balanced. Besides, other measures will be computed, such as Precision, Recall, and F1. All metrics will be also computed by language. In particular, Accuracy has been computed as follows:

$$Accuracy = \frac{\text{number of correctly predicted instances}}{\text{number of instances}}$$

In task 2, Sexism Categorization, we will use macro-averaged F1-score to rank the system outputs. Similarly, we will compute other measures such as Precision and Recall. The F1-score was computed as follows:

$$F_1 = \frac{F_1(\text{sexism categorization})}{6}$$

where F_1(sexism categorization) is calculated as the sum of all classes (including non-sexist):

F_1(sexism categorization) = F_1(non-sexist) + F_1(ideological-inequality) + F_1(misogyny-non-sexual-violence) + F_1(objectification) + F_1(sexual-violence) + F_1(stereotyping-dominance)

We propose two different baselines so that we can establish an expected performance of the submitted runs. First, we provided a benchmark (BASELINE_) based on Support Vector Machine (linear kernel) trained on tf-idf features built from the texts unigrams. Second, a model that labels each record based on the majority class (Majority Class).

3 Dataset

The EXIST 2022 shared task employs data from Twitter and Gab in English and Spanish. In particular, this edition uses the EXIST 2021 dataset for training and a new test set labeled by experts in the task for testing. Therefore, Twitter data was used for both training and testing while Gab was only included in the EXIST 2022 training set. This way, participants can analyse whether including data from a social network without "content control" in the training phase improves the performance of their systems. In order to build the testing data for both tasks, we employed the same terms used in EXIST 2021. In particular, the final set contains 116 seed terms for Spanish and 109 for English.

To create the new test set for this edition, we used the Twitter API to search for

[1] https://www.mturk.com/

[2] www.evall.uned.es

tweets written in English or Spanish containing some of the selected keywords. The setup of our crawler implies collecting 100 tweets for each term daily. Crawling was performed during the period from the 1st of January 2022 until the 31st of January 2022, gathering 170,210 tweets for Spanish and 206,549 for English. We have removed those with less than 60 tweets to ensure an appropriate balance between seeds. The final set of seeds used contains 91 seeds for Spanish and 94 seeds for English.

Regarding the sampling process, approximately 7 tweets were randomly selected for each seed term within the period from 1st to 31st of January 2022. We randomly resampled these tweets for each language to build the final sampled set composed of 600 tweets. The whole sampling process was defined taking into account different sources of bias. In particular, we considered three main sources: seed, temporal and user bias. We tried to mitigate seed bias by including a wide range of terms that are used in both sexist and non-sexist contexts (116 terms for Spanish and 109 for English). Temporal bias between training and testing data is mitigated since there is a temporal gap of almost one year between both sets. We also checked the temporal gap between tweets for each seed to ensure that data is spread all over the period. Finally, we checked messages generated by users to ensure an appropriate balance. We also took into account this principle to split the dataset into training and test sets and removed from the test set users who were also present in the training set to avoid user bias.

The sampled data set was labelled through a majority voting approach by six expert annotators trained to perform this task. Initially, we developed an annotation guide in English and Spanish in which we provided a clear explanation of each label along with a number of examples. We presented and explained the guidelines to ensure that all experts understood the task. Then, we did an annotation experiment proposing the 6 experts to annotate the 20% of the test set obtaining a 0.387 kappa for task 1 and 0.336 for task 2. These results indicated poor agreement and were used to modify the annotation guide and revise all the problems with the annotators. We repeated the experiment and obtained a 0.57 kappa for task 1 and 0.47 for task 2 showing a moderate agreement that aligns with the fact that the sexism detection task from a broad perspective is not simple. Sexism is even more subjective than misogyny or hate speech to women thus the labeling process is harder. The final labels were selected according to the majority vote between the 6 expert annotators in all cases. In the case of a tie, the tweet/gap was discarded. The final agreement for the whole dataset was 0.589 kappa for task 1 and 0.485 for task 2. Texts with disagreement for any of the classes were removed. The final EXIST 2022 test set consists of 1058 tweets, where all texts were randomly selected from the 1200 sampled set.

We have also tried to avoid gender bias in the annotation process by employing three female annotators and three men annotators. Gender bias may lead to algorithm bias.

The training data was provided as tab-separated, according to the following fields:

- test_case: contains the string "EXIST2021" or "EXIST2022" needed for the evaluation tool EvALL.
- id: denotes a unique identifier of the text.
- source: denotes the data source; it takes values "twitter" or "gab".
- language: denotes the language of the text; it takes values "en" or "es".
- text: contains the actual text.
- task1: defines whether the text is sexist or not; it takes values "sexist" and "non-sexist".
- task2: defines the type of sexism (if applicable); it takes values as:
  - "ideological-inequality": denotes the category "Ideological and inequality";
  - "misogyny-non-sexual-violence": denotes the category "Misogyny and non-sexual violence";
  - "objectification": denotes the category "Objectification";
  - "sexual-violence": denotes the category "Sexual violence";
  - "stereotyping-dominance": denotes the category "Stereotyping and dominance";

– "non-sexist": denotes that the tweet or gab does not express any sexist behaviours or discourses.

Concerning the test data, we removed "task1" and "task2" labels from the file that was provided to the participants.

The entire EXIST dataset contains 12,403 labeled texts, 11,345 for training corresponding to EXIST 2021 and a new test set consisting of 1,058 tweets. Table 2 summarizes the description of the dataset, as well as the number of texts per class for both training and test sets, and the distribution by language.

4 Overview of the Submitted Approaches

60 groups from 14 countries signed up for EXIST 2022, 19 of them submitted runs for task 1, and 15 for task 2. In this challenge, each team had the chance to submit a maximum of 6 runs, 3 runs for each task. We received a total of 45 runs for task 1 and 29 runs for task 2.

Regarding the classification approaches, all of the participants submitted their results using some sort of transformer-based system for both tasks with the exception of one team. In particular, 18 teams used some sort of transformer architecture, of which 8 teams used BERT (Devlin et al., 2019) (or multilingual BERT - mBERT), 5 used a Spanish version of BERT called BETO (Canete et al., 2020), 4 used RoBERTa (Liu et al., 2019), 3 used DeBERTa v3 (He, Gao, and Chen, 2021), 2 used a multilingual version of RoBERTa called XLM-R (Conneau et al., 2019) or other transformer versions. Traditional machine learning methods like decision trees or Logistic Regression (LR) have been adopted by only one team. This year, none of the teams experimented with other deep learning methods (i.e. Long short-term memory networks - LSTM) or libraries. Following, we list the participants and briefly describe the approaches used by each group.

2539404758 participated in both tasks and submitted one run for each task. They fine-tuned BERT for English texts and BETO for Spanish.

AI-UPV participated in both tasks and submitted 3 runs for each task. Their system was based on an ensemble of transformer models in a single-language and multilingual configuration. In particular, they used BERT, RoBERTa, ELECTRA (Clark et al., 2020) and GPT-2 (Radford et al., 2019) as transformer models.

AIT_FHSTP participated in both tasks and submitted 3 runs for each task. They experimented with two multilingual transformer models, such as mBERT and XLM-R, and a monolingual (English) T5 model (Raffel et al., 2019). To train the models, they used a two step approach. First, unsupervised pre-training with additional data and second, supervised fine-tuning with additional as well as augmented data. For these experiments, they employed the MeTwo dataset (Rodríguez-Sánchez, Carrillo-de Albornoz, and Plaza, 2020), HatEval 2019 dataset (Basile et al., 2019) and other hate speech related datasets.

avacaondata participated in both tasks and submitted 3 runs for each task. Their best approach to the task is based on an ensemble of different transformer models with BERTweet-large (Nguyen, Vu, and Nguyen, 2020), RoBERTa and DeBERTa v3 for English, and BETO, BERTIN (De la Rosa et al., 2022), MarIA-base (Gutiérrez-Fandiño et al., 2021) and RoBERTuito (Pérez et al., 2021) for Spanish. Models were trained in two phases. First, a validation set was used for hyperparameter optimization, second, models were trained using the whole training set.

besiguenza submitted one run for each task. Their best system was based on multilingual DeBERTa v3 and used back translation techniques to augment the EXIST dataset.

CIMATCOLMEX only participated in task 1 with three different runs. Their best approach consisted in an ensemble of 10 RoBERTuito and 10 BERT models each of them is trained individually using different seeds.

CompLingKnJ only participated in task 1 with two different runs. Their best run was based on transformers, where BETO was used for Spanish messages and BERT for English. They experimented with a system based on tf-idf features and traditional machine learning techniques.

ELiRF-VRAIN participated in both tasks and submitted three runs for each task. Their system was based in a ensemble of 5 different models for Spanish (XLM-R, RoBERTa and 3 BERT models) and other 5 models for En-

	Training				Testing		
	Twitter		Gab		Twitter		
	Spanish	English	Spanish	English	Spanish	English	Total
Sexist	2599	2494	265	300	254	215	6127
Non-sexist	2612	2658	225	192	271	305	6263
Ideological-inequality	695	619	73	100	97	64	1648
Misogyny-non-sexual-violence	600	436	58	63	32	25	1214
Objectification	368	377	50	29	18	21	863
Sexual-violence	304	494	71	48	44	43	1004
Stereotyping-dominance	632	568	13	60	60	55	1388

Table 2: Dataset distribution.

glish (XLM-R, RoBERTa, BERT, hateBERT (Caselli et al., 2020) and ALBERT (Lan et al., 2019)). Furthermore, they translated all English tweets to Spanish and vice versa and masked randomly selected tokens to augment the data available.

I2C participated with 3 runs for task 1 and one run for task 2. For their best system, they translated all Spanish tweets to English and created an ensemble of 3 models: RoBERTa, BETO and SiEBERT (Hartmann et al., 2020).

LPtower submitted 3 runs for each task. For their best run, they translated all tweets to 6 languages (French, Portuguese, Italian, German, Spanish and English) and created an ensemble of 6 models, each of them trained for a different language.

multiaztertest submitted two runs for task 1 and one run for task 2. In their best run, they fine-tuned RoBERTa for English texts and BETO for Spanish.

NIT Agartala NLP Team submitted one run for each task. Their system was based on Logistic Regression trained on tf-idf features built from the texts unigrams.

shm2022 submitted two runs for task 1 and one run for task 2. They trained the multilingual model LaBSE (Feng et al., 2020) to classify both English and Spanish tweets.

SINAI only participated in task 1 with three different runs. The best run was a system based on DistilBERT (Sanh et al., 2019). They experimented with other datasets for data augmentation.

SINAI-TL only participated in task 1 with three different runs. They followed a multi-task learning approach using different auxiliary tasks. BETO for Spanish and BERT for English were used as base models. Their best run used the emotion detection task as the auxiliary one by training a shared model with the Universal Joy dataset (Lamprinidis et al., 2021) for Spanish and, for English, they used a BERT model without auxiliary.

ThangCIC submitted 3 runs for each task. Their best system was based on an majority vote ensemble of 2 different models: mBERT and DeBERTa.

UMUTeam submitted 3 runs for each task. Their system combined linguistic features and state-of-the-art transformers using ensemble techniques. Their best model is based on a weighted ensemble model using transformers.

UNED-UPM submitted two runs for each task. For both tasks, they used a Multilingual Universal Sentence Encoder (Yang et al., 2019) as textual representation and computed the nearest neighbors to find the definitive class.

xaiTUD only participated in task 1 with a run. Their system was based on a combination of byte-level model ByT5 (Xue et al., 2022) with tabular modeling via TabNet (Arık and Pfister, 2021).

5 System Results

Tasks 1 and 2 were evaluated independently. In the following subsections, we show the results for each task and language. Teams were ranked by accuracy for task 1 and macro-averaged F1-score (F1) for task 2. However, we also report standard evaluation metrics such as Precision and Recall.

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task1_avacaondata_1	0.7996	0.7982	0.7975	0.7978
2	task1_CIMATCOLMEX_1	0.7949	0.7935	0.7952	0.794
3	task1_I2C_1	0.7883	0.7889	0.7912	0.788
4	task1_SINAI-TL_1	0.7845	0.7846	0.7868	0.7841
5	task1_multiaztertest_1	0.7836	0.7831	0.7853	0.783
6	task1_ELiRF-VRAIN_2	0.7694	0.7684	0.7704	0.7686
7	task1_UMU_1	0.7647	0.7647	0.7668	0.7642
8	task1_2539404758	0.7637	0.7619	0.7628	0.7623
9	task1_AI-UPV_3	0.7637	0.7652	0.7671	0.7635
10	task1_ThangCIC_3	0.7609	0.7598	0.7616	0.76
11	task1_LPtower_1	0.758	0.7561	0.7558	0.7559
12	task1_shm2022_1	0.7533	0.754	0.756	0.753
13	task1_AIT_FHSTP_3	0.7505	0.7494	0.7512	0.7496
14	task1_CompLingKnJ_1	0.7457	0.7446	0.7463	0.7448
15	task1_SINAI_1	0.7316	0.7353	0.7362	0.7315
16	task1_besiguenza_1	0.7306	0.729	0.726	0.7269
17	task1_NIT Agartala NLP Team_1	0.7098	0.7075	0.7059	0.7065
18	*task1_BASELINE*	*0.6928*	*0.6919*	*0.685*	*0.6859*
19	task1_UNED-UPM_1	0.6824	0.7131	0.6968	0.6792
20	*Majority Class*	*0.5444*	*0.5444*	*0.5*	*0.3525*
21	task1_xaiTUD_1	0.4811	0.5034	0.5026	0.46

Table 3: Results task 1 (best run).

5.1 Task 1

19 teams participated in task 1 for both English and Spanish, presenting 45 runs in total. In Table 3, the best run for each team is shown, as well as the two baselines: task1_BASELINE and Majority Class. All runs ranking is available at the task website[3].

Regarding the best run ranking, 14 teams achieved an Accuracy above the task1_BASELINE, while only 5 teams were below the baseline. For the Majority Class baseline, 16 teams achieved a higher Accuracy, whereas only 3 teams were below. The best performing team is *avacondata*, which achieved an overall F1 of 0.7996. This team exploited an ensemble of transformers models for different hyper-parameter configurations. The baseline based on majority vote was one of the worst performing solutions.

Although the official ranking considered both languages, we also presented two separate rankings by language (English and Spanish) for each task. Table 4 shows the top-10 runs for English and Table 5 for Spanish. Regarding the English results, the winning team *avacaondata* achieved the best results with an accuracy of 0.8422. Regarding the Spanish results, *CIMATCOLMEX* ranked first with an accuracy of 0.7801. They used an ensemble of 10 RoBERTuito and 10 BERT models, each of them trained individually using different seeds. The winning team *avacaondata* ranked fifth with more than 2% of difference in terms of accuracy.

As expected, transformer-based models performed better than the other techniques, since the top-10 teams are all based on these techniques. Traditional machine learning approaches did not perform well even using extra features based on external resources. Similarly, the use of external datasets has been explored by some teams with relative success. Specific-domain transformers have been successfully employed by the top-performed teams. This may suggest that transformer-based models benefit from training with data from the same source (e.g. Twitter).

It is interesting to highlight the performance difference (around 6%) between English and Spanish tasks. As we expected, transformer models perform better in English since they have been trained on corpus mainly composed of English texts. However, since Spanish is well-represented in these datasets, multilingual transformers perform very well for this language.

[3]http://nlp.uned.es/exist2022/

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task1_avacaondata_1.tsv_en	0.8422	0.8388	0.8365	0.8376
2	task1_SINAI-TL_1.tsv_en	0.8194	0.8148	0.8206	0.8166
3	task1_CIMATCOLMEX_3.tsv_en	0.8137	0.8087	0.8132	0.8103
4	task1_I2C_1.tsv_en	0.8137	0.8107	0.8182	0.8117
5	task1_AI-UPV_3.tsv_en	0.8118	0.807	0.8122	0.8087
6	task1_multiaztertest_2.tsv_en	0.8023	0.7981	0.794	0.7958
7	task1_ELiRF-VRAIN_3.tsv_en	0.7947	0.7893	0.7893	0.7893
8	task1_LPtower_1.csv_en	0.7852	0.7795	0.7811	0.7802
9	task1_ThangCIC_3.tsv_en	0.7852	0.7795	0.7805	0.78
10	task1_AI-UPV_1.tsv_en	0.7795	0.7789	0.7862	0.7779
11	*task1_BASELINE.tsv_en*	*0.7167*	*0.7092*	*0.7053*	*0.7068*
12	*Majority Class*	*0.5798*	*0.5798*	*0.5*	*0.367*

Table 4: Top-10 results task 1 English.

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task1_CIMATCOLMEX_1.tsv_es	0.7801	0.7808	0.7805	0.7801
2	task1_multiaztertest_1.tsv_es	0.7744	0.7753	0.7749	0.7744
3	task1_I2C_3.tsv_es	0.7707	0.7706	0.7707	0.7706
4	task1_I2C_1.tsv_es	0.7632	0.7652	0.7639	0.763
5	task1_UMU_3_es	0.7613	0.7614	0.7614	0.7613
6	task1_avacaondata_1.tsv_es	0.7575	0.7574	0.7574	0.7574
7	task1_ELiRF-VRAIN_1.tsv_es	0.7556	0.7608	0.7569	0.755
8	task1_ThangCIC_1.tsv_es	0.7556	0.7567	0.7549	0.755
9	task1_UMU_1_es	0.7556	0.757	0.7563	0.7556
10	task1_2539404758.tsv_es	0.7538	0.7539	0.7534	0.7535
11	*task1_BASELINE.tsv_es*	*0.6692*	*0.6747*	*0.6673*	*0.6649*
12	*Majority Class*	*0.5094*	*0.5094*	*0.5*	*0.3375*

Table 5: Top-10 results task 1 Spanish.

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task2_avacaondata_1	0.7013	0.5907	0.5351	0.5106
2	task2_ELiRF-VRAIN_3	0.7042	0.587	0.5057	0.4991
3	task2_UMU_2	0.6767	0.5552	0.492	0.4741
4	task2_multiaztertest_1	0.6786	0.5451	0.4826	0.4706
5	task2_ThangCIC_8	0.6626	0.5414	0.5001	0.4706
6	task2_I2C_1	0.6465	0.5255	0.518	0.47
7	task2_AIT_FHSTP_3	0.6522	0.5301	0.4999	0.4675
8	task2_LPtower_1	0.6569	0.5477	0.4748	0.4635
9	task2_AI-UPV_3	0.6267	0.519	0.5005	0.4516
10	task2_besiguenza_1	0.6285	0.4941	0.4231	0.4198
11	task2_2539404758	0.6153	0.4511	0.3939	0.3809
12	task2_UNED-UPM_1	0.5274	0.4141	0.4279	0.3708
13	*task2_BASELINE*	*0.5784*	*0.4299*	*0.3395*	*0.342*
14	task2_NIT Agartala NLP Team_1	0.6229	0.5736	0.281	0.3194
15	*Majority Class*	*0.5539*	*0.5539*	*0.1429*	*0.1018*
16	task2_shm2022_1	0.138	0.38	0.1637	0.056

Table 6: Results task 2 (best run).

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task2_avacaondata_1.tsv_en	0.7471	0.6184	0.5532	0.5337
2	task2_AI-UPV_3.tsv_en	0.6996	0.5718	0.5631	0.5133
3	task2_ELiRF-VRAIN_3.tsv_en	0.73	0.5954	0.5218	0.5084
4	task2_ThangCIC_8.tsv_en	0.6939	0.549	0.5058	0.4792
5	task2_UMU_2_en	0.7091	0.5461	0.4899	0.4751
6	task2_AI-UPV_1.tsv_en	0.673	0.5353	0.5177	0.474
7	task2_multiaztertest_1.tsv_en	0.711	0.552	0.4789	0.4689
8	task2_I2C_1.tsv_en	0.6654	0.5112	0.5193	0.4658
9	task2_AIT_FHSTP_3.tsv_en	0.6635	0.5196	0.4767	0.4545
10	task2_LPtower_1.csv_en	0.6806	0.5289	0.461	0.4515
11	*task2_BASELINE.tsv_en*	*0.5722*	*0.3836*	*0.3471*	*0.3276*
12	*Majority Class*	*0.5932*	*0.5932*	*0.1429*	*0.1064*

Table 7: Top-10 results task 2 English.

Ranking	Team_run	Accuracy	Precision	Recall	F1
1	task2_ELiRF-VRAIN_3.tsv_es	0.6786	0.5891	0.4881	0.4867
2	task2_avacaondata_1.tsv_es	0.656	0.5718	0.5169	0.4864
3	task2_UMU_1_es	0.6541	0.5985	0.5026	0.4855
4	task2_ELiRF-VRAIN_1.tsv_es	0.6767	0.5818	0.4964	0.4841
5	task2_AIT_FHSTP_3.tsv_es	0.641	0.5416	0.5215	0.4775
6	task2_I2C_1.tsv_es	0.6278	0.5432	0.5152	0.4714
7	task2_LPtower_1.csv_es	0.6335	0.5666	0.4889	0.4709
8	task2_ThangCIC_4.tsv_es	0.6316	0.5494	0.5131	0.4699
9	task2_multiaztertest_1.tsv_es	0.6466	0.5457	0.4863	0.4679
10	task2_AI-UPV_2.tsv_es	0.5545	0.4754	0.4405	0.3974
11	*task2_BASELINE.tsv_es*	*0.5846*	*0.4827*	*0.3317*	*0.3488*
12	*Majority Class*	*0.515*	*0.515*	*0.1429*	*0.0971*

Table 8: Top-10 results task 2 Spanish.

5.2 Task 2

15 teams participated in task 2 for both English and Spanish, for a total of 29 runs. In Table 6, the best run for each team is shown, as well as the two baselines. Among all the runs, 11 teams achieved an F1 above the task2_BASELINE, while only 4 teams were below it. For the Majority Class baseline, 14 teams achieved a higher F1, whereas only 1 team is below the baseline.

It is interesting to highlight the strong difference between the best and the worst systems. The best performing team for task 2 is again *avacaondata*. The worst results have been obtained by teams that employ traditional machine learning methods. Furthermore, the difference between the first and second team is more significant. This could be due to the fact that, unlike the first task, the second team does not employ a domain-adapted transformer.

Tables 7 and 8 show results for the top-10 teams in English and Spanish, respectively. Again, the task winner *avacaondata* performed better in English than in Spanish, they ranked first and second respectively. Interestingly, *ELiRF-VRAIN* performed well in Spanish by using an ensemble of 5 different models for Spanish and other 5 models for English.

In this task, the difference in performance between English and Spanish is very similar to task 1. However, it is important to notice that most participants achieved relatively low results, demonstrating the difficulty of this task and the need for further research.

6 Conclusion

In this paper, we have presented the results of the second shared task on sexism detection in a broad sense, from explicit misogyny to other subtle expressions that involve implicit sexist behaviours. The task setup provided an opportunity to test classification systems in multilingual scenarios (English and Spanish) along with different social

networks (Twitter and Gab) and labelling procedures (crowdsourcing and expert annotators). Perhaps the main contribution of this new edition is its high-quality dataset, which comprises more than 1,000 tweets labelled by experts trained to perform both tasks in the competition. We think that this dataset is a useful resource for researchers in online sexism detection.

Compared to the previous EXIST edition, the runs submitted show that the problem of sexism identification can be better addressed by using transformer-based models adapted to the Twitter domain. However, the sexism categorization still remains a challenging problem. Like in the previous edition, we found out that modern transformer-based models considerably overcome traditional machine learning approaches. Overall, the results confirm that sexism detection in social networks is challenging but there is still room for improvement.

Again, the high number of participating teams at EXIST 2022 confirms the growing interest of the community around sexism detection in social networks.

Acknowledgments

This work was partially supported by the Spanish Ministry of Science and Innovation under the project "FairTransNLP: Midiendo y Cuantificando el sesgo y la justicia en sistemas de PLN"(PID2021-124361OB-C31 and PID2021-124361OB-C32). This work was also partially funded by the Spanish Ministry of Economy and Competitiveness, as part of the research cooperation project "Space for Observation of AI in Spanish" (UNED and RED.ES, M.P., ref. C039/21-OT). The work of Paolo Rosso was in the framework of the research project PROMETEO/2019/121 (DeepPattern) by the Generalitat Valenciana.

References

Amigó, E., J. Carrillo-de Albornoz, M. Almagro-Cádiz, J. Gonzalo, J. Rodríguez-Vidal, and F. Verdejo. 2017. Evall: Open access evaluation for information access systems. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1301–1304.

Amigó, E., J. Gonzalo, S. Mizzaro, and J. Carrillo-de Albornoz. 2020. An effectiveness metric for ordinal classification: Formal properties and experimental results. *arXiv preprint arXiv:2006.01245*.

Amigó, E., D. Spina, and J. Carrillo-de Albornoz. 2018. An axiomatic analysis of diversity evaluation metrics: Introducing the rank-biased utility metric. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 625–634.

Arık, S. O. and T. Pfister. 2021. Tabnet: Attentive interpretable tabular learning. In *AAAI*, volume 35, pages 6679–6687.

Basile, V., C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Berg, S. H. 2006. Everyday sexism and posttraumatic stress disorder in women: A correlational study. *Violence Against Women*, 12(10):970–988.

Canete, J., G. Chaperon, R. Fuentes, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR*, 2020.

Caselli, T., V. Basile, J. Mitrović, and M. Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Chiril, P., V. Moriceau, F. Benamara, A. Mari, G. Origgi, and M. Coulomb-Gully. 2020. He said "who's gonna take care of your children when you are at ACL?": Reported sexist acts are not sexist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4055–4066, Online, July. Association for Computational Linguistics.

Clark, K., M.-T. Luong, Q. V. Le, and C. D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

De la Rosa, J., E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, and M. Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68:13–23.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Donoso-Vázquez, T. and Rebollo-Catalán. 2018. Violencias de género en entornos virtuales. Ediciones Octaedro.

Feng, F., Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Frenda, S., B. Ghanem, M. Montes-y Gómez, and P. Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, and M. Villegas. 2021. Spanish language models. *arXiv preprint arXiv:2107.07253*.

Hartmann, J., M. Heitmann, C. Siebert, and C. Schamp. 2020. More than a feeling: Accuracy and application of sentiment analysis.

He, P., J. Gao, and W. Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Lamprinidis, S., F. Bianchi, D. Hardt, and D. Hovy. 2021. Universal joy: A data set and results for classifying emotions across languages. In *The 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Manne, K. 2017. *Down girl: The logic of misogyny*. Oxford University Press.

Mills, S. 2008. *Language and sexism*. Cambridge University Press.

Nguyen, D. Q., T. Vu, and A. T. Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Pérez, J. M., D. A. Furman, L. A. Alemany, and F. Luque. 2021. Robertuito: a pre-trained language model for social media text in spanish. *arXiv preprint arXiv:2111.09453*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rodríguez-Sánchez, F., J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Rodríguez-Sánchez, F., J. Carrillo-de Albornoz, and L. Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Swim, J., L. Hyers, L. Cohen, and M. Ferguson. 2001. Everyday sexism: Evidence for its incidence, nature, and psychological impact from three daily diary studies. *Journal of Social Issues*, 57:31 – 53.

Waseem, Z. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November. Association for Computational Linguistics.

Waseem, Z. and D. Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June. Association for Computational Linguistics.

Xue, L., A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Yang, Y., D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 241-253 recibido 05-07-2022 revisado 22-07-2022 aceptado 25-07-2022

Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of the LivingNER shared task and resources

Detección, normalización y clasificación de especies, patógenos, humanos y alimentos en documentos clínicos: resumen de la tarea y los recursos LivingNER.

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima-López, Darryl Estrada, Luis Gascó, Martin Krallinger
Barcelona Supercomputing Center, Spain
antoniomiresc@gmail.com

Abstract: There is a pressing need to generate tools for finding mentions of species, pathogens, or food from medical texts. To promote the development of such tools we organized the LivingNER task. LivingNER relied on a large Gold Standard corpus of 2000 carefully selected clinical cases in Spanish covering diverse specialties. It was manually annotated with species mentions that were also carefully mapped to their corresponding NCBI Taxonomy identifiers. Besides, we have generated Silver Standard versions of LivingNER for 7 languages: English, Portuguese, Galician, Catalan, Italian, French, and Romanian. LivingNER had three subtasks: LivingNERSpecies NER (species mention detection sub-task), LivingNER-Species Norm (species mention detection and normalization to NCBI taxonomy Ids), and LivingNERClinical IMPACT (a document classification task related to the detection of pets, animals-causing injuries, food, and nosocomial entities). We received and evaluated 62 systems from 20 teams from 11 countries worldwide, obtaining highly competitive results. Successful approaches typically modified pre-trained transformer-like language models (BERT, BETO, RoBERTa, etc.) and employed embedding distance metrics for entity linking. LivingNER corpus: doi.org/10.5281/zenodo.6376662
Keywords: named entity recognition, pathogens text mining, entity linking, NCBI Taxonomy.

Resumen: Existe la necesidad de generar herramientas para encontrar y normalizar menciones de especies, patógenos o alimentos en textos médicos. Para promover el desarrollo de tales herramientas hemos organizado la tarea LivingNER. La tarea LivingNER se basó en un corpus en español de 2000 casos clínicos cuidadosamente seleccionados, representando una diversidad de especialidades. El corpus fue anotado manualmente por expertos que también asignaron a las menciones sus correspondientes identificadores de la NCBI Taxonomy. Además, hemos generado versiones de LivingNER para otros 7 idiomas: inglés, portugués, gallego, catalán, italiano, francés y rumano. LivingNER se estructuró en tres subtareas: 1) LivingNER-Species NER (subtarea de detección de menciones de especies), 2) LivingNER-Species Norm (detección de especies y normalización a identificadores de NCBI Taxonomy) y 3) LivingNER-Clinical IMPACT (tarea de clasificación relacionada con la detección de mascotas, animales causantes de lesiones, alimentos y entidades nosocomiales). Recibimos y evaluamos 62 sistemas de 20 equipos de 11 países a nivel mundial, obteniendo resultados altamente competitivos. Generalmente, los enfoques más exitosos hicieron modificaciones a modelos de lenguaje basados en transformers (BERT, BETO, RoBERTa, etc.) y emplearon métricas de distancia de embeddings para la normalización de entidades. Corpus LivingNER: doi.org/10.5281/zenodo.6376662
Palabras clave: reconocimiento de entidades nombradas, minería de textos de patógenos, normalización de entidades, NCBI Taxonomy.

ISSN 1135-5948. DOI 10.26342/2022-69-21

1 Introduction

The semantic annotation of species or living organisms is critical to scientific disciplines like medicine, biology, ecology/biodiversity, nutrition, and agriculture. For instance, detecting species in clinical records underscores the burden of disease caused by pathogens in the case of infectious diseases; and identifying organisms and foods can reveal the cause of allergy-related conditions. Despite this undisputed relevance, organisms/species have relatively scarcely featured in NLP studies, particularly for non-English content.

Because of the significance of this task, hierarchical taxonomic relations have been developed over 250 years to determine rules and conventions to catalog species. And they have been recently transformed into computer-based terminological resources such as NCBI taxonomy (Schoch et al., 2020; Federhen, 2012), the Thompson scientific name list, the Catalogue of Life, the Global Names Index database, and the ITIS Catalogue. However, these efforts have not been adequately aligned with the development of automatic systems for semantic analysis of species mentions in text, especially when considering documents beyond English. Common challenges encountered are name changes (obsolete species names); homonymy with commonly used words (e.g., "spot" refers to the species *Leiostomus xanthurus* or "permit" to *Trachinotus falcatus*); abbreviations and acronyms (sometime highly ambiguous like EC, which can be used for the bacteria "*Escherichia coli*" and "*Enterobacter cloacae*," among others); misspelled names (*Escerichia coli* for *Escherichia coli*); coordinations and nested expressions ("human immunodeficiency viruses types 1 and 2"); vernacular forms (common names); and role names (e.g., athletes, responders).

To overcome these limitations, corpora and tools are already available for species identification in the English-language biomedical literature and their standardization to controlled vocabularies. For example, LINNAEUS (Gerner, Nenadic, and Bergman, 2010) and the SPECIES tool (Pafilis et al., 2013) are capable of detecting species mentions. Additionally, there have been shared tasks on information related to microorganisms/species, such as the Infectious Diseases (ID) task of BioNLP 2011 (Pyysalo et al., 2011). And the importance of detecting species mentions for gene mention entity linking to database records has been addressed using biomedical literature data in English (Krallinger, Leitner, and Valencia, 2010).

However, adapting these resources to languages other than English and document types different from biomedical literature is not trivial. This is aggravated by the lack of resources, common evaluation scenarios, and shared tasks in other languages.

The LivingNER task addressed these issues through (1) a challenge on Named Entity Recognition (NER) of species mentions, entity linking, and document classification and; (2) providing a manually, exhaustively annotated large corpus of Spanish clinical cases. All annotated organism/species mentions were manually mapped to the NCBI taxonomy and classified into four information axes related to relevant use cases.

The National Center for Biotechnology Information (NCBI) Taxonomy includes names of organisms classified primarily based on a phylogenetic hierarchy. The NCBI Taxonomy is a universal database, used by the International Nucleotide Sequence Database Collaboration (INSDC), which includes GenBank, the European Molecular Biology Laboratory (EMBL), and DNA Data Bank of Japan (DDBJ) as a single source of taxonomic classification to maintain consistency between databases. In NCBI, each unique code identifies a specific type of organism (e.g., Taxonomy ID: 5476 for *Candida Albicans*) or groups of organisms (Taxonomy ID: 40674 for mammals). NCBI Taxonomy was the controlled vocabulary chosen in the LINNAEUS corpus to standardize citations.

The corpus also distinguishes between antibiotic-resistant pathogens and hospital-acquired (nosocomial) infections, an increasing cause of morbidity and mortality when existing drugs become ineffective in eliminating some bacteria. It also references and standardizes mentions of the different floras of the human organism in preparation for the literature and clinical cases related to the human microbiome. For clinical and microbiological use, all forms of parasitic cycles are also noted. Some of the most relevant applications are associated with extracting information about highly prevalent sexually transmitted diseases, animals causing injuries, and animal-transmitted diseases (zoonoses) originating from pets and animal husbandry. The

correct extraction of species and infectious diseases facilitates the classification of bacteria in the context of antibiotic resistance and nosocomial pathogens and diseases. Another potential application relates to food (infection, intoxication, healthy and unhealthy diets, etc.), allergy triggers, and epidemiologically relevant mentions such as close contacts, people living in the same household, and relatives.

LivingNER is the first track on comprehensive species mention recognition and grounding of non-English content with a clear potential for multilingual adaptation, particularly for pathogens, to generate high-quality living being mention recognition components. The LivingNER annotation guidelines and corpus are indispensable resources for detecting and classifying species and infectious diseases in Spanish-language literature and medical reports.

2 Task Description

2.1 Shared Task goal

The LivingNER shared task explores the automatic recognition of species mentions in clinical documents in the Spanish language, the assignment of NCBI Tax IDs, and the classification of each mention into four categories. Notably, LivingNER incorporates a subtask in which participants solve four real-world health use cases.

2.2 Sub-tasks

The LivingNER track contains three independent subtasks that are built one on top of the other:

LivingNER-Species NER track (Species mention entity recognition): given a plain text clinical case report document collection, participants must return the exact character offsets of all species mentions, both human and non-human.

LivingNER-Species Norm track (Species mention normalization): given a plain text clinical case report document collection, participating systems have to return all species mentions, together with their corresponding NCBI taxonomy concept identifiers.

LivingNER-Clinical IMPACT track given a collection of plain text documents, systems must (1) Perform a document classification according to information relevant to high-impact, real-world clinical use cases. The classification is multi-label, meaning that a single document may belong to several categories. And (2) Retrieve the list of NCBI Tax IDs that support the binary classification. Systems have to categorize the documents into the following information axes:

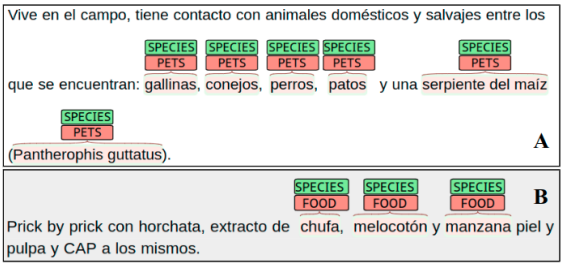


Figure 1: LivingNER example sentences annotated with Clinical Impact entities. (A) for Pets and farm animals, (B) for food species.

- *Pets and farm animals* in close contact with the patient (important for detecting animal-transmitted diseases such as toxoplasmosis, salmonellosis, cat-scratch disease, etc.).
- *Animals causing injuries*. Parasites are NOT included.
- *Food species*. It includes ingested aliments and any other food mentioned in the document. It excludes ingested items that are not food.
- *Nosocomial entities*: mentions corresponding to nosocomial/healthcare-associated infections.

2.3 Evaluation metrics

The micro-average f1-score has been the primary evaluation metric in the three subtasks. Additionally, micro-average precision and recall have been computed. The LivingNER evaluation library is available on GitHub (github.com/tonifuc3m/livingner-evaluation-library).

2.4 Baseline

For the LivingNER-Species NER subtask, we have employed the PathoTagIt-Base system. This competitive baseline is a deep neural network system trained with the LivingNER training dataset. The network is a customization of the BiLSTM-CRF architecture, and it employs word embeddings optimized for biomedical Spanish language (Soares et al., 2019). For a more in-depth description of the

system, check the PharmaCoNER tagger paper (Armengol-Estapé et al., 2019). The code is available on GitHub (github.com/TeMU-BSC/PharmaCoNER-Tagger). There is also a web demo of the PathoTagIt-Base system (see temu.bsc.es/livingner/).

Finally, we have followed an indirect approach to create the document classifier of the LivingNER-Clinical Impact subtask. We have trained four different NER systems. The first NER system recognizes pet and farm animal mentions; the second mentions animals causing injuries; the third, food mentions; and the last, nosocomial entities. The four NER systems were run on the test set documents. The document containing it is automatically classified into the mention category if a mention is detected. For instance, in Figure 1 B, as soon as the NER system of food mentions recognizes "chufa", "melocotón", or "manzana", the document would be classified as a "food document".

3 Corpus and Resources

3.1 LivingNER Gold Standard Corpus

The LivingNER corpus is a collection of 2,000 clinical cases in Spanish from 20 medical specialties: infectious diseases (including Covid-19 cases), cardiology, neurology, oncology, ENT, dentistry, pediatrics, endocrinology, primary care, allergology, radiology, psychiatry, ophthalmology, psychiatry, urology, internal medicine, emergency and intensive care medicine, radiology, tropical medicine, and dermatology annotated with species [SPECIES] (including living organisms and microorganisms) and infectious diseases [ENFERMEDAD] mentions. Each mention in the corpus has been standardized to NCBI Taxonomy terminology. Finally, the species mentions have been classified into four classes of clinical interest to improve their usability (companion animals, animals causing injuries, food, and nosocomial entities).

The infectious diseases annotations are not used in the LivingNER shared task.

Document selection. The objective of document selection was to obtain a sufficient diversity of mentions representative of species in the clinical domain. We were mainly limited by the availability of relevant documents for certain specialties. For instance, obtaining clinical reports on tropical diseases was much easier than on pediatric allergies. The documents were also selected based on the richness of mentions, favoring the reports with a larger variety of species. Finally, we revised that certain diseases of great interest, notably COVID-19, but also zoonoses and parasite infections, AIDS, hepatitis C and others, were not excluded from our selection.

Corpus annotation. The LivingNER corpus has been annotated and standardized by a domain specialist with the support of a clinical specialist, who was also in charge of reviewing the mentions and their associated codes to arrive at a final version. The process of annotation and normalization of the corpus took place between 2020 and 2021, lasting approximately five months using the brat tool. Before starting the annotation, a first draft of these guides was created based on our previous annotation experiences MEDDOCAN (Marimon et al., 2019), CANTEMIST (Miranda-Escalada, Farré, and Krallinger, 2020) or MEDDOPROF (Lima-López et al., 2021) among others), and previous related work (Pafilis et al., 2013; Gerner, Nenadic, and Bergman, 2010). The annotation guidelines were refined by several rounds of inter-annotator agreement (IAA) consisting of parallel annotation of 5% of the corpus. After several rounds, a total IAA score of 0.942 for species and 0.885 for infectious diseases was reached. In addition, during the remainder of the LivingNER annotation, a random 10% of the papers were thoroughly reviewed to ensure that quality was maintained. There was also ongoing discussion about the content of the corpus, especially about difficult and ambiguous cases, with the aim of achieving the highest possible quality and refining these guidelines as much as possible.

The NCBI Taxonomy terminology was used to assign an identifier to each manual annotation, ensuring the usability of the corpus citations. The final version of the LivingNER corpus includes 30886 species mentions, of which 43.9% correspond to humans, 4580 are unique, and 29411 are normalized to NCBI Taxonomy. In addition, it contains 11841 infectious disease mentions, 4093 of which are unique, and 2283 are normalized to NCBI Taxonomy. The total is 42727 mentions. Finally, all species entries have been classified into four classes of clinical interest to improve their use (companion animals, animals causing injuries, food, nosocomial enti-

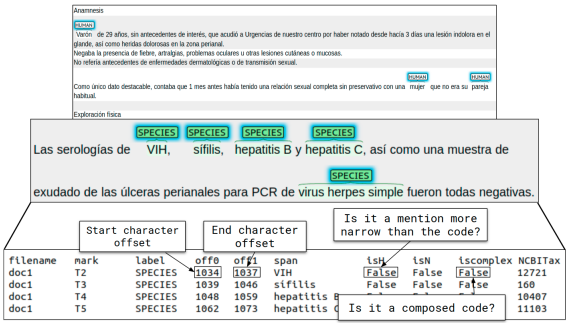


Figure 2: Annotated clinical case visualized with Brat tool and annotation tab-separated format.

ties, and antibiotic-resistant bacteria).

Corpus format. The LivingNER clinical case documents are released in plain text format with UTF-8 encoding. The annotations are included in a tab-separated document. In the LivingNER-SPECIES NER task, the annotations file has the following columns: filename, mark (identifier mention mark), label (SPECIES or HUMAN), off0 (starting position of the mention in the document), off1 (ending position of the mention in the document) and span. The LivingNER-Species Norm file, in addition to these columns, includes four more columns: isH (whether the span is narrower than the NCBITax assigned code), isN (whether the mention corresponds to a nosocomial infection), iscomplex (whether the span has assigned a combination of NCBITax codes) and NCBITax (mention code in the NCBI Taxonomy). Finally, the LivingNER-Clinical Impact annotation file has the following columns: filename, isPet, PetIDs (NCBITaxonomy codes of pet and farm animals present in document), isAnimalInjury, AnimalInjuryIDs (NCBITaxonomy codes of animals causing injuries present in document), IsFood, FoodIDs, (NCBITaxonomy codes of food mentions present in document), isNosocomial and NosocomialIDs (NCBITaxonomy codes of nosocomial species mentions present in document) (see Figure 3).

Corpus statistics. The LivingNER corpus contains 1,985 documents, which amounts to 65,373 sentences and 1,234,579 tokens. The corpus was randomly split into three subsets: training, validation, and test set. The test set is used for evaluation purposes of participating teams and consists of 485 records (15 extra records will be released shortly). Species and human mentions are found in all 1,985 documents. There are 30,604 such mentions (17158 species and 13446 human mentions) manually mapped to an NCBI Taxonomy ID. All human mentions have the 9606 NCBI Taxonomy ID, and there are 2,672 other unique codes. See Table 1 for the LivingNER corpus general statistics.

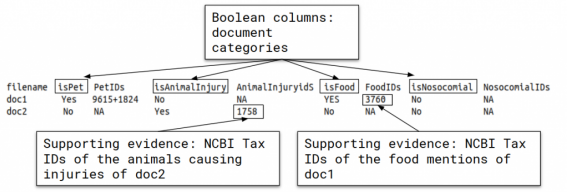


Figure 3: LivingNER Clinical Impact data format.

The 15 most common SPECIES mentions are shown in Figure 4B. It is noteworthy that seven out of the ten most common have the HUMAN label, despite there being fewer HUMAN annotations. This is because it is a more homogeneous entity type. Indeed, there are 707 different HUMAN mentions, while there are 3818 different SPECIES mentions.

In Figure 4.A, the 15 most common SPECIES NCBI Tax IDs are displayed. The main term of the code is shown instead of the numeric ID for clarity. While some terms are general (prokaryotes, viruses, eukaryotes), others are specific (HIV, *Enterobius vermicularis*, etc.) HIV appears very frequently partially because it is commonly mentioned in the context of patient serology results.

	Training	Validation	Test	Total
Documents	1000	500	485	1,850
SPECIES Annotations	9090	3817	4251	17158
HUMAN Annotations	7007	3289	3150	13446
Total Annotations	16097	7106	7401	30604
Unique codes	6738	2833	3101	12672
Sentences	34261	15107	16005	65373
Tokens	642813	296161	295605	1234579
Pets and farm animals	45	14	21	80
Animal causing injuries	107	12	22	141
Food species	255	107	163	525
Nosocomial entities	67	21	10	98

Table 1: DrugProt Gold Standard corpus statistics.

3.2 LivingNER Annotation Guidelines

The annotation guidelines posed many challenges, since many mentions of species in clinical documents are not identical to what

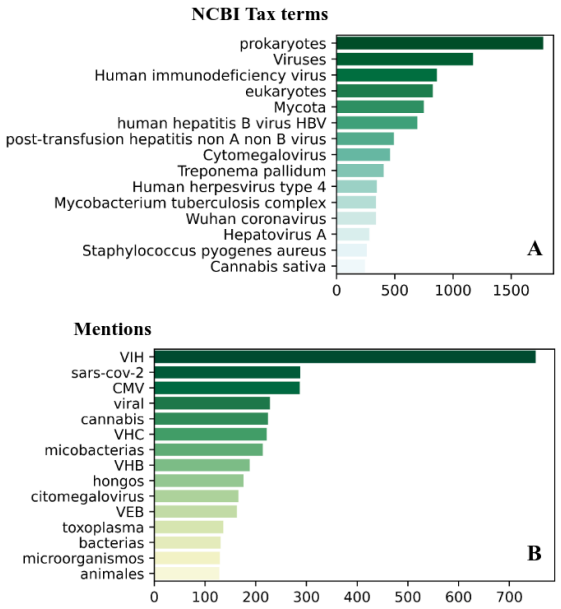


Figure 4: Number of appearances of (A) the main terms of the 15 most common codes, and (B) the 15 most common entities in the LivingNER Gold Standard.

we can find in the terminologies (for instance, hepatitis C virus is usually found as acronym, i.e., HCV, and *Staphilococcus aureus* as Staph A. and even its vernacular form, i.e., estafilococo). However, language was not the only challenge. We had to determine whether to include mentions undoubtedly related to infectious diseases and thus to pathogens, such as the term *vaccine*, if practically pathognomonic laboratory tests could be equated to the infection and thus the microorganism (i.e., VDRL to syphilis and thus to *Treponema pallidum*), and if we should distinguish between humans since their prevalence and responses to infection might greatly differ (for instance, neonates, children, males, females, the elderly). Naturally, these findings and decisions weighed heavily on the normalisation. We also wanted to include references to various parasite phases, since they are very important for microscopic diagnosis, and to the human microbiome, particularly the gut microbiome, since its study has imploded in the last decades, and the use of faecal transplant is already used to treat resistant *Clostridium difficile* infections and a large number of clinical trials to treat other conditions with human flora are under way. The current annotation guidelines are well adapted to capture pathogens and species, and also to expand with the advance of molecular microbiology and scientific knowledge.

3.3 LivingNER Multilingual Silver Standard

To foster the development of multilingual tools and generate systems not only for Spanish but also for content in English and various Romance languages, we have developed the annotated (and normalized to NCBI Taxonomy) LivingNER corpus in 7 languages: English, Portuguese, Galician, Catalan, Italian, French, and Romanian. The overview statistics of the Silver Standard are shown in Table 2. We refer to the DisTEMIST overview paper (Miranda-Escalada et al., 2022) for a complete description of the generation process since it is equivalent to that corpus. Find the Multilingual Silver Standard at Zenodo (doi.org/10.5281/zenodo.6376662)

		Documents	Annotations	Unique NCBI Tax IDs	Sentences	Tokens
Catalan	**Training**	1000	14803	832	34173	642926
	Valid	500	6724	533	15073	297012
	Test	485	7709	548	15979	296124
English	**Training**	1000	13772	776	34430	624437
	Valid	500	6332	493	15180	287164
	Test	485	7225	513	16075	286419
French	**Training**	1000	12419	754	34552	697869
	Valid	500	5540	471	15225	322198
	Test	485	6428	498	16107	321766
Italian	**Training**	1000	12945	759	34373	649831
	Valid	500	5846	470	15130	299907
	Test	485	6703	499	16038	299470
Portuguese	**Training**	1000	12420	727	34330	641095
	Valid	500	5642	470	15143	295942
	Test	485	6738	490	16038	295587
Romanian	**Training**	1000	10522	699	34334	651595
	Valid	500	4799	427	15130	300773
	Test	485	5617	478	16029	300297
Galician	**Training**	1000	16633	875	34188	616216
	Valid	500	7319	555	15065	284023
	Test	485	7672	546	15983	283808

Table 2: LivingNER Multilingual Silver Standard corpus statistics.

3.4 LivingNER Terminology

It is the official NCBI Taxonomy FTP dump (ftp.ncbi.nlm.nih.gov/pub/taxonomy/) with the terms translated to Spanish by a Neural Machine Translator fine-tuned for the biomedical domain. It is a tab-separated file with the following columns: *tax_id* (the NCBI Taxonomy ID of node associated with this name), *name_txt* (the NCBI Taxonomy name), *unique name* (the unique variant of this name if the name is not unique), *name class* (synonym, common name, scientific name, ...), *Spanish name* (the NCBI Taxonomy name in Spanish).

Besides, we have added the following terms: 2560602 (Mumps orthorubulavirus), 2560526 (Human orthorubulavirus 4),

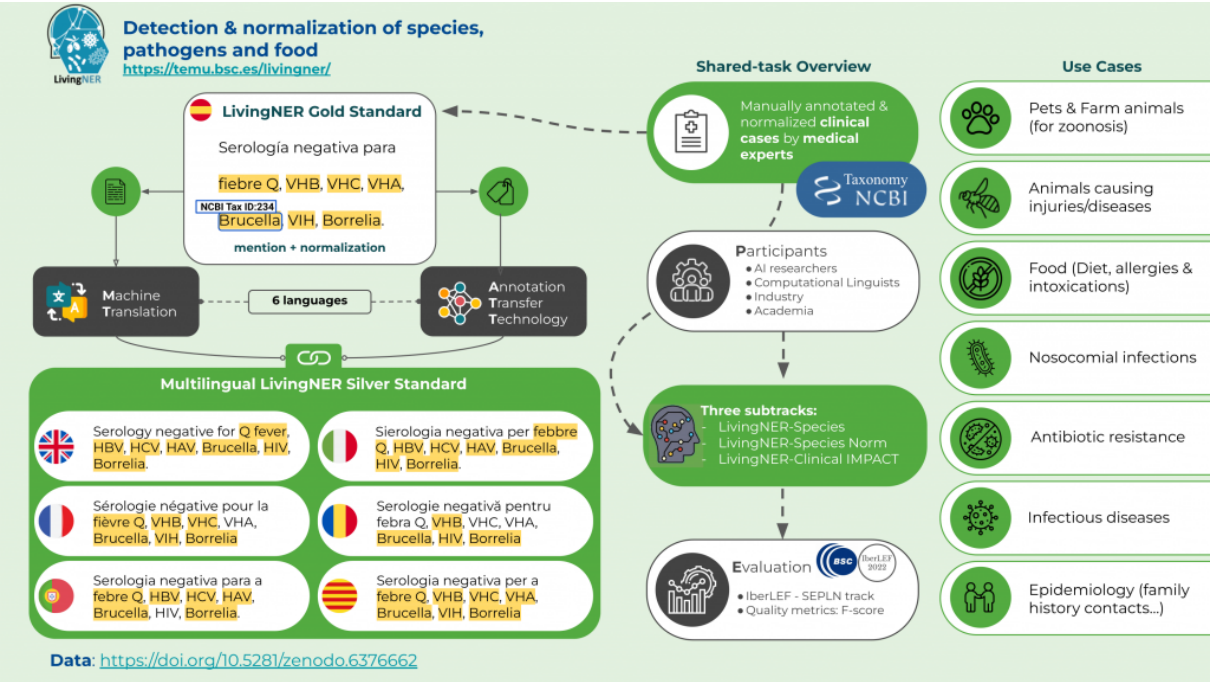


Figure 5: LivingNER multilingual corpus overview.

2847144 (hepatitis C virus genotype 1a), _NOCODE_ (out of NCBI Taxonomy scope). The first three were added because they appear in the LivingNER corpus, and are present in the browser version of NCBI Taxonomy. The last one (_NOCODE_) is added to identify terms in the LivingNER corpus that are not present in the NCBI Taxonomy.

The terminology is available at Zenodo (doi.org/10.5281/zenodo.6390506).

4 Results

4.1 Participation Overview

The community has shown an active interest in LivingNER. There were 56 teams registered in LivingNER, and 20 successfully submitted their system results, totaling 62 submissions. 20 teams participated in LivingNER-SPECIES NER [41 runs], 8 also submitted their system predictions for LivingNER-SPECIES Norm [15 runs], and 5 did it for the LivingNER Clinical Impact track [6 runs]. Besides, as Table 3 shows, participants belonged to institutions (industry or academia) from different countries, including Spain, Romania, China and México.

4.2 System Results

Table 5 shows the best-run results by all teams for subtasks LivingNER-Species NER and LivingNER-Species Norm. In LivingNER-Species NER, the Vicomtech NLP team obtained the highest micro-average F1-score, 0.951. Team RACAI F1-score was almost tied with Vicomtech (0.9503), and it reached the highest precision (0.9622) and recall (0.9439) in different submissions.

In LivingNER-Species Norm, the highest F1-score (0.9304) and recall (0.9234) were obtained once again by the Vicomtech NLP team. The highest precision was obtained by the ClaC team (0.9641).

Table 4 contains the best-run results of the third subtask, LivingNER-Clinical Impact. In this case, participants had to classify the test set documents into four categories and include the NCBI Taxonomy codes justifying the classification. Results were computed for the document classification task and the document classification + code justification. The baseline system was available for the first task (document classification), and none of the participant teams outperformed it. We discuss this in the Discussion section. We must outline that only 4 test set documents were positive Nosocomial documents. Therefore, the results for this fourth classification axis are challenging to interpret.

Finally, the complete results of all runs, plus the disaggregated results

Team Name	Affiliation	Country	Tasks	Ref.	Tool URL
Vicomtech NLP	Vicomtech	Spain	NE/No/C	(Zotova et al., 2022)	–
racai	Research Institute for Artificial Intelligence "Mihai Draganescu"	Romania	NE	(Avram, Mitrofan, and Pais, 2022)	–
READ-Biomed	RMIT University	Australia	NE	(Jimeno Yepes and Verspoor, 2022)	–
SINAI	Universidad de Jaén	Spain	NE/No/C	(Chizhikova et al., 2022)	–
plncmm	CMM, University of Chile	Chile	NE/No/C	(Rojas et al., 2022)	(plncmm, 2022)
Sumam Francis	KU Leuven	Belgium	NE	(Francis and Moens, 2022)	–
Clac	Concordia University	Canada	NE/No	(Bagherzadeh, Verma, and Bergler, 2022)	–
john_snow_labs	John Snow Labs	USA	NE	(Kocaman et al., 2022)	–
avacaondata	IIC (ADIC)	Spain	NE/No/C	(Vaca, 2022)	–
Pumas	Universidad Nacional Autónoma	México	NE/No/C	(del Moral et al., 2022)	–
IAM	University of Bordeaux	France	NE	(Cossin, Diallo, and Jouhet, 2022)	(IAM, 2022)
IGES	IGES Institut GmbH	Germany	NE/No	(Chapman, Schwarz, and Häussler, 2022)	–
NLP-CIC-WFU	Instituto Politécnico Nacional Wake Forest University	México & USA	NE/No	(Tamayo, Burgos, and Gelbukh, 2022)	(NLP-CIC-WFU, 2022)
Vitor	Universidade Federal do Rio de Janeiro	Brasil	NE	–	–
zzz	Yunnan University	China	NE	(Zhu and Wang, 2022)	(zzz, 2022)
Kformer-OEG	Universidad Politécnica de Madrid	Spain	NE	–	–
Mark	–	–	NE	(Hanjie and Xiaobing, 2022)	(Mark, 2022)
Han	Yunnan University	China	NE	(Han and Ding, 2022)	(tutorial, 2022)
Sapphire	–	–	NE	–	–
boun-ner	Bogazici University	Turkey	NE	–	–

Table 3: LivingNER team overview. In the Tasks column, NE stands for LivingNER-Species NER, No for LivingNER-Species Norma and C for LivingNER-Clinical Impact.

	Pets and farm animals			Animals causing injuries			Food species			Nosocomial entities		
Team Name	**MiP**	**MiR**	**MiF**	**MiP**	**MiR**	**MiF**	**MiP**	**MiR**	**MiF**	**MiP**	**MiR**	**MiF**
LivingNER-Clinical Impact with codes												
Vicomtech	0	0	0	.0006	.125	.0012	.0088	.1154	.0164	0	0	0
SINAI	0	0	0	0	0	0	0	0	0	0	0	0
plncmm	.0317	.3636	.0584	0	0	0	.02	.3846	.038	0	0	0
avacaondata	0	0	0	0	0	0	0	0	0	0	0	0
Pumas	.024	.25	.0438	0	0	0	.0211	.2692	.0391	0	0	0
LivingNER-Clinical Impact												
Vicomtech	.0326	.25	.0577	.0058	.5	.0115	.0235	.3077	.0437	.0016	.75	.0032
SINAI	0	0	0	0	0	0	0	0	0	0	0	0
plncmm	.0397	.4167	.0725	.0282	.5	.0533	.0479	.9231	.0911	.006	.5	.0118
avacaondata	0	0	0	.0021	.125	.0041	0	0	0	0	0	0
Pumas	.024	.25	.0438	.0167	.25	.0312	.0211	.2692	.0391	0	0	0
PathoTagIt-Base	1	.1667	.2857	.032	1	.062	.8	.9231	.8571	.0513	.5	.093

Table 4: Results of LivingNER-Clinical Impact systems. MiP, MiR and MiF stands for micro-averaged precision, recall and F1-score.

by label (HUMAN and SPECIES), are published on a dedicated webpage (temu.bsc.es/livingner/results/).

4.3 Methodologies

Table 5 briefly describe the methodologies used by LivingNER participants, and for an in-depth description, we refer to their scientific articles, listed in Table 3. We have observed that the most successful approaches to LivingNER-Species NER included non-standard fine-tuning of pre-trained transformer-based language models. Typically, these language models are domain and language-specific, such as bsc-bio-es RoBERTa (Carrino et al., 2021), employed by teams READ-Biomed and SINAI, among others; or cross-lingual, such as XLM-RoBERTa (Conneau et al., 2019), chosen by team racai. The highest-scoring participant of LivingNER-Species NER, Vicomtech, has fine-tuned a transformer-based language model using a sliding windows technique that avoids hard, meaningless segmentation cuts that typically occur in these scenarios (Zotova et al., 2022).

In LivingNER-Species Norm, participants with the highest scores used a robust NER system to detect the species mentioned. And they were mapped to NCBI Taxonomy using traditional approaches such as string matching using Levenshtein distance and setting a heuristic cutoff. Additionally, other participants used, for instance, word embeddings similarity (Pumas) or TF-IDF matching (SINAI).

Finally, in LivingNER-Clinical Impact, the most successful approach has been the baseline: to train a simple NER system to recognize the entities of interest and label as positive any document with a detected named entity.

	SPECIES NER				SPECIES Norm			
Team Name	**MiP**	**MiR**	**MiF**	**Description**	**MiP**	**MiR**	**MiF**	**Description**
Vicomtech NLP	.9583	.9438	**.951**	sophisticated fine-tune transformer model	0.9376	**0.9234**	**0.9304**	Semantic Text Search approaches
racai	.9569	**.9439**	.9503	fine-tune XLM-RoBERTa with lateral inibitory layer	-	-	-	-
READ-Biomed	.954	.9411	.9475	Fine-tune RoBERTa (bsc-bio-es)	-	-	-	-
SINAI	.9571	.9346	.9457	fine-tune RoBERTa (roberta-base-bne, bsc-bio-es & roberta-biomedical-clinical-es)	.8733	.8527	.8629	character-level TF-IDF matching and string matching w. Levenshtein distance
plncmm	.9455	.9373	.9414	Fine-tune RoBERTa (bsc-bio-es) w. FLERT	.9139	.906	.9099	string matching w. Levenshtein distance
Sumam Francis	.9443	.9307	.9375	Fine-tune BERT (BETO) pre-trained w. contrastive loss	-	-	-	-
Clac	.9385	.9256	.932	mi-RIM model	.9495	.891	.9193	string matching w. Levenshtein distance
john_snow_labs	.916	.9327	.9243	Bi-LSTM-CNN-Char & BertForTokenClassification	-	-	-	
avacaondata	.9228	.908	.9153	Domain adaptation of MarIA-Large	.512	.4799	.4954	-
Pumas	.9284	.8899	.9087	fine-tune RoBERTa (bsc-bio-es)	.9389	.8075	.8682	word embedding similarity
IAM	.9209	.8733	.8965	Complex dictionary lookup	-	-	-	-
IGES	.9112	.8638	.8869	SAPBert-XLMR + CRF	.8979	.8512	.874	FAISS indexes containing encoded synonyms
NLP-CIC-WFU	.8303	.8704	.8499	fine-tune mBERT & post-processing rules	.7768	.8143	.7951	dictionary lookup
Vitor	.9492	.5634	.7071	-	-	-	-	-
zzz	.8012	.6138	.6951	fine-tune BERT+BiLSTM	-	-	-	-
Kformer-OEG	.7306	.6057	.6623	-	-	-	-	-
Mark *pw	.8214	.6145	.703	BERT(BETO)+BiGRU+ CRF + adversarial learning	-	-	-	-
Han *pw	.5399	.1965	.2881	fine-tune BERT (BETO)	-	-	-	-
Sapphire	.6875	.0149	.0291	-	-	-	-	-
Boun-ner	0.126	0.078	0.0963	fine-tune BERT	-	-	-	-
PathoTagIt-Base	0.9461	0.8507	0.8958	Section 2.5	-	-	-	-

Table 5: Results of LivingNER systems, subtasks SPECIES NER and SPECIES Norm. **pw* means post-workshop submissions. MiP, MiR and MiF stands for micro-averaged precision, recall and F1-score.

4.4 LivingNER Spanish Silver Standard

The LivingNER test set was released together with a background set: an additional collection of 13,000 clinical case documents from various medical disciplines, all Spanish. The background set helps examine whether systems could scale to more extensive data collections and avoid manual annotation correction. Participants have generated automatic predictions for the test and the background set, although they were only evaluated on the test set predictions in the three subtasks.

Therefore, the background set predictions include automatic mention annotations (LivingNER-Species NER predictions), normalized to NCBI Taxonomy (LivingNER-Species Norm predictions) and document classifications with evidence (LivingNER-Clinical Impact predictions). The background set predictions from all participants will be harmonized and constitute the LivingNER Spanish Silver Standard corpus, similar to the CALBC initiative (Rebholz-Schuhmann et al., 2010), to the Cantemist (Miranda-Escalada, Farré, and Krallinger, 2020), CodiEsp (Miranda-Escalada et al., 2020), MESINESP2021 (Gasco et al., 2021), ProfNER (Miranda-Escalada et al., 2021), and PharmaCoNER (Gonzalez-Agirre et al., 2019) shared tasks.

Considering the large precision and recall of most LivingNER systems, the LivingNER Spanish Silver Standard will be a

high-quality collection of annotated, normalized, and classified clinical documents in Spanish. Besides, it will serve to foster the development of species recognition and linking resources, as well as to generate more annotated data. The LivingNER Spanish Silver Standard will be released on the Zenodo Medical NLP community.

5 Discussion

There is a clear need to generate, extend and provide access to multilingual terminologies and glossaries for the biomedical domain. Providing access to bilingual medical glossaries such as MeSpEN, curated for species information and other clinical entities, might be helpful to foster exploitation for multilingual semantic annotation efforts (Villegas et al., 2018).

In this direction, the LivingNER initiative pioneers to structure the species information in clinical documents written in languages other than English. To foster the development of species NER and linking resources, we have released the LivingNER corpus: the first Gold Standard corpus of Spanish clinical documents with species mentions, manually mapped to the NCBI Taxonomy.

The LivingNER corpus was created following strict annotation guidelines that are made public to allow the corpus extension and adaptation to other languages or domains. It contains HUMAN annotations (a building block to collect relevant information from patient history, hereditary diseases, etc.) and SPECIES annotations. The latter is essential for diverse clinical applications, such as epidemiology.

To enhance the interoperability between different data sources, and taking into account (1) multilingual scenarios, (2) the multilingual potential of species mentions, and (3) the general lack of annotated data in other languages, we have released the LivingNER Multilingual Corpus. It contains the LivingNER corpus documents, translated to 7 languages (English, French, Italian, Portuguese, Catalan, Romanian, and Galician), and automatically generated species mention annotations mapped to NCBI Taxonomy.

The resources and the task have generated considerable interest in the community. Participant teams have developed 62 competitive systems based on pre-trained transformer language models evaluated against

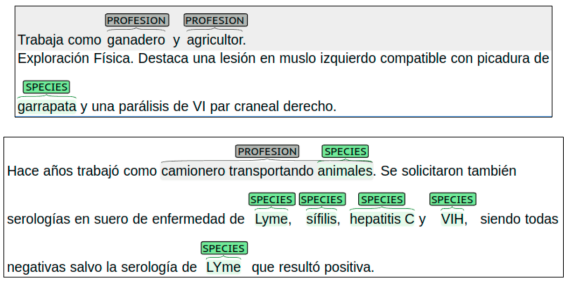


Figure 6: Actual examples of annotated species mentions and automatically recognized profession mentions.

the LivingNER corpus manual annotations. Additionally, they have generated automatic predictions for nearly 13,000 documents that will be harmonized to create the LivingNER Spanish Silver Standard.

These resources can be used to obtain actionable information from clinical narratives. An example would be linking the species with the text's occupational information to fine-tune the work-related disease statistics. This linking is seen in Figure 6, in which Gold Standard SPECIES annotations are combined with an automatic system that recognizes profession mentions (trained with MEDDOPROF (Lima-López et al., 2021) corpus).

As future directions, we plan to generate more granular annotations for the HUMAN mentions that are needed for real-world applications. In addition, the third subtask on Clinical Impact applications lacked enough training and test data, and we plan to correct this issue in the future. Finally, the Multilingual Silver Standard will be manually reviewed to generate manually-generated parallel annotations in eight languages.

Acknowledgements

We acknowledge the Encargo of Plan TL (SEDIA) to BSC for funding and the scientific committee for their guidance and help. Due to the relevance of species for biomaterials and implants, this project is supported by the European Union's Horizon Europe Coordination & Support Action under Grant Agreement No 101058779. We acknowledge the support from the AI4PROFHEALTH project (PID2020-119266RA-I00). We thank the organization of IberLEF and SEPLN, and Bitac for collaboration during the corpus and guidelines construction.

References

Armengol-Estapé, J., F. Soares, M. Marimon, and M. Krallinger. 2019. Pharmaconer tagger: a deep learning-based tool for automatically finding chemicals and drugs in spanish medical texts. *Genomics & informatics*, 17(2).

Avram, A.-M., M. Mitrofan, and V. Pais. 2022. Species entity recognition using a neural inhibitory mechanism.

Bagherzadeh, P., H. Verma, and S. Bergler. 2022. Multi-input rim for named-entity recognition in spanish clinical reports.

Carrino, C. P., J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, and M. Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *arXiv preprint arXiv:2109.03570*.

Chapman, K., M. Schwarz, and B. Häussler. 2022. Multilingual medical entity recognition and cross-lingual zero-shot linking with faiss.

Chizhikova, M., J. Collado-Montañez, P. López-Úbeda, M. C. Díaz-Galiano, L. A. Ureña-López, and M. T. Martín-Valdivia. 2022. Sinai at livingner shared task 2022: Species mention recognition and normalization using transfer learning and string matching techniques.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Cossin, S., G. Diallo, and V. Jouhet. 2022. Iam at iberlef 2022: Ner of species mentions.

del Moral, R., J. Reyes-Aguillón, O. Ramos-Flores, H. Gómez-Adorno, and G. Bel-Enguix. 2022. Species mention entity recognition, linking and classification using roberta in combination with spanish medical embeddings.

Federhen, S. 2012. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143.

Francis, S. and M.-F. Moens. 2022. Task-aware contrastive pre-training for spanish named entity recognition in livingner challenge.

Gasco, L., A. Nentidis, A. Krithara, D. Estrada-Zavala, R. T. Murasaki, E. Primo-Peña, C. Bojo Canales, G. Paliouras, M. Krallinger, et al. 2021. Overview of bioasq 2021-mesinesp track. evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials. CEUR Workshop Proceedings.

Gerner, M., G. Nenadic, and C. M. Bergman. 2010. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):1–17.

Gonzalez-Agirre, A., M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, and M. Krallinger. 2019. Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10.

Han, S. and H. Ding. 2022. Named entity recognition for livingner-species based on bert and span detection.

Hanjie, M. and Z. Xiaobing. 2022. Clinical text entity recognition based on pretrained model and bigru-crf.

IAM. 2022. Iamsystem. https://github.com/scossin/IAMsystem.

Jimeno Yepes, A. and K. Verspoor. 2022. The read-biomed team in livingner task 1 (2022): Adaptation of an english annotation system to spanish.

Kocaman, V., G. Pirge, B. Polat, and D. Talby. 2022. Biomedical named entity recognition in eight languages with zero code changes.

Krallinger, M., F. Leitner, and A. Valencia. 2010. Analysis of biological processes and diseases using text mining approaches. *Bioinformatics Methods in Clinical Research*, pages 341–382.

Lima-López, S., E. Farré-Maduell, A. Miranda-Escalada, V. Brivá-Iglesias, and M. Krallinger. 2021. Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classification and normalization of

professions and occupations from medical texts. *Procesamiento del Lenguaje Natural*, 67:243–256.

Marimon, M., A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodriguez, J. L. Martin, M. Villegas, and M. Krallinger. 2019. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@ SEPLN*, pages 618–638.

Mark. 2022. 33da. https://github.com/33Da/.

Miranda-Escalada, A., E. Farré, and M. Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. *IberLEF@ SEPLN*, pages 303–323.

Miranda-Escalada, A., E. Farré-Maduell, S. Lima-López, L. Gascó, V. Briva-Iglesias, M. Agüero-Torales, and M. Krallinger. 2021. The profner shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, pages 13–20.

Miranda-Escalada, A., L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, and M. Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources.

Miranda-Escalada, A., A. Gonzalez-Agirre, J. Armengol-Estapé, and M. Krallinger. 2020. Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. In *CLEF (Working Notes)*.

NLP-CIC-WFU. 2022. Nlp-cic-wfu- contribution-to-livingner-shared-task-2022. https://github.com/ajtamayoh/NLP-CIC-WFU- Contribution-to-LivingNER-shared-task-2022.

Pafilis, E., S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen. 2013. The species and organisms resources for fast and accurate identification of taxonomic names in text. *PloS one*, 8(6):e65390.

plncmm. 2022. Livingner. https://github.com/maranedah/LivingNER.

Pyysalo, S., T. Ohta, R. Rak, D. Sullivan, C. Mao, C. Wang, B. Sobral, J. Tsujii, and S. Ananiadou. 2011. Overview of the infectious diseases (id) task of bionlp shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 26–35.

Rebholz-Schuhmann, D., A. J. J. Yepes, E. M. Van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, E. Beisswanger, and U. Hahn. 2010. Calbc silver standard corpus. *Journal of bioinformatics and computational biology*, 8(01):163–179.

Rojas, M., J. Barros, M. Araneda, and J. Dunstan. 2022. Flert-matcher: A two-step approach for clinical named entity recognition and normalization.

Schoch, C. L., S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, et al. 2020. Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020.

Soares, F., M. Villegas, A. Gonzalez-Agirre, M. Krallinger, and J. Armengol-Estapé. 2019. Medical word embeddings for Spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 124–133, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Tamayo, A., D. A. Burgos, and A. Gelbukh. 2022. Partner: Paragraph tuning for named entity recognition on clinical cases in spanish using mbert + rules.

tutorial. 2022. ner. https://github.com/songhan123123/ner.

Vaca, A. 2022. Named entity recognition for humans and species with domain-specific and domain-adapted transformer models.

Villegas, M., A. Intxaurrondo, A. Gonzalez-Agirre, M. Marimon, and M. Krallinger. 2018. The mespen resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. *LREC MultilingualBIO: multilingual biomedical text processing*.

Zhu, Z. and L. Wang. 2022. Bert-bilstm model for entity recognition in clinical text.

Zotova, E., A. García-Pablos, N. Perez, P. Turón, and M. Cuadros. 2022. Vicomtech at livingner 2022.

zzz. 2022. 2251821381. https://github.com/2251821381.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 255-263 recibido 05-07-2022 revisado 18-07-2022 aceptado 21-07-2022

Overview of PAR-MEX at Iberlef 2022: Paraphrase Detection in Spanish Shared Task

Resumen de PAR-MEX en IberLEF 2022: Tarea Compartida para la Detección de Paráfrasis en Español

Gemma Bel-Enguix[1], **Gerardo Sierra**[1], **Helena Gómez-Adorno**[2], **Juan-Manuel Torres-Moreno**[3], **Jesus-German Ortiz-Barajas**[4], **Juan Vásquez**[4]
[1] Instituto de Ingeniería (UNAM)
[2] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (UNAM)
[3] Laboratoire Informatique d'Avignon (Avignon Université)
[4] Posgrado en Ciencia e Ingeniería de la Computación
{gbele,gsierram}@iingen.unam.mx, helena.gomez@iimas.unam.mx,
juan-manuel.torres@univ-avignon.fr, {jgermanob,juanmv}@comunidad.unam.mx

Abstract: Paraphrase detection is an important unresolved task in natural language processing; especially in the Spanish language. In order to address this issue, and contribute to the creation of high-performance paraphrase detection automated systems, we propose a shared task called PAR-MEX. For this task, we created a corpus, in Spanish, with topics in the domain of Mexican gastronomy. Afterwards, the participants in this task submitted their classification results on our corpus. In this paper we explain the steps followed for the creation of the corpus, we summarize the results obtained by the various participants, and propose some conclusions regarding the paraphrase-detection task in Spanish.
Keywords: PAR-MEX, paraphrase detection, Iberlef.

Resumen: La detección de paráfrasis es una tarea importante no resuelta en procesamiento del lenguaje natural; especialmente en la lengua española. Para atacar este problema, y para contribuir a la creación de sistemas de detección automática que obtengan resultados competitivos, proponemos la tarea compartida llamada PAR-MEX. Para esto, creamos un corpus en español con temas dentro del campo semántico de gastronomía mexicana. Después los participantes en esta tarea enviaron los resultados de sus sistemas de clasificación sobre nuestro corpus. En este paper explicamos los pasos seguidos para la creación del corpus, resumimos los resultados obtenidos por los participantes, y proponemos algunas conclusiones al respecto de la detección de paráfrasis en español.
Palabras clave: PAR-MEX, detección paráfrasis, Iberlef.

1 Introduction

Two texts, or two sentences, are paraphrase when they are semantically equivalent, regardless of the cause that led to that equivalence (Das and Smith, 2009). Detecting paraphrased text is a task that has aroused the interest of the Natural Language Processing (NLP) community, due to the fact that it has multiple applications, such as plagiarism detection, question-answering and machine translation (Kong et al., 2020).

Paraphrase construction includes different mechanisms, such as lexical changes through synonymy, sentence rearrangement, breaking of a sentence into several parts, and joining more than one phrase into another. Therefore, addressing the problem of paraphrase detection requires an analysis that encompasses different levels, both lexical and semantic, as well as syntactic.

To deal with the problem of paraphrase detection using supervised machine learning methods, researchers use data sets that typically include pairs of sentences that are identified as paraphrase or non-paraphrase. There are various ways of elaborating or compiling these corpora: news collections, plagiarism pairs, manual creation, relational ac-

ISSN 1135-5948. DOI 10.26342/2022-69-22

Topic of the document	No. of lines
sushi	28
molecular cuisine	21
tequila	25
kebab	25
day of the dead	25
vegan food	25
street food	25

Table 1: Topic and number of lines in each of the seven original documents.

quisition, back-translation, multiple translations.

The PARMEX task has been organized for the first time at IberLeF 2022 (Montes-y-Gómez et al., 2002), a shared evaluation campaign for NLP systems in Spanish and other iberian languages, which is part of the SEPLN congress. The task is based on the Gastronomy Corpus, elaborated by the Language Engineering Group, which is divided into seven sub-corpora that deal with different topics related to cuisine, preferably, but not exclusively, Mexican. The corpus has been manually compiled in Mexico and, therefore, contains some terms and expressions specific to the Mexican variant of Spanish.

The rest of their paper is organised as follows. Section 2 presents the evaluation framework used at PARMEX 2022. Section 3 shows an overview of different approaches taken to tackle the problem. Section 4 reports and analyses the results obtained by the teams that have participated. Finally, Section 5 presents our conclusions from this shared task.

2 PARMEX 2022 Corpus and evaluation framework

For the PAR-MEX at Iberlef 2022 task, we created a corpus comprised of sentence pairs in Mexican Spanish. For the creation of the sentence pairs, first we produced seven original texts with gastronomical topics. Each one of these seven texts had a variable number of lines. On Table 1 the exact number of lines and topics per document are shown.

The second step in the creation of the corpus was the generation of the paraphrased documents. These new documents were created by humans who were tasked with writing one document with identical semantic content and same number of lines as in the original document. For example, for the document `sushi.txt`, an original document with 28 lines, seven paraphrased documents were created. The 28 lines in each one of these seven paraphrased documents contained the exact same meaning as the 28 lines in the original document.

Topic	No. of paraphrased documents
sushi	7
molecular cuisine	31
tequila	7
kebab	7
day of the dead	8
vegan food	6
street food	6
$\sum$	**72**

Table 2: Topics of the original seven documents, and their respective number of paraphrased documents.

The process described above was repeated for every one of the seven original documents. Then, we generated a total of 72 paraphrased documents. The exact numbers can be seen on Table 2.

The next step in the elaboration of the task's corpus was the creation of the sentence pairs, and their respective labels. For this, we paired each line in every original document with each line in every paraphrased document. If the sentence pair was made up of a line in an original document with an index of i, and one line in a paraphrased document with an index i, then it would be labeled as "paraphrase". In the opposite case, the one in which a sentence in the original document with index i was matched with a sentence from another document but with an index of j (given that $i \neq j$), then that sentence pair would be labeled as "not paraphrase". It is important to mention that even if the index of an original document and the index of a paraphrased document were equal, it was also verified that the line from the paraphrased document belonged to the same topic as the line from the original document. For example, if line i from document `vegan_food.txt` was paired with line i from a paraphrased document related to `tequila.txt`, this pair would not be labeled as paraphrase since their semantic contents would differ due to their topics even though their indices were

Topic	No. of high-level sentence-pairs
sushi	41
molecular cuisine	214
tequila	84
kebab	63
day of the dead	75
vegan food	42
street food	51
$\sum$	**750**

Table 3: Number of high-level paraphrase pairs per original document.

the same. Therefore, in order to obtain the paraphrase sentence-pairs, the topic and the indices were compared.

The final step in the creation of the corpus was the addition of the high-level paraphrase pairs. For this, we requested humans to write several original documents with high-level paraphrase. During this step, we did not ask them to write paraphrased documents with the same number of lines as the original documents. Once created these novel documents with high-level paraphrases, we extracted some lines and paired them with the sentences in the original documents. This process generated less paraphrase pairs than the initial step with low-level paraphrases, and the exact number of high-level paraphrase-pairs can be observed in Table 3.

After the pairing of the sentences, and the creation of their respective labels, a total of 10,298 sentence-pairs were obtained. From this set, 1,844 sentence-pairs were labeled as paraphrase, while the remaining 8,454 sentence-pairs were labeled as non-paraphrase. This represented an approximate of 20% of sentence-pairs labeled as paraphrase, with the remaining 80% labeled as not paraphrase. From this set, we created the training, validation and test partitions. The distribution of these sets is shown on Table 4.

3 Overview of the Submitted Approaches

In this edition, six teams submitted one or more solutions to the task through the codalab platform[1]. CodaLab Competitions is

[1]https://codalab.lisn.upsaclay.fr/competitions/2345

Partition	Total sentence-pairs	Paraphrase sentence-pairs
Training	7,382	1,282
Validation	97	20
Test	2,819	542
Total	10,298	1,844

Table 4: Number and distribution of sentence-pairs in the training, validation and evaluation sets.

a robust open-source framework for running competitions that involve results or code submission. The evaluation methodology of a competition in this platform consists of receiving as input the predictive outputs of systems. It returns a performance evaluation based on the metrics defined for each task.

This section presents a summary of the submitted systems in terms of preprocessing, feature extraction, and classification algorithms. In Table 5 we indicate the general approach used by each team. It can be appreciated that participants used two general approaches: transformers and traditional ML. Following this, we briefly describe each of the participants methods.

Approach	NLP-CIC-TAGE	Tü-Par	Thang CIC	Abu	FRSCIC	UC3M-DEEPNLP
Transformers	X		X			X
Traditional ML		X		X	X	

Table 5: General approach of each participating team.

- *Using Transformers on Noisy vs. Clean Data for Paraphrase Identification in Mexican Spanish* (Tamayo, Burgos, and Gelbukh, 2022)
  - **Team name: NLP-CIC-TAGE**
  - **Summary:** The participants presented a transfer learning approach using transformers to tackle paraphrase identification on noisy vs. clean data in Spanish. They used BERTIN, a pre-trained model on the Spanish portion of a massive

multilingual web corpus. The fine-tuning and parameter tunning of BERTIN was performed on noisy data and used to identify paraphrase on clean data.

- *PAR-MEX Shared Task Submission Description: Identifying Spanish Paraphrases Using Pretrained Models and Translations* (Girrbach, 2022)
  - **Team name: Tü-Par**
  - **Summary:** The participants proposed an approach based on a classical machine learning pipeline consisting of feature extraction, supervised learning, and evaluation. The feature extraction consists in encoding Spanish sentences (or their English translations) by a pretrained sentence encoder, then concatenating the sentence embeddings or representing the sentences by a similarity score. Different classifiers were used depending on the feature type—a logistic regression model and a random forest model on the similarity features, and multi-layer perceptrons on the sentence embeddings features.

- *GAN-BERT, an Adversarial Learning Architecture for Paraphrase Identification* (Ta et al., 2022)
  - **Team name: Thang CIC**
  - **Summary:** The participants used text embeddings from pre-trained transformer models for training by GAN-BERT, adversarial learning. They modified noises for the generator, which have a random rate and the exact size of the hidden layer of transformers. They also included a rule of thumb based on the pair similarity to remove possible wrong sentence pairs in positive examples and additional unlabelled data in the same domain to improve the model performance.

- *Paraphrase Identification: Lightweight effective methods based features from pre-trained models* (Rahman et al., 2022)
  - **Team name: Abu**
  - **Summary:** The participants introduced two lightweight methods: linear regression and multilayer perceptron, trained on six features: the difference in sentences' length, common lemmas between 2 sentences, sentences' similarity, etc. After performing Component Analysis (PCA) to reduce the dimension, they filter noises in the positive examples by introducing a rule of thumb on the pair similarity.

- *Mexican Spanish Paraphrase Identification using Data Augmentation* (Meque et al., 2022)
  - **Team name: FRSCIC**
  - **Summary:** The participants performed a data augmentation step on the training set using translation. The text vectorization process consisted of sentence transformers, spaCy vectors, traditional word n-grams, and bi-tri syntactic n-grams using TF-IDF. They proposed a similarity vector using three different similarity algorithms for the final representation: Jaccard, Cosine, and spaCy. For the classification step, they used a soft-voting ensemble model with three estimators.

- *UC3M at PAR-MEX@IberLef 2022: From Cosine Distance to Transformer Models for Paraphrase Identification in Mexican Spanish* (Brando-Le-Bihan, Karbushev, and Segura-Bedmar, 2022)
  - **Team name: UC3M-DEEPNLP**
  - **Summary:** The participants evaluated a baseline method based on the cosine similarity of two text pairs representation: TF-IDF model on bag-of-words and word embedding models provided by spaCy. For the final submission, they used the "bert-base-cased-finetuned-mrpc" model, which is

fine-tuned for paraphrase detection by using the MRPC corpus. They also proposed strategies such as class balancing or data augmentation to improve the generalization capability. However, they did not present these strategies in the final submission.

4 *Experimental Evaluation and Analysis of the Results*

This section reviews the results obtained by the participants of PAR-MEX at Iberlef 2022: Paraphrase Detection in Spanish Shared Task. For this purpose, we analyse and compare the submitted solutions' performance on the test partition. We used the F1-score metric on the paraphrase (P) as the primary performance measure and to rank all the participants. We launched a Codalab competition to manage the shared task stages and compute the performance metric for all submissions.

We propose a transformers-based approach as a baseline. It consists of the Bidirectional Encoder Representation from Transformer (BERT) model (Devlin et al., 2019). We use the base model for our baseline, consisting of twelve Transformer blocks and the pre-trained model BETO (Cañete et al., 2020), a BERT model trained on an enormous Spanish corpus. We use four epochs and the Adam optimizer for the fine-tuning stage with a learning rate of 2e-5. We use the HuggingFace implementation (Wolf et al., 2020) for Tensorflow (Abadi et al., 2015). In order to have comparable results with the participant submissions, we report the best result in five runs using different random seeds.

Table 6 summarises the results obtained by each team and our baseline in the PAR-MEX shared task. We report the F1 score in both Paraphrase and Non-paraphrase classes, the macro F1 score, and the accuracy. In this edition of the PAR-MEX shared task, the approach submitted by the NLP-CIC-TAGE team outperformed all the other approaches and the baseline. The NLP-CIC-TAGE team used an approach based on a transformer architecture; they fine-tuned the RoBERTa model pre-trained in a Spanish corpus. In contrast, the second-best approach proposed by the Tü-Par team used a Random Forest classifier using similarity-based features. These results show that classic approaches are still competitive for this task compared to deep learning.

We use the Maximum Possible Accuracy (MPA) and Coincident Failure Diversity (CFD) metrics (Tang, Suganthan, and Yao, 2006) to analyse the complementariness and the diversity of the predictions of the submitted approaches. The MPA is analogous to accuracy, defined as the correct classified instances divided into the total number of instances. To consider an instance correctly classified, at least one of the teams needs to assign the correct label to it. Using the MPA metric, we can detect the misclassified instances by all teams. The CFD metric has a minimum value of 0 when all classifiers are always correct or when all classifiers are either correct or wrong. On the other hand, it has a maximum value of 1 when at most one classifier will fail on any randomly chosen instance (Kuncheva and Whitaker, 2003). The CFD is defined in equation 1.

$$CFD = \begin{cases} 0, & p_0 = 1.0; \\ \frac{1}{1-p_0} \sum_{i=1}^{L} \frac{L-i}{L-1} p_i & p_0 < 1 \end{cases} \quad (1)$$

Table 7 shows the results of these metrics by grouping the proposed approaches based on their similar features. We create four groups: all teams, all teams who send their paper, Transformers-based approaches, traditional-machine-learning-based approaches. All of the groups mentioned above have at least two members. All participants sent their papers but one. Transformers-based approaches include the following teams: NLP-CIC-TAGE, Thang CIC, and UC3M-DEEPNLP. Tü-Par, FRSCIC, and ThangCIC conform traditional machine learning based approaches group.

In terms of the general approach, traditional machine learning performs better in terms of MPA than Transformers-based solutions. The above suggests that the different features used to train these machine learning models complement each other. In the same way, the combination of transformers and machine learning approaches obtain the highest MPA performance and have an average increment of 0.66% compared to those

Team	F1-score (P)	F1-score (NP)	Accuracy	Macro F1-score
NLP-CIC-TAGE	0.9424	0.9869	0.9787	0.9647
Tü-Par	0.9373	0.9853	0.9762	0.9613
Thang CIC	0.9022	0.9775	0.9635	0.9399
Abu	0.8867	0.9751	0.9592	0.9309
FRSCIC	0.8754	0.9730	0.9557	0.9242
UC3M-DEEPNLP	0.8450	0.9679	0.9468	0.9065
temu_bsc	0.8441	0.9567	0.9322	0.9004
baseline	0.834936	0.953075	0.926924	0.894006

Table 6: Result summary for the PAR-MEX shared task on the test set.

Approach	Best accuracy	MPA	CFD	Number of systems
All teams	0.9787	0.9936	0.0408	7
all teams (with submission)	0.9787	0.9915	0.0341	6
Transformers	0.9787	0.9847	0.0331	3
Traditional ML	0.9762	0.9883	0.0373	3

Table 7: MPA and CFD comparison results among the different proposed approaches.

individual approaches. Finally, the values for the CFD score are comparable among all approaches, which means that their predictions are complementary to an extent; this leads us to conclude that traditional and transformer-based approaches learn different information from text pairs.

Table 8 shows the results of the F1 score for the paraphrase class divided by topic in the test set. The kebab category achieved the highest performance with an average F1 score of 0.9483; on the other hand, the sushi topic had the worst performance with an average F1 score of 0.7659. The NLP-CIC-TAGE team obtained the best performance in two of the seven topics. In contrast, the Tü-Par team obtained the best performance in four topics, including the sushi, which is the hardest. Nevertheless, the difference was in the food truck topic. The NLP-CIC-TAGE obtained a 0.9153 F1 score, while the Tü-Par team obtained 0.8673. For this result, the NLP-CIC-TAGE achieved first place in the PAR-MEX shared task.

Tables 9 and 10 show the performance of each team by topic and low-level paraphrase and high-level paraphrase, respectively. In order to compute these metrics, we filtered the paraphrase examples and kept the non-paraphrase examples unchanged. Only day of the dead, vegan food, and food truck topics have examples of high-level paraphrase. Regarding high-level paraphrase, the food truck topic obtains the highest performance while the sushi topic obtains the lowest; however, the sushi topic only has one example of this type of paraphrase. When comparing high-level and low-level paraphrase performance, only the food truck topic performs better on high-level paraphrase than on low-level paraphrase. These results suggest that, in general, detecting high-level paraphrase examples is more challenging for the proposed approaches. The most substantial difference is in the vegan food topic; the average result in high-level paraphrases is 0.6509, while in low-level paraphrases is 0.9095, which means a 0.2586 between both levels. This topic has 41 high-level paraphrase examples and 36 low-level paraphrase examples; because the examples of this topic are nearly balanced, we can conclude that the performance difference is due to the difficulty of identifying high-level paraphrase features.

In terms of proposed approaches, Transformers-based models outperform all teams in two of the three topics with high-level paraphrase examples; in the remaining topic, Transformers-based and traditional machine learning approaches have the same performance. Therefore, we can conclude that Transformers can learn better features to identify high-level paraphrases. On the other hand, when dealing with low-level paraphrases, a traditional machine learning approach outperform all teams in 4 of 7 topics. A Transformers-based approach has the highest performance in the remaining

Team	Molecular cusine	Day of the dead	Kebab	Tequila	Vegan food	Sushi	Food truck
NLP-CIC-TAGE	0.9878	0.9714	0.9792	0.9231	0.8261	0.8333	0.9153
Tü-Par	0.9762	0.9859	0.98	0.9362	0.8252	0.8772	0.8673
Thang CIC	0.9687	0.8400	0.9216	0.9091	0.8444	0.7692	0.8468
Abu	0.9495	0.9489	0.9574	0.8864	0.8000	0.6567	0.7573
FRSCIC	0.9254	0.8806	0.9293	0.8764	0.7852	0.8077	0.7810
UC3M-DEEPNLP	0.9010	0.8000	0.9462	0.8989	0.7576	0.6818	0.7358
temu_bsc	0.8460	0.8675	0.9245	0.7132	0.8591	0.7353	0.9167
Baseline	0.7871	0.9863	0.8596	0.7833	0.8387	0.7692	0.918
Average	0.9177	0.9096	0.9372	0.8658	0.8181	0.7654	0.8412

Table 8: Results for the PAR-MEX shared task on the test set by topic.

Team	Molecular cusine	Day of the dead	Kebab	Tequila	Vegan food	Sushi	Food truck
NLP-CIC-TAGE	0.9878	0.9636	0.9792	0.9231	0.9333	0.8511	0.9189
Tü-Par	0.9762	1	0.98	0.9362	0.9114	0.8727	0.8406
Thang CIC	0.9687	0.8099	0.9216	0.9091	0.9577	0.7843	0.806
Abu	0.9495	0.9541	0.9574	0.8864	0.9429	0.6667	0.6885
FRSCIC	0.9254	0.8571	0.9293	0.8764	0.8919	0.8	0.7302
UC3M-DEEPNLP	0.901	0.7473	0.9462	0.8989	0.8919	0.6977	0.6769
temu_bsc	0.846	0.8382	0.9245	0.7132	0.9	0.7273	0.9067
Baseline	0.7871	0.9828	0.8596	0.7833	0.8471	0.7619	0.9091
Average	0.9177	0.8941	0.9372	0.8658	0.9095	0.7702	0.8096

Table 9: Results for the PAR-MEX shared task on the test set by topic and low-level paraphrases.

three topics. With these results, we can conclude that machine learning models can handle low-level paraphrasing better than complex models like transformers when using similarity-based features as the primary type of characteristics.

Finally, Table 11 shows each team's performance only on the paraphrase type. Again, the results are consistent with what we show in tables 7 and 8. Although the NLP-CIC-TAGE team does not obtain the best result in every topic in the test set, their overall performance is the best on both levels of paraphrasing.

5 Conclusions

This paper described the design and results of the PAR-MEX shared task collocated with IberLef 2022. PAR-Mex is focused in paraphrase identification in Mexican Spanish texts. This has been the first edition of the task.

The data set of PAR-MEX included both, low-level and high-level pairs of paraphrases, although they were not distinguished for the participants. The analysis of the results shows that, whereas low-level paraphrase is currently an easy task for natural language processing (0.90 of average), high-level paraphrase is a problem that has not been conveniently approached yet.

The best results in this shared task were obtained by a team that proposed to approach the problem with a method based on transformers. However, traditional machine learning strategies obtained very similar results. Indeed, while deep learning techniques have the best scores in the sub-corpora of molecular cuisine, vegan food sushi and food truck, traditional methods lead in day of the dead, kebab and tequila. The only topic in which transformers reach a clearly better score is food truck. This shows this is a complex task and that collaboration between models and the use of multiple variables can improve the final outcome of the research.

Acknowledgments

We acknowledge the support of the projects CONACyT CB A1-S-27780, and DGAPA-UNAM PAPIIT references TA400121 and TA101722. The authors thank CONACYT for the computing resources provided through the Deep Learning Platform for Lan-

Team	Day of the dead	Vegan food	Sushi	Food truck
NLP-CIC-TAGE	1	0.6567	0	0.9091
Tü-Par	0.9286	0.6479	0.2222	0.9091
Thang CIC	0.6364	0.7077	0	0.9091
Abu	0.9286	0.623	0	0.8571
FRSCIC	0.875	0.6061	0.25	0.8571
UC3M-DEEPNLP	0.9655	0.5397	0	0.7727
temu_bsc	0.5769	0.7273	0.1	0.913
Baseline	0.9375	0.6988	0.1176	0.8936
Average	0.8561	0.6509	0.0862	0.8776

Table 10: Results for the PAR-MEX shared task on the test set by topic and high-level paraphrases. Molecular cuisine, kebab and tequila do not have high-level paraphrase examples.

Team	F1-score high-level paraphrase	F1-score low-level paraphrase
NLP-CIC-TAGE	0.7755	0.9602
Tü-Par	0.6951	0.9538
Thang CIC	0.6552	0.9137
Abu	0.6494	0.905
FRSCIC	0.6795	0.8884
UC3M-DEEPNLP	0.6575	0.8605
temu_bsc	0.4167	0.8378
Baseline	0.3976	0.8265
Average	0.6158	0.8927

Table 11: Results for the PAR-MEX shared task on the test set by paraphrase type.

guage Technologies of the INAOE Supercomputing Laboratory.

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Brando-Le-Bihan, A., R. Karbushev, and I. Segura-Bedmar. 2022. UC3M at PAR-MEX@IberLef 2022: from cosine distance to transformer models for paraphrase identification in mexican spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Das, D. and N. A. Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 468–476, Suntec, Singapore, August. Association for Computational Linguistics.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Girrbach, L. 2022. PAR-MEX shared task submission description: Identifying spanish paraphrases using pretrained models and translations. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*.

Kong, L., Z. Han, Y. Han, and H. Qi. 2020. A deep paraphrase identification model interacting semantics with syntax. *Complexity*, 2020:14 pages.

Kuncheva, L. and C. Whitaker. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207, 05.

Meque, A., F. Balouchzahi, G. Sidorov, and A. Gelbukh. 2022. Mexican spanish paraphrase identification using data augmentation. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*.

Montes-y-Gómez, M., J. Gonzalo, F. Rangel, M. Casavantes, M. Álvarez-Carmona, G. Bel-Enguix, H. Escalante, L. Freitas, A. Miranda-Escalada, F. Rodríguez-Sánchez, A. Rosá, M. Sobrevilla-Cabezudo, M. Taulé, and R. Valencia-García. 2002. *Proceedings of IberLeF 2002*.

Rahman, A., H. Ta, L. Najjar, and A. Gelbukh. 2022. Paraphrase identification: Lightweight effective methods based features from pre-trained models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*.

Ta, H., A. Rahman, L. Najjar, and A. Gelbukh. 2022. GAN-BERT, an adversarial learning architecture for paraphrase identification. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*.

Tamayo, A., D. A. Burgos, and A. Gelbukh. 2022. Using transformers on noisy vs. clean data for paraphrase identification in mexican spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*.

Tang, E. K., P. N. Suganthan, and X. Yao. 2006. An analysis of diversity measures. *Machine learning*, 65(1):247–271.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 265-272 recibido 05-07-2022 revisado 18-07-2022 aceptado 21-07-2022

Overview of PoliticEs 2022: Spanish Author Profiling for Political Ideology

Resumen de la tarea PoliticEs 2022: Perfilado del Autor Español por su Ideología Política

José Antonio García-Díaz[1], **Salud María Jiménez-Zafra**[2], **María-Teresa Martín Valdivia**[2], **Francisco García-Sánchez**[1], **L. Alfonso Ureña-López**[2], **Rafael Valencia-García**[1]
[1]Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain
[2]Computer Science Department, SINAI, CEATIC, Universidad de Jaén, 23071, Spain
{joseantonio.garcia8,frgarcia,valencia}@um.es
{sjzafra,maite,laurena}@ujaen.es

Abstract: This paper presents the PoliticEs 2022 shared task, organized at IberLEF 2022 workshop, within the framework of the 38th International Conference of the Spanish Society for Natural Language Processing. This task aims to extract the political ideology from a given user's set of tweets. Specifically, it focused on the identification of the gender and the profession, as demographic traits, and the political ideology from a binary and multi-class perspective, as a psychographic trait. The PoliticEs task attracted 63 teams that registered through CodaLab. Finally, 20 submitted results and 14 presented working notes describing their systems. Most of the teams proposed transformer-based approaches, although some of them also used traditional machine learning algorithms or even a combination of both approaches.
Keywords: Author profiling, political ideology, author analysis, demographic and psychographic traits.

Resumen: Este artículo presenta la tarea PoliticEs 2022, organizada en el taller IberLEF 2022, en el marco de la 38 edición del Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural. Esta tarea tiene como objetivo extraer la ideología política de un usuario a partir de un conjunto de tuits publicados por él. En concreto, se centró en la identificación del género y la profesión, como rasgos demográficos, y la ideología política desde una perspectiva binaria y multiclase, como rasgo psicográfico. La tarea PoliticEs atrajo a 63 equipos que se inscribieron a través de CodaLab. Finalmente, 20 enviaron resultados y 14 presentaron artículos describiendo sus sistemas. La mayoría de los equipos propusieron enfoques basados en transformers, aunque algunos de ellos también utilizaron algoritmos tradicionales de aprendizaje automático o incluso una combinación de ambos enfoques.
Palabras clave: Perfilado de usuarios, ideología política, análisis de autores, rasgos demográficos y psicográficos.

1 Introduction

Political ideology is a psychographic trait that can be used to understand individual and social behaviour, including moral and ethical values as well as inherent attitudes, appraisals, biases, and prejudices (Verhulst, Eaves, and Hatemi, 2012). The relationship between personality traits and political ideology was demonstrated in Fatke (2017). The author gathered data from 21 countries and found a correlation between political ideology and the big five personality traits. For instance, he found that conscientiousness is strongly correlated with the right wing, whereas openness to experience and agreeability were notably more correlated to the left wing. Moreover, our political ideology has a great influence in our daily lives. For example, Baumgaertner, Carlisle, and Justwan (2018) found a correlation between political ideology and the attitude of citizens to vaccination campaigns of infectious diseases.

The PoliticEs shared task organized at IberlEF 2022 (Montes-y Gómez et al., 2022)

ISSN 1135-5948. DOI 10.26342/2022-69-23

aims to extract political ideology information from texts. For this, an author profiling task is proposed. It is focused on the identification of the gender, the profession, and the political spectrum from a binary and multi-class perspective.

In recent years, several shared tasks have been organized on author analysis under the PAN workshop series (Bevendorff et al., 2021). The novelty of the PoliticEs task is that, to the best of our knowledge, none of these previous tasks have focused on political ideology.

The rest of the paper is organized as follows. Section 2 describes the PoliticEs shared task. Section 3 presents the dataset provided in the competition. Section 4 summarized the participant approaches. Section 5 shows the results and a discussion thereof. Finally, Section 6 concludes the paper with some insights and future works.

2 Task description

The PoliticEs shared task consists of extracting the gender and the profession as demographic traits, and the political ideology as a psychographic trait from a given user's set of tweets. Political ideology is considered as a binary (pib) and as a multiclass problem (pim). The possible categories of each trait are as follows:

- gender: male, female.
- profession: political, journalist.
- pib: left, right.
- pim: left, moderate left, right, moderate right.

The challenges involved in this shared task are:

1. Extracting political ideology from a text collection. To the best of our knowledge, this is the first Spanish shared task focused on this.
2. Multi-class classification. The author profiling task should be addressed from a binary and multi-class perspective with four different classes.

The competition was organized through CodaLab and is accessible at the following link: `https://codalab.lisn.upsaclay.fr/competitions/1948`. It was divided into 3 phases: Practice, Evaluation and Post-evaluation. In the Practice phase, the participants were provided with a subset of the training data to familiarize with the training data format, and with a notebook with a baseline based on Bag of Words (BoW) to have a starting point for system development. Later, they were provided with the full training set to develop their approaches. For this, they were allowed to make a maximum of 100 submissions in CodaLab. It should be mention that in the Evaluation phase, the test partition was provided for the participants to label it using the developed systems. This partition was used to evaluate the teams. They were allowed to make a maximum of 10 submissions through CodaLab, from which each team had to select the best one for ranking. The ranking was determined using the arithmetic mean of the macro f1-score of the gender, profession, binary political ideology, and multi-class political ideology.

3 Dataset

The dataset for this shared task is an extension of the Spanish PoliCorpus 2022 (García-Díaz, Colomo-Palacios, and Valencia-García, 2022), which consists of a set of tweets from the timelines of the Twitter accounts of politicians and journalists in Spain. The politicians are members of the government, congress and senate of Spain along with mayors, presidents of the autonomous communities, former politicians, and collaborators whereas the journalist accounts belong to journalists associated to political press, from Spanish newspapers such as ABC, El País, ElDiario, El Mundo or La Razón among others. The dataset was compiled using the UMUCorpusClassifier tool (García-Díaz et al., 2020).

The users of the dataset are labelled with their gender (male, female), profession (political, journalist), and their political spectrum on a binary axis (left, right) and a multiclass axis (left, moderate left, moderate right, right).

Regarding the tweets collected from each user, we discarded retweets and tweets that contains headlines from news sites. We also removed tweets written in languages other than Spanish. Moreover, we anonymised them by replacing all mentions with the token @user, except for the real users, that were encoded with the token @user and a correlative

number. We did this to hinder the author's traits identification.

The final dataset is composed of around 400 different users with at least 120 tweets. For the shared task, training and test sets were released (80%-20%). We released the dataset in two splits: training and testing. However, in the first stages of the competition, we released an early birds dataset composed by a subset of 5,000 tweets from the training dataset. It is worth noting that the accounts from training and testing are completely independent in order to prevent automatic classifiers learn to identify the authors rather than the traits. The number of users per set and trait are shown in Table 1.

4 Participant approaches

The PoliticEs shared task attracted 63 teams that registered in CodaLab, of which 20 submitted results and 14 presented working notes describing their systems. The following is a brief summary of the participants' proposals:

- **(1st) LosCalis (Carrasco and Rosillo, 2022).** This system is based on transformers (Vaswani et al., 2017). It combines BETO (Cañete et al., 2020) and MarIA (Gutiérrez Fandiño et al., 2022), and employs both architectures for document level characteristics extraction together with a Multi-Layer Perceptron for labels decoding.
- **(2nd) NLP-CIMAT (Villa-Cueva et al., 2022).** The authors propose PolitiBETO, a pretrained BETO model (Cañete et al., 2020) in the political domain, based on the use of domain adaptation and ensemble learning. They compose an ensemble using several instances of pretrained adapted BETO models, which predicts the test data at a tweet level. These predictions are then merged through a majority vote to determine the labels of a given author based on their tweets.
- **(3rd) Alejando Mosquera (Mosquera, 2022).** He explores the use of L2-regularized logistic regression model based on word and character n-grams features along with readability features. This work is notable for the analysis of adversarial attacks on the author profiling challenge.
- **(4th) CIMAT 2021 (Santibáñez-Cortés et al., 2022).** This team defines different classification models per each trait. Specifically, they use fine-tuned BERT (Kenton and Toutanova, 2019) models for the gender and profession, XGBoost for binary ideology, and Logistic Regression for multiclass ideology.
- **(5th) HalBERT (Holgado and Sinha, 2022).** The authors evaluate multiple feature sets, and deep learning and machine learning models. They also explore data augmentation and ensemble learning techniques. They find that GloVE embedding features and term-frequency based features, like TF-IDF, can be very helpful and can provide comparative results to deep learning approaches.
- **(7th) I2C (Ramos et al., 2022).** Their proposal is based on the used of transformers (Vaswani et al., 2017). For gender extraction, they build an ensemble as a set of pre-trained transformers models (RoBERTa (Liu et al., 2019), ALBERTI[1] and BERTIN (De la Rosa et al., 2022)). For the identification of the profession, the tweets of each user are merged to optimize the models. Finally, for the binary and multi-class classification of political ideology, the ROBERTA model was fine-tuned.
- **(8th) TeamMX (Ochoa-Hernández and Alemán, 2022).** The authors analyze several methods for feature selection and machine learning (Random Forest and SVM) and deep learning (Multi-Layer Perceptron) classifiers. Finally, for determining the gender, they select the best 200 Pearson's correlation words, using TF-IDF and SVM classifier. For the identification of the profession, they use transition point analysis with lemmas using TF-IDF and Random Forest classifier. For the binary classification of the ideology, they select the set based study using TF-IDF and SVM classifier. For the multi-class classification of the ideology they use an average analysis with lemmas using frequency and Multi-Layer Perceptron classifier.

[1] https://huggingface.co/flax-community/alberti-bert-base-multilingual-cased

Trait		Training	Test	Total
Gender	***Male***	177	69	**246**
	Female	136	36	**172**
Profession	***Politician***	251	80	**331**
	Journalist	61	26	**87**
Binary ideology	***Left***	178	57	**235**
	Right	135	48	**183**
Multiclass ideology	***Moderate left***	102	36	**138**
	Left	76	21	**97**
	Moderate right	94	31	**125**
	Right	41	17	**58**

Table 1: Corpus statistics per trait.

- **(9th) UniRetro (Manea and Dinu, 2022).** The authors propose two approaches: the first one based on using TF-IDF on SentencePiece pretrained and custom tokens obtained by Named Entity Encapsulation, and the second one consisting of fine-tuning BETO (Cañete et al., 2020) and DistilBETO (Cañete et al., 2022).
- **(13th) UNED (Rodrigo, Fabregat, and Centeno, 2022).** This team explores two approaches. The first is based on approximate nearest neighbours, which obtains low scores for individual results but a great score when combining several outputs. The second uses some fine-tuned BERT systems, obtaining the best results.
- **(14th) THANGCIC (Ta et al., 2022).** They present a system based on multilingual BERT (Kenton and Toutanova, 2019), fine-tuned for sentiment analysis, which has been trained with product reviews written in different languages.
- **(15th) URJC-Team (Rodríguez-García, Montalvo Herranz, and Martínez Unanue, 2022).** This team explores two machine learning algorithms (Logistic Regression and SVM) using a pre-processing module that cleans the tweets and a feature extractor module that combines character and word features with two different settings, with and without stopwords. Finally, they select SVM classifier with stopwords, which provides the best results.
- **(16th) SINAI (Espin-Riofrio, Ortiz-Zambrano, and Montejo-Ráez, 2022).** The authors propose a voting classifier model that leverages the use of several classical classifiers (Logistic Regression, Random Forest, Decision Trees, Multi-Layer Perceptron, and Gradient Tree Boosting) using as features the combination of stylometry measures with embeddings obtained from MarIA (Gutiérrez Fandiño et al., 2022), a Spanish RoBERTa model for text representation.
- **(19th) UC3MDeep (García-Ochoa Martín-Forero, Massotti López, and Segura-Bedmar, 2022).** The authors explore several machine learning approaches (K-Nearest Neighbours, Random Forest and Logistic Regression) with different configurations. They obtain the best scores using Logistic Regression without penalty and with a saga solver.
- **INFOTEC-LaBD (Cabrera, Tellez, and Miranda, 2022).** The proposal of these authors is based on a low-dimensional stacking model approach, which was designed to create both transparent and competitive user profiling models. The results of this team were late in the challenge, due to a confusion with the deadline.

5 Results and discussion

The official leaderboard of the PoliticEs shared task is shown in Table 2. It can be seen the results of the 19 participants that submitted results in time, plus the results of the baseline provided as a notebook, plus the results of the INFOTEC-LaBD team, which submitted a few hours late due to a mistake.

Team	Average Macro-F1	F1-gender	F1-profession	F1-ideology (binary)	F1-ideology (m-class)
LosCalis	0.90226 (01)	0.90287 (01)	0.94433 (01)	0.96162 (01)	0.80023 (04)
NLP-CIMAT	0.89096 (02)	0.78484 (06)	0.92125 (03)	0.96148 (02)	0.89628 (01)
Alejandro Mosquera	0.88918 (03)	0.82671 (03)	0.93345 (02)	0.95152 (03)	0.84504 (03)
CIMAT_2021	0.87976 (04)	0.83683 (02)	0.89500 (05)	0.94167 (04)	0.84553 (02)
HalBERT	0.82532 (05)	0.72602 (13)	0.89776 (04)	0.92176 (05)	0.75574 (06)
Bernardo	0.81996 (06)	0.79178 (04)	0.84982 (08)	0.91315 (06)	0.72511 (08)
I2C	0.79998 (07)	0.74377 (11)	0.86756 (07)	0.86215 (09)	0.72646 (07)
TeamMX	0.79849 (08)	0.78222 (07)	0.82681 (11)	0.82143 (11)	0.76349 (05)
UniRetro	0.78694 (09)	0.73798 (12)	0.88346 (06)	0.90220 (07)	0.62412 (12)
joseluisUS	0.78164 (10)	0.79178 (04)	0.79532 (13)	0.91315 (06)	0.62631 (11)
ErHulio	0.77040 (11)	0.75278 (09)	0.71208 (16)	0.89207 (08)	0.72466 (09)
AzaelCC	0.75180 (12)	0.78050 (08)	0.89500 (05)	0.79935 (14)	0.53234 (18)
UNED	0.74089 (13)	0.74716 (10)	0.83331 (09)	0.81827 (12)	0.56482 (15)
THANGCIC	0.72724 (14)	0.69146 (15)	0.81471 (12)	0.75769 (16)	0.64511 (10)
URJC-Team	0.72192 (15)	0.65987 (16)	0.83298 (10)	0.80811 (13)	0.58672 (13)
SINAI	0.72147 (16)	0.78571 (05)	0.75395 (15)	0.78469 (15)	0.56154 (16)
UC3M-DEEPLNLP-2	0.64315 (17)	0.69388 (14)	0.47324 (17)	0.82917 (10)	0.57629 (14)
probatzen	0.61084 (18)	0.59167 (18)	0.77987 (14)	0.67453 (18)	0.39729 (20)
UC3MDeep	0.58644 (19)	0.64892 (17)	0.40341 (19)	0.74638 (17)	0.54704 (17)
BASELINE	0.51123 (20)	0.57621 (19)	0.43243 (18)	0.59567 (19)	0.44060 (19)
INFOTEC-LaBD	0.72426 (15)	0.71275 (14)	0.61111 (17)	0.95152 (03)	0.72426 (10)

Table 2: PoliticEs official leaderboard (ranking per metric is shown between parenthesis).

The system that obtained the overall highest performance was LosCalis, with an average macro-f1 of 0.90226, combining BETO and MarIA for document level characteristics extraction together with a Multi-Layer Perceptron classifier for labels decoding. It was followed by NLP-CIMAT and Alejandro Mosquera with an average macro-f1 of 0.89096 and 0.88918, respectively. The NLP-CIMAT team proposed PolitiBETO, based on domain adaptation and ensemble learning. Alejandro Mosquera used word and character n-grams features along with readability features with a L2-regularized logistic regression classifier.

Regarding the results per trait, on the one hand, in relation to the demographic traits, gender has been the most difficult for the participants to classify and, on the other hand, with respect to the psychographic trait, political ideology, the multi-class classification has been the most complex.

Concerning the approaches used, most of the teams propose approaches based on transformers (BETO, MarIA, RoBERTa, ALBERTI, BERTIN, DistilBERT, and multilingual BERT), mainly fine-tuning the pretrained models. Some of them also use traditional machine learning algorithms, being SVM and Logistic Regression the most frequent. There are teams that define different models for the identification of each trait, although most use a single model for all of them. Some of them also combine different approaches through ensemble learning and only one team explores data augmentation techniques.

6 Conclusions

This paper presents the first edition of the PoliticEs task at IberLEF 2022. It is an author profiling task for political ideology in Spanish. So far, several tasks on authorship analysis have been organized in the PAN workshop series (Bevendorff et al., 2021), but none of them focuses on political ideology. Political ideology is a psychographic trait that can be used to understand individual and social behavior. Because of its relevance, we intend to promote author profiling research for political ideology in Spanish through the organization of this shared task.

We are very pleased with the impact of the PoliticEs task, as 63 teams registered for it through CodaLab, the platform on which the competition was organized, which is accessible at the following link: `https://codalab.lisn.upsaclay.fr/competitions/1948`. Finally, of all the registered teams, 20 submitted results and 14 presented working notes to describe their systems, which are summarized in this paper.

As expected, approaches based on transformers are the trend solutions presented by participating teams, but some of them also used traditional machine learning systems or even a combination of them. Finally, it should be mentioned that gender and political ideology multi-class have been the traits most difficult to classify for the participants.

As future work, we plan to extend the dataset by including more users who are neither politicians nor journalists. For this, we ask users to voluntarily sent their tweets at the same time they define their political spectrum. Another idea is to include more subtasks concerning author analysis. For example, we are planing to add a subtask related to stance detection, in order to determine which authors are in favor of certain topics and which users are against. We can use this information to define clusters of users and to observe whether there is a relationship between the topics and the political ideology.

Acknolwedgements

This work was supported by Project LaTe4PSP (PID2019-107652RB-I00) funded by MCIN/AEI/10.13039/501100011033, Project AlInFunds (PDC2021-121112-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project LIVING-LANG (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe, and Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government. In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the industrial doctorate programme, and Salud María Jiménez-Zafra has been partially supported by a grant from Fondo Social Europeo and Administración de la Junta de Andalucía (DOC_01073).

References

Baumgaertner, B., J. E. Carlisle, and F. Justwan. 2018. The influence of political ideology and trust on willingness to vaccinate. *PloS one*, 13(1):e0191728.

Bevendorff, J., B. Chulvi, G. L. D. L. Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, et al. 2021. Overview of PAN 2021: authorship verification, profiling hate speech spreaders on twitter, and style change detection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 419–431. Springer.

Cabrera, H., E. S. Tellez, and S. Miranda. 2022. INFOTEC-LaBD at PoliticES 2022: Low-dimensional Stacking Model for Political Ideology Profiling. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10.

Cañete, J., S. Donoso, F. Bravo-Marquez, A. Carvallo, and V. Araujo. 2022. Albeto and distilbeto: Lightweight spanish language models. *arXiv preprint arXiv:2204.09145*.

Carrasco, S. S. and R. C. Rosillo. 2022. LosCalis at PoliticEs 2022: Political Author Profiling using BETO and MarIA. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

De la Rosa, J., E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, and M. Grandury. 2022. Bertin: Efficient pretraining of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68:13–23.

Espin-Riofrio, C., J. Ortiz-Zambrano, and A. Montejo-Ráez. 2022. SINAI at PoliticEs 2022: Exploring Relative Frequency of Words in Stylometrics for Profile Discovery. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Fatke, M. 2017. Personality traits and political ideology: A first global assessment. *Political Psychology*, 38(5):881–899.

García-Díaz, J. A., Á. Almela, G. Alcaraz-Mármol, and R. Valencia-García. 2020. UMUCorpusClassifier: Compilation and

evaluation of linguistic corpus for Natural Language Processing tasks. *Procesamiento del Lenguaje Natural*, 65(0):139–142.

García-Díaz, J. A., R. Colomo-Palacios, and R. Valencia-García. 2022. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Generation Computer Systems*, 130:59–74.

García-Ochoa Martín-Forero, Á., A. Massotti López, and I. Segura-Bedmar. 2022. UC3MDeep at PoliticEs 2022: Exploring Traditional Machine Learning Algorithms for Political Ideology Detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Gutiérrez Fandiño, A., J. Armengol Estapé, M. Pàmies, J. Llop Palao, J. Silveira Ocampo, C. Pio Carrino, C. Armentano Oller, C. Rodriguez Penagos, A. Gonzalez Agirre, and M. Villegas. 2022. MarIA: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Holgado, C. G. and A. Sinha. 2022. HalBERT at PoliticEs 2022: Are Machine Learning Algorithms better for Author Profiling? In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Kenton, J. D. M.-W. C. and L. K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

Manea, A.-A. and L. P. Dinu. 2022. UniRetro at PoliticEs@IberLef 2022: Political Ideology Profiling using Language Models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Montes-y Gómez, M., J. Gonzalo, F. Rangel, M. Casavantes, M. Á. Álvarez-Carmona, G. Bel-Enguix, H. Jair Escalante, L. Freitas, A. Miranda-Escalada, F. Rodríguez-Sánchez, A. Rosá, M. A. Sobrevilla-Cabezudo, M. Taulé, and R. Valencia-García, editors. 2022. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022).*

Mosquera, A. 2022. Alejandro Mosquera at PoliticEs 2022: Towards Robust Spanish Author Profiling and Lessons Learned from Adversarial Attacks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Ochoa-Hernández, J. L. and Y. Alemán. 2022. TeamMX at PoliticEs 2022: Analysis of Feature Sets in Spanish Author Profiling for Political Ideology. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Ramos, P. C., J. M. Vázquez, V. P. Álvarez, and J. L. D. Olmedo. 2022. I2C at PoliticEs 2022: Using Transformers to Identify Political Ideology in Spanish Tweets. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Rodrigo, Á., H. Fabregat, and R. Centeno. 2022. UNED at PoliticEs 2022: Testing Approximate Nearest Neighbors and Spanish Language Models for Author Profiling in Political Ideology. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Rodríguez-García, M. Á., S. Montalvo Herranz, and R. Martínez Unanue. 2022. URJC-Team at PoliticEs 2022: Political Ideology Prediction using Linear Classifiers. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Santibáñez-Cortés, E., A. Carrillo-Cabrera, Y. A. Castillo-Castillo,

D. Moctezuma, and V. Muñiz-Sánchez. 2022. CIMAT_2021 at PoliticEs 2022: Ensemble Based Classification Algorithms for Author Profiling in Spanish Language. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Ta, H. T., A. B. S. Rahman, L. Najjar, and A. Gelbukh. 2022. THANGCIC at PoliticEs 2022: Term-based BERT for Extracting Political Ideology from Spanish Author Profiling. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Verhulst, B., L. J. Eaves, and P. K. Hatemi. 2012. Correlation not causation: The relationship between personality traits and political ideologies. *American journal of political science*, 56(1):34–51.

Villa-Cueva, E., I. González-Franco, F. Sanchez-Vega, and A. P. López-Monroy. 2022. NLP-CIMAT at PoliticEs 2022: PolitiBETO, a Domain-Adapted Transformer for Multi-class Political Author Profiling. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022). CEUR Workshop Proceedings, CEUR-WS, A Coruña, Spain.*

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 273-280 recibido 05-07-2022 revisado 21-07-2022 aceptado 24-07-2022

Overview of QuALES at IberLEF 2022: Question Answering Learning from Examples in Spanish

Overview de QuALES en IberLEF 2022: Preguntas y Respuestas Automáticas sobre Ejemplos en Español

Aiala Rosá[1], Luis Chiruzzo[1], Lucía Bouza[1], Alina Dragonetti[1], Santiago Castro[2], Mathias Etcheverry[1], Santiago Góngora[1], Santiago Goycoechea[1], Juan Machado[1], Guillermo Moncecchi[1], Juan José Prada[1], Dina Wonsever[1]
[1]Universidad de la República, Montevideo, Uruguay
{aialar, luischir, lucia.bouza, alina.dragonetti, mathiase, sgongora, sgoycoechea, juan.machado, gmonce, prada, wonsever}@fing.edu.uy
[2]University of Michigan, Anne Arbor, USA
sacastro@umich.edu

Abstract: We present the results of the QuALES task, which addresses the problem of Extractive Question Answering from texts. For both training and evaluation we use the QuALES corpus, a corpus of Uruguayan media news about the Covid-19 pandemic and related topics. We describe the systems developed by seven participants, all of them based on different BERT-like language models. The best results were obtained using the multilingual RoBERTa model pre-trained with SQUAD-Es-V2, with a fine tuning on the QuALES corpus.
Keywords: Question Answering for Spanish, Language Models, Datasets for Question Answering.

Resumen: Presentamos los resultados de la tarea QuALES, que aborda el problema de Búsqueda de Respuestas extractiva a partir de textos. Tanto para entrenamiento como para evaluación utilizamos el corpus QuALES, un corpus de noticias de medios uruguayos sobre la pandemia por Covid-19 y temas relacionados. Describimos los sistemas desarrollados por siete participantes, todos ellos basados en diferentes modelos de lenguaje tipo BERT. Los mejores resultados se obtuvieron usando el modelo RoBERTa multilingüe preentrenado con SQUAD-Es-V2, con una fine tuning sobre el corpus QuALES.
Palabras clave: Búsqueda de Respuestas en Español, Modelos de Lenguaje, Corpus para Búsqueda de Respuestas.

1 Introduction

Question Answering (QA) is a classical Natural Language Processing task that is currently gaining great relevance. QA can be roughly divided into two main categories (Jurafsky and Martin, 2021): semantic analysis, where the question is transformed to a query to a knowledge database; and open domain question answering, where, starting from a question written in natural language and a set of documents, the answer to the question is obtained using information retrieval and information extraction techniques.

Open domain question answering involves two main stages: a) getting the relevant documents, generally using methods from the Information Retrieval field (IR) (Manning, Raghavan, and Schütze, 2010), possibly one of the most widely studied topics in NLP, with web search engines as their most noticeable product, b) extracting the answer from those documents. Each of these stages has its own challenges, and the whole task requires a successful outcome for each of them and for their integration.

In this task we address the problem of extractive QA in Spanish, based on a corpus of a specific domain: press news about the Covid-19 pandemic. We focus on the second stage of the task: given a text, extracting the

ISSN 1135-5948. DOI 10.26342/2022-69-24

answer to a question, if there is one.

The rest of the paper is structured as follows: section 2 describes the background of QA, focusing on QA for Spanish; section 3 describes the corpus created for this task and some other resources; section 4 describes the QuALES task; section 5 presents the participants systems and analyzes the results; and, finally, section 6 shows some conclusions.

2 Background

Starting last decade, along with the popularization of distributional semantic methods based on neural networks (Le and Mikolov, 2014; LeCun, Bengio, and Hinton, 2015), this kinds of methods started to be applied to the QA task, achieving significant results improvement (Yu et al., 2014; Min et al., 2018; Xiong, Zhong, and Socher, 2017; Seo et al., 2016).

All these supervised learning approaches were possible due to the existence of research oriented publicly available datasets. These datasets have enabled not only model training, but also constant monitoring of this area's state of the art. Probably the most popular is SQuAD (Rajpurkar et al., 2016). To build this dataset, annotators were presented with a Wikipedia paragraph and asked to write questions that could be answered from the given text. Natural Questions (Kwiatkowski et al., 2019b) was created from actual Google Search queries, where annotators marked the answer into Wikipedia article snippets. TriviaQA (Joshi et al., 2017) contains a set of Trivia questions and answers. CuratedTREC (Baudiš and Šedivý, 2015) dataset generated by the QA track of the NIST TREC conferences contains questions and answers. NewsQA (Trischler et al., 2016) is a machine comprehension dataset of over 100,000 human-generated question-answer pairs, based on set of over 10,000 news articles from CNN.

In the last few years, after the publication of models based on the Transformers architecture (Vaswani et al., 2017) for solving sequence to sequence transformation problems, and particularly language models such as BERT (Devlin et al., 2018) and ALBERT (Lan et al., 2019), there has been a new push in system performance, particularly for the English language. These kinds of models are trained in an self-supervised way, using large volumes of data and computing power. After that stage (called pretraining), they can be easily fine-tuned to apply them to different tasks. Regarding this shared task, we are particularly interested in fine-tuning them to the open domain question answering task.

The study of the QA area is currently very active, as evidenced by the inclusion of a tutorial[1] on this topic in ACL 2020, the main NLP event worldwide.

Throughout the last few years, several QA related tasks have been proposed. Since 2015, one of the tasks of each SemEval annual international workshop on Semantic Evaluation has been related to some form of the QA Task. For example, SemEval-2015 Task 3: "Answer Selection in Community Question Answering" (Nakov et al., 2019b) proposed, given a question, to classify a certain answer as good, bad, or potential, and answer yes/no questions. The challenge was proposed for Arabic and English. SemEval-2017 Task 3: "Community Question Answering" (Nakov et al., 2019a) proposed three different subtasks: Question-Comment Similarity, Question-Question Similarity, and Question-External Comment Similarity. Additionally, for the Arabic language, another task was added: reranking correct answers for a new question.

SemEval-2022 includes the task "Competence-based Multimodal Question Answering" (Task 09)[2], designed to query how well a system understands the semantics of recipes derived from the R2VQ dataset, a multimodal dataset of cooking recipes and videos.

In (Reddy, Chen, and Manning, 2019) the CoQA (Conversational Question Answering) dataset is presented as a challenge. The dataset includes 127k questions with answers, obtained from 8k conversations about text passages. The goal of the CoQA challenge is to measure the ability of machines to understand a text passage and answer a series of interconnected questions that appear in a conversation.

The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) website includes the highest performing systems on the dataset, measuring Exact Match and

[1] https://github.com/danqi/acl2020-openqa-tutorial

[2] https://competitions.codalab.org/competitions/34056

F1 values (see the next section for a description of these metrics). These systems should answer reading comprehension questions, including questions that do not have an answer on the dataset.

Based on Google's Natural Questions Dataset (Kwiatkowski et al., 2019a), the Tensorflow 2.0 machine learning platform includes a Question Answering competition, where the goal is to predict short and long answer responses to real questions about Wikipedia articles. Using the same dataset, the 2020 NeurIPS Conference an open domain question answering challenge (Min et al., 2021) was also proposed, including three tracks where the objective is to build self-contained question answering systems.

For English, the BioASQ challenge for 2022 proposes a task relative to the Covid-19 domain, using a dataset composed by biomedical articles.

QA research for Spanish has evolved much more slowly. However, similar language resources have been created for this language, which makes us think it is possible to study and fine-tune current architectures to obtain competitive results. In particular, there is a recently developed version of BERT for Spanish, dubbed BETO (Cañete et al., 2020), and a version of SQuAD (the main dataset for training and evaluating open domain QA systems) translated to Spanish (Rajpurkar et al., 2016; Carrino, Costa-jussà, and Fonollosa, 2019). The Spanish Question Answering Corpus (SQAC) is an extractive QA dataset created from texts extracted from a mix from different news-wire and literature sources, and it includes 18,817 questions with the annotation of their answer spans from 6,247 textual contexts (Gutiérrez Fandiño et al., 2022).

From 2003 to 2014, the CLEF Question Answering Track has proposed different campaigns related to question answering, some of which included Spanish datasets. For example, together with the CLEF 2009 forum, ResPubliQA, a Question Answering Task over European legislation was proposed (Peñas et al., 2009). The task consisted of extracting a relevant paragraph of text that included the answer to a natural language question. During CLEF 2010, the task was expanded (Peñas et al., 2010) to include an answer selection task (i.e. besides retrieving the relevant paragraph, systems were required to identify the exact answer to the question). It also proposed several cross-lingual tasks, working on two multilingual parallel corpus: the JRC-ACQUIS Multilingual Parallel Corpus (10,700 parallel and aligned documents), and the Europarl collection (150 parallel and aligned document per language), with 200 question-answer pairs provided for evaluation.

Unlike the task we present here, the CLEF tasks have addressed domain-general questions, or questions for some specific domains, but different from the one selected for QuALES. In addition, they have worked with smaller amounts of training and testing data. Some of these CLEF tasks have some characteristics that differ from our proposal, such as datasets oriented to answer multiple choice questions, or natural language questions to be answered from DBPedia structured data (instead of plain text), among other.

3 Corpus

We provided a corpus of around 2,600 question-answer pairs (the QuALES corpus). The training set contains 1,000 of these pairs, while the dev and test sets have around 800 pairs each. Participants could use any other data for training as well, in particular SQuAD (Rajpurkar et al., 2016) or NewsQA (Trischler et al., 2016). The data is available at the Codalab competition site[3].

The QuALES corpus is original and it was created manually by the members of the team and students. It is a Question-Answering corpus in Spanish obtained from a set of Covid-19 related news published in two important news media from Uruguay (La Diaria[4] and Montevideo Portal[5]). It consists of a set of factoid questions mostly about Covid-19 and its repercussions in Uruguay and the world. Table 1 shows the statistics of the dataset.

The corpus annotation was made in two stages: first, we annotated questions by reading only the title and first sentence of the article; then, we thought of questions derived from the reading of the whole article. For each question, we annotated the answer found (if there was any) and the whole sentence context which included it. For the annotation of the answer, we selected the shor-

[3]https://codalab.lisn.upsaclay.fr/competitions/2619

[4]https://ladiaria.com.uy/

[5]https://www.montevideo.com.uy/

Split	Train	Dev	Test
Articles	176	146	143
Questions	948	773	759
Answers	1000	800	821
Empty answers	165	132	103

Table 1: Statistics of the QuALES corpus showing number of articles, number of questions, total number of answers and total number of questions without answers (empty answers) by split.

test span of text contained in the sentence that consisted in a complete answer for the question. All the answers were directly extracted from the text. Some questions may have more than one answer in a given text, in such cases, a set of answers is generated for this question.

We measured inter-annotator agreement between six pairs of annotators. Each pair answered a set of 25 questions, generating a total of 150 questions with two different annotator answers. We obtained an average Exact Match of 0.61 and an average F1 of 0.76. These results are quite low, which shows the complexity of the task, even for humans. The difference between Exact Match and F1 shows the difficulty in defining the limits of the answer, in general the differences are due to the inclusion or not of elements such as prepositions or determiners. The low F1 shows that selecting the fragment that contains the answer, or deciding that a certain fragment has no answer in the text, is also a highly complex task.

We also published some resources to automatically generate a Spanish version of the NewsQA (Trischler et al., 2016) corpus. The complete NewsQA corpus was translated using a machine translation model and after that we aligned the answers. This alignment stage is necessary because, when translating each fragment with its associated question and answer, the substring corresponding to the answer within the fragment, can be translated differently from the associated answer, which is translated decontextualized. In our translation of the corpus, this alignment problem was detected in 49 % of the cases. To solve this problem we worked on two approaches: on the one hand we trained a neural model from pairs of aligned texts, and, on the other hand, we tested some heuristics defined from the analysis of different examples. In order to evaluate the two approaches, we performed a manual evaluation of a subset of 2,000 question-answer pairs. A portion of this curated corpus was used for parameter tuning of the neural model. The neural model for alignment achieved better results than the heuristics approach. Due to licensing issues, it is not possible to provide a link to this dataset, but the resources to recreate this process are available at our github repository[6].

4 Task

The aim of the QuALES task is to develop question answering systems that can answer questions based on news articles written in Spanish. The systems get a full news article and a question, and must find the shortest span of text in the article (if it exists) that answers the question. It should be noted that for some questions there may not be an answer in the given text. está hablando. The training, development and test datasets were generated from the QuALES corpus, as mentioned above. Originally, we planned to have two separate corpora for evaluation, but seeing that the texts often contain Covid-19 related news mixed with other topics, we decided to annotate only one set. Most of the questions in the dataset are about Covid-19 matters, but some of them are also about other topics.

Table 2 shows a sample text with two questions. The answer to one of the questions can be found in the text, while the other is not present.

As one of our evaluation metrics, we measure average Exact Match for all the dataset instances, following the approach of SQuAD (Rajpurkar et al., 2016). We also report, following (Reddy, Chen, and Manning, 2019), the macro-average F1 score of word overlap: we compare each individual prediction against the different human gold standard answers and select the maximum value as system F1 score for that instance; the system performance is the macro-average of all those F1 scores. Some determiners, specifically, definite articles, and punctuation marks were ignored when calculating this evaluation metric.

Some of the questions in the dataset have more than one possible answer, but the systems are expected to generate at most only

[6]https://github.com/pln-fing-udelar/newsqa-es

Comenzaron las clases presenciales en 344 escuelas rurales, con baja asistencia. A las 8.45 dos perros paseaban por el patio de la escuela rural 27 de La Macana, en Florida. Dos maestras con túnicas blancas y tapabocas esperaban a los alumnos que reanudarían las clases presenciales luego de cinco semanas de conexión virtual. Ya estaba instalado el micrófono y el parlante en el patio, habían llegado los inspectores regionales junto con la directora general del Consejo de Educación Inicial y Primaria (CEIP), Irupé Buzzetti, que junto a la prensa local esperaban a los niños. De los 28 alumnos que asisten regularmente, 14 habían dicho que no iban a ir y los otros no habían confirmado. A las 9.00, cuando debían comenzar las clases en la escuela de La Macana, no había ningún niño. (...) La situación de La Macana se repitió en varias de las escuelas que abrieron este miércoles. De las 547 escuelas habilitadas abrieron 344, confirmó a la diaria Limber Santos, director del departamento de Educación Rural del CEIP. De esas escuelas, cerca de 90 no recibieron alumnos; Santos estimó que en la mañana del miércoles 1.030 niños concurrieron a las escuelas, de un total de 3.900 que concurren a las 547 habilitadas y de 2.838 alumnos que tienen matriculadas las 344 escuelas que abrieron. La asistencia, por tanto, llegó a 36 % en el primer día.
Q1: *¿Cuántas escuelas rurales hay en Uruguay?* A1: *De las [547] escuelas habilitadas abrieron 344, confirmó a la diaria Limber Santos, director del departamento de Educación Rural del CEIP.*
Q2: *¿Cuándo vuelven las clases presenciales a todas las escuelas?* A2: –not found in the text–

Table 2: Example of a short text that could be found in the corpus, and two possible questions for the text. Q1 has the answer 547, found in the text, but Q2 does not have an answer in the text.

one answer. Because of this, when there are multiple answers for a question, the metrics evaluate the answer candidate provided by the system against all the possible answers, and get the maximum value.

5 Competition

The competition was run in two phases: a development phase, for which we released the training dataset with annotations and development dataset without annotations; and an evaluation phase, for which we released the annotations of the development dataset and a test dataset without annotations. Participants could train their models using other available corpora, such as the Spanish version of SQuAD or NewsQA.

5.1 Description of the systems

Eighteen participants registered for the competition in our Codalab site, eight of them submitted results for the development phase (73 submissions in total), and seven of them submitted results for the evaluation phase (46 submissions in total). All of the participants that sent results in the evaluation phase used BERT-like models, analyzing if fine-tuning them with proper data improved their performance.

The language model most commonly used by the participants was RoBERTa for Spanish, trained with the corpus from the Biblioteca Nacional de España[7]. BETO[8], multilingual RoBERTa[9], multilingual BERT[10], and distill BERT for Spanish[11] were also used.

The corpora used, in addition to the QuALES corpus, were SQuAD 2.0, NewsQA and SQAC (Spanish versions).

The participant `smaximo` (Máximo, 2022) followed a curriculum learning strategy consisting of fine-tuning BETO and RoBERTa for Spanish on a series of QA datasets. The author found out that the top performance was achieved using RoBERTa first trained on SQAC, then on the Spanish version of SQuAD (SQuAD-ES-v2) and finally on the QuALES corpus.

[7] https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne

[8] https://github.com/dccuchile/beto

[9] https://huggingface.co/docs/transformers/main/en/model_doc/roberta

[10] https://github.com/google-research/bert/blob/master/multilingual.md

[11] https://huggingface.co/mrm8488/distill-bert-base-spanish-wwm-cased-finetuned-spa-squad2-es

The participant `alvarory` (Rodrigo and Peñas, 2022) tried three main approaches. In the first one they fine-tuned RoBERTa for Spanish (base and large versions) and BETO for 10 epochs on the QuALES training set with datasets containing both training and development splits of the task, for a total of 1,800 question-answer pairs. For the second approach they used even more data than the available in QuALES, in order to study the transferability among different datasets when using two pretrained models: RoBERTa and multilingual BERT. The third approach was based on combining different models for returning a single output using two voting schemes.

The participant `avacaondata` (Vaca-Serrano, 2022) addresses extractive QA through an ensemble system composed of three large pre-trained language models in Spanish: MarIA-base, MarIA-large and RigoBERTa. These models were fine-tuned on data from the Spanish version of SQuAD (SQUAD-ES-v2), a Spanish version of NewsQA, generated by the author, and QuALES. The best model is an ensemble that gives scores to each answer based on multiple criteria such as the number of models that predict it and the models' scores. The final predictions were performed by aggregating the output of the resulting models, referred as a meta-ensemble. A number of ensemble strategies were tried, where finally *Grouped Score Aggregation* perform best. This strategy consists on selecting the answer by the count of each answer multiplied by a scaling factor based on the validation scores of the models.

The participant `Bernardo` fine-tuned RoBERTa for Spanish for 3 epochs using th está hablando.e train subsets of SQAC, SQUAD-ES-v2 and QuALES. `ichramm` performed experiments using RoBERTa for Spanish pretrained with the SQAC corpus, and distill-BERT pretrained with SQUAD-ES-v2. His submitted outputs were calculated using the RoBERTa model, fine-tuned on QuALES. He also experimented including the NewsQA version for Spanish, obtaining lower results (one point less in each metric). The participant `gberger` also used the distill-BERT model.

`sebastianvolti` ranked first in both metrics of the competition. He reached his top performance using XML-RoBERTa, a multilingual model, pretrainend with SQUAD-ES-v2 and fine tuned using the QuALES corpus. He also tested a model that included a fine tuning stage with 2,000 examples from the Spanish translation of the NewsQA corpus, prior to fine tuning with the QuALES corpus, which yielded slightly lower results.

5.2 Results

We show the best result for each user for each metric. Please notice that the best exact match and F1 scores might have been obtained in different submissions by the same user.

Table 3 shows the best exact match scores for each user:

User	EM
sebastianvolti	0.5349
ichramm	0.4677
smaximo	0.4598
Bernardo	0.4427
avacaondata	0.3992
gberger	0.3715
alvarory	0.3175

Table 3: Results for the exact match metric.

Table 4 shows the best F1 overlap scores for each user.

User	F1
sebastianvolti	0.7282
Bernardo	0.6159
smaximo	0.6142
avacaondata	0.5877
ichramm	0.5581
gberger	0.4500
alvarory	0.4293

Table 4: Results for the overlap F1 metric.

As can be seen in the tables, the best results achieved (F1: 0.73 and EM: 0.53) are far from those reported on the SQuAD corpus for English on the official SQuAD site (F1: 0.93 and EM: 0.91). Our task differs from what is reported there in that the evaluation texts belong to a specific domain (news about the Covid-19 pandemic), and also in the size of the context provided to search for the answers. In our case, the context is a complete news article, which are longer than the contexts included in the SQuAD dataset.

The best results were obtained by `sebastianvolti`, whose best model is based on RoBERTa pretrained on SQuAD 2.0, fine tuned on the QuALES corpus, and was the only participant who used the multilingual

version of RoBERTa, four other participants used RoBERTa for Spanish (trained on the BNE corpus).

Also note that none of the systems have reached the inter-annotator agreement levels, both for EM and F1, although for F1 the best submission by `sebastianvolti` is the closest by around 5 %.

6 Conclusions

We presented the results of the QuALES competition on Question Answering Learning from Examples in Spanish. Seven participants submitted systems to the competition, and the best systems achieved 0.53 in exact match and 0.73 in average F1 overlap.

The extractive Q&A task, although showing very good results on the main available benchmark, the SQuAD corpus, still presents great challenges when working with different data and searching for answers in larger contexts.

The QuALES corpus, despite its rather small size, provided significant improvements in training, complementing other larger corpora taken as a base, mainly SQuAD (Spanish version) and SQAC.

References

Baudiš, P. and J. Šedivý. 2015. Modeling of the question answering task in the yodaqa system. In J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. Jones, E. San Juan, L. Capellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 222–228, Cham. Springer International Publishing.

Carrino, C. P., M. R. Costa-jussà, and J. A. Fonollosa. 2019. Automatic spanish translation of the squad dataset for multilingual question answering. *arXiv preprint arXiv:1912.05200*.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:1–10.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gutiérrez Fandiño, A., J. Armengol Estapé, M. Pàmies, J. Llop Palao, J. Silveira Ocampo, C. Pio Carrino, C. Armentano Oller, C. Rodriguez Penagos, A. Gonzalez Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Joshi, M., E. Choi, D. S. Weld, and L. Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

Jurafsky, D. and J. H. Martin. 2021. Speech and language processing. 3rd edition draft. *US: Prentice Hall*.

Kwiatkowski, T., J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. 2019a. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Kwiatkowski, T., J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, M. Kelcey, J. Devlin, K. Lee, K. N. Toutanova, L. Jones, M.-W. Chang, A. Dai, J. Uszkoreit, Q. Le, and S. Petrov. 2019b. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Le, Q. and T. Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Manning, C., P. Raghavan, and H. Schütze. 2010. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103.

Min, S., J. Boyd-Graber, C. Alberti, D. Chen, E. Choi, M. Collins, K. Guu, H. Hajishirzi, K. Lee, J. Palomaki, et al. 2021. Neurips 2020 efficientqa competition: Systems,

analyses and lessons learned. In *NeurIPS 2020 Competition and Demonstration Track*, pages 86–111. PMLR.

Min, S., V. Zhong, R. Socher, and C. Xiong. 2018. Efficient and robust question answering from minimal context over documents. *arXiv preprint arXiv:1805.08092*.

Máximo, S. 2022. Supervised domain adaptation for extractive question answering in spanish.

Nakov, P., D. Hoogeveen, L. Màrquez, A. Moschitti, H. Mubarak, T. Baldwin, and K. Verspoor. 2019a. Semeval-2017 task 3: Community question answering. *arXiv preprint arXiv:1912.00730*.

Nakov, P., L. Màrquez, W. Magdy, A. Moschitti, J. Glass, and B. Randeree. 2019b. Semeval-2015 task 3: Answer selection in community question answering. *arXiv preprint arXiv:1911.11403*.

Peñas, A., P. Forner, Á. Rodrigo, R. Sutcliffe, C. Forăscu, and C. Mota. 2010. Overview of respubliqa 2010: Question answering evaluation over european legislation. In *CLEF*.

Peñas, A., P. Forner, R. Sutcliffe, Á. Rodrigo, C. Forăscu, I. Alegria, D. Giampiccolo, N. Moreau, and P. Osenova. 2009. Overview of respubliqa 2009: Question answering evaluation over european legislation. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 174–196. Springer.

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Reddy, S., D. Chen, and C. D. Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Rodrigo, A. and A. Peñas. 2022. Uned@quales 2022: Testing the performance of transformer-based language models for spanish question-answering.

Seo, M., A. Kembhavi, A. Farhadi, and H. Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Trischler, A., T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Vaca-Serrano, A. 2022. Adversarial question answering in spanish with transformer models.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xiong, C., V. Zhong, and R. Socher. 2017. Dcn+: Mixed objective and deep residual coattention for question answering. *arXiv preprint arXiv:1711.00106*.

Yu, L., K. M. Hermann, P. Blunsom, and S. Pulman. 2014. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 281-287 recibido 05-07-2022 revisado 18-07-2022 aceptado 21-07-2022

Overview of ReCoRES at IberLEF 2022: Reading Comprehension and Reasoning Explanation for Spanish

Overview de ReCoRES en IberLEF 2022: Comprensión de Lectura y Explicación de Razonamiento en Español

Marco Antonio Sobrevilla Cabezudo[1], Diego Diestra[2], Rodrigo López[2], Erasmo Gómez[2], Arturo Oncevay[3], Fernando Alva-Manchego[4]
[1]University of São Paulo
[2]Department of Engineering, Pontificia Universidad Catolica del Perú
[3]School of Informatics, University of Edinburgh
[4]Cardiff University
msobrevillac@usp.br, {ddiestra, a20112387, hector.gomez}@pucp.pe,
a.oncevay@ed.ac.uk, alvamanchegof@cardiff.ac.uk

Abstract: This paper presents the ReCoRES task, organized at IberLEF 2022, within the framework of the 38th edition of the International Conference of the Spanish Society for Natural Language Processing. The main goal of this shared-task is to promote the task of Reading Comprehension and Verbal Reasoning. This task is divided into two sub-tasks: (1) identifying the correct alternative in reading comprehension questions and (2) generating the reasoning used to select an alternative. In general, 3 teams participated in this event, mainly proposing transformer-based neural models in conjunction with additional strategies. The results of this event, insights and some challenges are presented, opening a range of possibilities for future work.
Keywords: Reading Comprehension, Reasoning Explanation, Spanish.

Resumen: Este artículo presenta la tarea ReCoRES, organizada en IberLEF 2022, en el marco de la 38 edición de la Conferencia Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural. El objetivo de esta tarea es promover la tarea de Comprensión de Lectura y Razonamiento Verbal. Esta tarea es dividida en dos sub-tareas: (1) la identificación de la alternativa correcta en preguntas de comprensión de lectura y (2) la generación del razonamiento usado para seleccionar una alternativa. En general, 3 equipos participaron de este evento proponiendo mayormente modelos neuronales basados en transformers con algunas estrategias adicionales. Los resultados de este evento así como aprendizajes y algunos desafíos son presentados, abriendo un abanico de posibilidades como trabajos futuros.
Palabras clave: Comprensión de Lectura, Explicación del Razonamiento, Español.

1 Introduction

Question Answering (QA) consists of returning an accurate and short answer given a Natural Language question. According to Rogers et al. (2020), QA can be approached from two main perspectives: Open QA, in which responses are recovered from several sources such as Web pages and knowledge bases, and Reading Comprehension (RC), where the answer is recovered from a single document.

RC datasets are classified into three categories according to their answer type: (1) span-selection datasets, where the text explicitly includes the answer, (2) multiple-choice datasets, where systems have to select an answer from a list of candidates; and (3) freeform answers dataset, where answers are written in freeform. Most RC datasets are in the first category, with the most popular being SQuAD (Rajpurkar et al., 2016). An explicit limitation of these span-selection datasets is that they can only target information explicitly mentioned in the text and often get solved with shallow lexical match-

ISSN 1135-5948. DOI 10.26342/2022-69-25

ing (Rogers et al., 2020).

Using a multiple-choice dataset is a common and realistic way to measure reading comprehension in humans (Echegoyen, Álvaro Rodrigo, and Peñas, 2020). In addition, Rogers et al. (2020) point out that multiple-choice is a better format to assess language understanding of automatic systems. It is because it requires a high degree of textual inference and the development of strategies for selecting the correct answer.

For English, there are diverse Multiple-Choice QA datasets, such as RACE (Lai et al., 2017), Entrance Exams (Peñas et al., 2011) and QuAIL (Rogers et al., 2020). However, that is not the case for most languages. For Spanish, in particular, there are two QA datasets available: SQuAD-es (Carrino, Costa-jussà, and Fonollosa, 2020) and Entrance Exams (EE) (Peñas et al., 2011). However, these datasets present some limitations originated by the nature of the dataset or some aspects like the size. For example, SQuAD-es is a span-based QA dataset, i.e., the answers are included in the text explicitly. In the case of EE, it is a multiple-choice QA dataset in which, even though questions demand a certain level of reasoning, the dataset size is quite small (43 texts and 191 questions), constraining the exploration of current State-of-the-Art approaches.

In order to contribute to the development of research in Question-Answering/Reading Comprehension for Spanish, this shared-task aims to:

- Introduce a new and more extensive multiple-choice QA dataset for Spanish based on university entrance examinations, where questions aim to evaluate humans instead of computers and include extra information about the reasoning used to choose an alternative.
- Evaluate multiple-choice question answering, and reasoning generation approaches on this dataset.

2 Task Description

This shared-task consists of two sub-tasks:

- Sub-task 1 - Machine Reading Comprehension: given a text, a question, and a set of candidate answers, a system must select the correct answer.
- Sub-task 2 - Reasoning Explanation: given a text and a question, a system must generate an explanation for its answer selection

3 Dataset

The dataset used in this shared-task was extracted from actual university entrance examinations provided by Peruvian institutions that train students for entrance examinations and includes diverse topics and question types that require a certain level of reasoning. Source documents that compose the dataset were initially available in PDF format. This way, we built the dataset by applying two strategies: (1) using an OCR to convert the PDF documents to TXT format and then manually correcting them to fix possible OCR problems, and (2) transcripting the PDF files. Eight collaborators and two organization committee members performed manual revision and transcription.

The whole dataset comprises 439 texts, and 1,822 questions with 2-7 candidate answers each, divided into training, development, and test sets with 257 texts (1,073 questions), 91 texts (363 questions), and 91 texts (386 questions), respectively. Additionally, each question-answer pair instance includes a short explanation as reasoning support for choosing a candidate answer [1].

Figure 1 shows an example of a long text, a question with five alternatives, and the corresponding reasoning. It is worth noting that texts are long, and questions are not described most typically -using question markers and wh-questions; instead, these are described as a sentence that needs to be completed.

4 Experimental Setup

4.1 Baseline

We use two baselines for sub-task 1. The first consists of randomly choosing an answer among the alternatives for each question, and the second is a BERT-based baseline[2], similar to the one used by Rogers et al. (2020). It works this way: for each answer option, the context, question, and choice are joined and

[1]The dataset is available at https://github.com/ddiestra/mrc-dataset.

[2]We use the BERT model available at https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased.

Text: "El trabajo es en primer término un proceso entre la naturaleza y el hombre, proceso en que este realiza, regula y controla mediante su propia acción su intercambio de materias con la naturaleza. En este proceso, el hombre se enfrenta como un poder natural con la materia de la naturaleza. Pone en acción las fuerzas naturales que forman su corporeidad, los brazos y las piernas, la cabeza y la mano, para de ese modo asimilarse, bajo una forma útil para su propia vida, las materias que la naturaleza le brinda. Y a la par que de ese modo actúa sobre la naturaleza exterior a él y la transforma, transforma su propia naturaleza, desarrollando las potencias que dormitan en él y sometiendo el juego de sus fuerzas a su propia disciplina. Aquí no vamos a ocuparnos de las primeras formas de trabajo, formas instintivas y de tipo animal. Aquí, partimos del supuesto del trabajo plasmado ya bajo una forma en la que pertenece exclusivamente al hombre. Una araña ejecuta operaciones que semejan a las manipulaciones del tejedor, y la construcción de los panales de las abejas podría avergonzar, por su perfección, a más de un maestro de obras. Pero, hay algo en que el peor maestro de obras aventaja, desde luego, a la mejor abeja, y es el hecho de que, antes de ejecutar la construcción, la proyecta en su cerebro. Al final del proceso de trabajo, brota un resultado que antes de comenzar el proceso existía ya en la mente del obrero; es decir, un resultado que tenía ya existencia ideal. El obrero no se limita a hacer cambiar de forma la materia que le brinda la naturaleza, sino que, al mismo tiempo, realiza en ella su fin, fin que él sabe que rige como una ley las modalidades de su actuación y al que tiene necesariamente que supeditar su voluntad. Y esta supeditación no constituye un acto aislado. Mientras permanezca trabajando, además de esforzar los órganos que trabajan, el obrero ha de aportar esa voluntad consciente del fin a que llamamos atención, atención que deberá ser tanto más reconcentrada cuanto menos atractivo sea el trabajo, por su carácter o por su ejecución, para quien lo realiza, es decir, cuanto menos disfrute de él el obrero como de un juego de sus fuerzas físicas y espirituales."

Question: Medularmente, el autor intenta dilucidar:

Alternatives:

A. las diferencias entre lo instintivo y lo planificado.
B. la naturaleza del trabajo exclusivamente humano.
C. el carácter pernicioso del trabajo en la actualidad.
D. la supremacía de la naturaleza frente a la humanidad.
E. las etapas que componen el proceso productivo.

Answer: B

Reason: El autor busca caracterizar el trabajo humano frente a lo instintivo, señala así que el trabajo humano está supeditado a un fin.

Figure 1: ReCoRES's Example.

used as input, and the output is its probability, and the most likely option is selected as the answer. Among the settings, we train the model for 1 epoch and use a learning rate of 3e-5 with Adam optimizer.

The baseline for sub-task 2 is a T5-based one that receives the text and the question as inputs and returns the reason[3]. In addition, we evaluate a two-stage approach. Firstly, we select the two most important sentences[4] for a specific question according to cosine similarity. [5]. Then we train a T5-based model, similar to the first baseline. The parameters used were: input length and output length o 512 and 100 tokens, respectively, a learning rate of 0.003 with Adafactor optimizer, and a batch size of 8 with gradient accumulation of 4 steps. Besides, we freeze the embedding layer. Finally, we select the model with the best perplexity in the development set after 7 epochs. During prediction, we use a beam size of 5.

4.2 Evaluation

Sub-task 1 is evaluated in two ways. Firstly, we will evaluate the accuracy, i.e., the number of correct answers in relation to the total number of questions. The second measure is c@1 (Peñas and Rodrigo, 2011), used at CLEF (Rodrigo et al., 2015). c@1 is a conservative metric that penalizes incorrect answers, encouraging systems to not choose an answer unless they are certain.

Sub-task 2 is evaluated in two ways as

[3]We use the T5 model available at https://huggingface.co/flax-community/spanish-t5-small.

[4]We used the two most important sentences because it produced the best results in the development set.

[5]This strategy is inspired by query-based automatic summarization (Hovy, 2005).

well. The first one will consist of running automatic semantic metrics BERTScore (Zhang et al., 2020) to measure the similarity between the generated explanation and its manual reference. We will use this metric instead of classical BLEU (Papineni et al., 2002), or METEOR (Banerjee and Lavie, 2005) because "reasons" can be open and diverse. The second one is a manual evaluation of three quality criteria:

- Accuracy, to measure how accurate is the output system in relation to the original output;
- Fluency (Howcroft et al., 2020), that measures the degree to which a text "flows welländ is not e.g. a sequence of unconnected parts.
- Readability (Howcroft et al., 2020), that measures if the output system is understandable or easy to read.

To perform the manual evaluation, we recruit some crowdworkers. In particular, these were undergraduate students who had experience in this task (Reading Comprehension). The crowdworkers were guided to rate each criteria using an interval of 1-5, being 1 the worst and 5 the best.

5 Participants

In this edition, 3 teams registered on the task and submitted results. However, two of them presented working notes describing their systems. The following is a brief summary of the final proposals submitted:

5.1 MRCPUCP

This team only participated in the sub-task 1[6]. They proposed a BERT-based approach in which all text, alternatives, and reasoning are concatenated and used as input, and the output is one of the alternatives. They used BETO as BERT-based model for Spanish (similar to the baseline) and finetune the model on the dataset they built.

5.2 SADA (Baggetto et al., 2022)

This team only participated in the sub-task 1. The authors explore using encoder models, generative models, clue generation systems, and dataset expansion. In experiments, the best model was a pre-trained multilingual T5 model finetuned on an expanded multilingual dataset.

[6]This team did not present working notes for the present shared-task.

Sub-task 1		
Team	**Accuracy**	**c@1**
Baseline (Random)	0.2514	0.2514
Baseline (BERT)	0.1917	0.1917
Baseline (BERT) + Threshold	0.0492	0.0896
Versae & Nandezgarcia	0.4067	0.4067
SADA	0.7254	0.7254
MRCPUCP	**0.7591**	**0.7591**
Sub-task 2		
Team	**BERTScore**	
Baseline T5	0.6579	
T5 - 2 sentences	0.6652	
Versae & Nandezgarcia	**0.6867**	

Table 1: Results of Automatic Evaluation.

5.3 Versae & Nandezgarcia (De la Rosa and Fernández, 2022)

This team participated in both sub-tasks. The authors tested several methods for classic fine-tuning of encoder-only language models for the task of reading comprehension and a zero-shot approach for reasoning explanation using a decoder-only model.

6 Results and Discussion

Table 1 presents the results for the sub-task 1 and sub-task 2. For sub-task 1 (Reading Comprehension), the best performance was obtained by the MRCPUCP team, and the SADA team obtained the second-best one, only 3 points lower than the first one.

It is worth noting that all teams used pretrained models such as BERT (Devlin et al., 2019) or T5 (Raffel et al., 2020) in conjunction with some additional strategies. Among the strategies, we can highlight the use of reasoning as part of the input and its helpfulness in getting the correct answers in the reading comprehension task. On the other hand, multilingual information has proven to be helpful, even when the domains are different.

Concerning sub-task 2, the best performance in the automatic evaluation was obtained by the work of Versae & Nandezgarcia, being almost 2 points higher than the strong T5-based baseline. It is worth noting that the winning proposal used a zero-shot approach, i.e., no training data of this task was used for learning to generate the reasoning.

Due to input texts in our dataset being long, we wonder how much do text length influence the performance? To verify it, we

divide the test set in text subsets according to its length, as shown in the X-axis in Figures 2 and 3.

Figure 2 shows how the accuracy changes according to the text length for all proposals (baseline is the BERT-based one). We can note that, as was expected, the performance decreases when the texts are longer, except for the cases where the length is higher than 500 tokens. This result is suspicious as most proposals were BERT-based models. Thus, the maximum length was defined as 512 tokens. However, the proposal of SADA uses a T5-based model that can deal with these lengths. In the case of the longest texts (between 850 and 900 tokens), we must note that the performance was almost 0.25 because the models usually chose an alternative by chance, and it was correct for all questions that had the same alternative as correct.

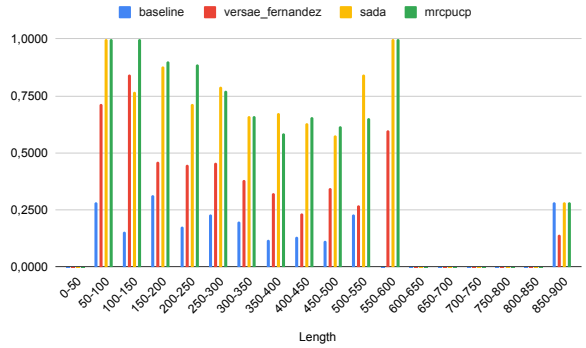


Figure 2: Analysis of the performance on the Machine Comprehension task according different text lengths.

Figure 3 shows how BERTScore changes according to the text length for all proposals. We note that even when the values obtained by Versae & Nandezgarcia are a bit higher in all subsets, these are almost the same (a bit higher than 0.60). These results can suggest that models can deal with different text lengths in the same way or that metric is not good enough to determine what is the best proposal. However, a deeper study is necessary to determine the actual reason for getting these results.

Finally, Table 2 presents the human evaluation results. It shows that fluency and readability achieve similar scores (almost 4) for all proposals, being a bit better for the proposal of Versae & Nandezgarcia. This is expected as all models are based on big language models that can usually generate fluent and readable texts. In the case of accuracy, we can see that the proposal of Versae & Nandezgarcia obtained the best results. However, results are still lower than 3, proving that this task is harder and the automatic evaluation metric could not be suitable.

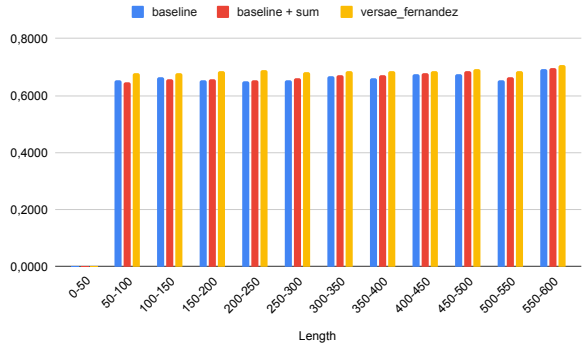


Figure 3: Analysis of the performance on the Reasoning Explanation task according different text lengths.

	Accuracy	Fluency	Readability
Baseline T5	1.08 ±0.40	4.04 ±1.16	3.95 ±1.26
T5 - 2 sentences	1.20 ±0.51	4.08 ±1.07	3.93 ±1.22
Versae & Nandezgarcia	2.35 ±1.39	4.33 ±0.90	4.33 ±0.92

Table 2: Human Evaluation. Accuracy, Fluency and Readability were rated in an interval of 1-5. Results are shown in terms of mean ± standard deviation.

7 *Conclusion*

We presented the first edition of the ReCoRES task at IberLEF, including two sub-tasks: reading comprehension and reasoning explanation.

In general, three teams participated in this shared-task: three for sub-task 1 and one for sub-task 2. However, only two teams sent their working notes. All proposals were based on pre-trained language models with some additional strategies.

Overall, the winner of sub-task 1 was the MRCPUCP team, and the winner of sub-task 2 was the Versae-Nandezgarcia team. About the results, some interesting findings about the helpfulness of incorporating reasoning information and multilingual datasets in the reading comprehension task and the need to use more suitable metrics and other strategies to deal with the reasoning explanation task as this one has proven to be complicated.

As future work, we plan to extend the cur-

rent corpus for both sub-tasks and annotate different question types according to the taxonomy proposed by Rogers et al. (2020) to verify what are the actual abilities of pretrained language models. Besides, we plan to annotate text segments that explain the reasoning to build an extractive reasoning explanation dataset instead of an abstractive one like the one used in this shared-task.

Acknowledgements

The authors acknowledge the support of the undergraduate students at the department of Software Engineering at the Universidad Nacional Mayor de San Marcos in the manual evaluation.

References

Baggetto, P., S. Ramos, J. García, and J. R. Navarro. 2022. Study on text comprehension and MCQA in spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*, A Coruña, Spain. CEUR Workshop Proceedings.

Banerjee, S. and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Carrino, C. P., M. R. Costa-jussà, and J. A. R. Fonollosa. 2020. Automatic Spanish translation of SQuAD dataset for multi-lingual question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5515–5523, Marseille, France, May. European Language Resources Association.

De la Rosa, J. and A. Fernández. 2022. Zero-shot Reading Comprehension and Reasoning for Spanish with BERTIN GPT-J-6B. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*, A Coruña, Spain. CEUR Workshop Proceedings.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Echegoyen, G., Álvaro Rodrigo, and A. Peñas. 2020. Cross-lingual training for multiple-choice question answering. *Procesamiento del Lenguaje Natural*, 65(0):37–44.

Hovy, E. 2005. Text summarisation. *The Oxford Handbook of computational linguistics*, pages 583–598.

Howcroft, D. M., A. Belz, M.-A. Clinciu, D. Gkatzia, S. A. Hasan, S. Mahamood, S. Mille, E. van Miltenburg, S. Santhanam, and V. Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland, December. Association for Computational Linguistics.

Lai, G., Q. Xie, H. Liu, Y. Yang, and E. Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark, September. Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Peñas, A., E. H. Hovy, P. Forner, Á. Rodrigo, R. F. E. Sutcliffe, C. Forascu, and C. Sporleder. 2011. Overview of QA4MRE at CLEF 2011: Question answering for machine reading evaluation. In V. Petras, P. Forner, and P. D. Clough, editors, *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands*, volume 1177 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Peñas, A. and A. Rodrigo. 2011. A simple measure to assess non-response. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1424, Portland, Oregon, USA, June. Association for Computational Linguistics.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Rodrigo, Á., A. Peñas, Y. Miyao, E. H. Hovy, and N. Kando. 2015. Overview of CLEF QA entrance exams task 2015. In L. Cappellato, N. Ferro, G. J. F. Jones, and E. SanJuan, editors, *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*, volume 1391 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rogers, A., O. Kovaleva, M. Downey, and A. Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731, Apr.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 289-299 recibido 05-07-2022 revisado 18-07-2022 aceptado 21-07-2022

Overview of Rest-Mex at IberLEF 2022: Recommendation System, Sentiment Analysis and Covid Semaphore Prediction for Mexican Tourist Texts

Resumen de la tarea Rest-Mex en IberLEF 2022: Sistema de Recomendación, Análisis de Sentimiento y Predicción de Semáforo Covid para Textos Turísticos Mexicanos

Miguel Á. Álvarez-Carmona[1,2], **Ángel Díaz-Pacheco**[1], **Ramón Aranda**[1,2], **Ansel Y. Rodríguez-González**[1,2], **Daniel Fajardo-Delgado**[3], **Rafael Guerrero-Rodríguez**[4], **Lázaro Bustio-Martínez**[5]
[1]Centro de Investigación Científica y de Educación Superior de Ensenada
[2]Consejo Nacional de Ciencia y Tecnología
[3]Tecnológico Nacional de México Campus Ciudad Guzmán
[4]Universidad de Guanajuato
[5]Universidad Iberoamericana, Ciudad de México
{malvarez, diazpacheco, aranda, ansel}@cicese.edu.mx
daniel.fd@cdguzman.tecnm.mx, r.guerrero-rodriguez@ugto.mx,
lazaro.bustio@ibero.mx

Abstract: This paper presents the framework and results from the Rest-Mex task at IberLEF 2022. This task considered three tracks: Recommendation System, Sentiment Analysis and Covid Semaphore Prediction, using texts from Mexican touristic places. The Recommendation System task consists in predicting the degree of satisfaction that a tourist may have when recommending a destination of Nayarit, Mexico, based on places visited by the tourists and their opinions. On the other hand, the Sentiment Analysis task predicts the polarity of an opinion issued and the attraction by a tourist who traveled to the most representative places in Mexico. We have built corpora for both tasks considering Spanish opinions from the TripAdvisor website. As a novelty, the Covid Semaphore Prediction task aims to predict the color of the Mexican Semaphore for each state, according to the Covid news in the state, using data from the Mexican Ministry of Health. This paper compares and discusses the participants' results for all three tacks.
Keywords: Rest-Mex 2022, Sentiment Analysis, Covid Prediction, Mexican Tourist Text.

Resumen: Este artículo presenta el marco y los resultados de la tarea Rest-Mex en IberLEF 2022. Esta tarea consideró tres sub tareas: Sistema de recomendación, Análisis de sentimiento y Predicción de semáforo Covid, utilizando textos de lugares turísticos mexicanos. La tarea del Sistema de Recomendación consiste en predecir el grado de satisfacción que puede tener un turista al recomendar un destino de Nayarit, México, con base en los lugares visitados por los turistas y sus opiniones. Por otro lado, la tarea de Análisis de Sentimiento predice la polaridad de una opinión emitida y la atracción por parte de un turista que viajó a los lugares más representativos de México. Hemos construido corpus para ambas tareas teniendo en cuenta las opiniones en español de TripAdvisor. Como novedad, la tarea de Predicción de Semáforo Covid tiene como objetivo predecir el color del Semáforo Mexicano para cada estado, de acuerdo a las noticias Covid en el estado, utilizando datos de la Secretaría de Salud de México. Este documento compara y discute los resultados de los participantes para las tres sub tareas.
Palabras clave: Rest-Mex 2022, Análisis de sentimientos, Predicción de covid, Textos Turísticos Mexicanos.

ISSN 1135-5948. DOI 10.26342/2022-69-26

1 Introduction

Tourism is a social, cultural, and economic phenomenon related to people's movement to places outside their usual place of residence for personal or business/professional reasons (Guerrero-Rodriguez et al., 2021). This activity is vital in various countries, including Mexico (Álvarez-Carmona et al., 2022)[1], where tourism represents 8.7% of the national GDP, generating around 4.5 million direct jobs (Arce-Cardenas et al., 2021).

In 2021, Rest-Mex emerged, which is an evaluation forum (Álvarez-Carmona et al., 2021). This forum is the first that seeks to specialize in text analysis from tourism to provide solutions to different tasks for Mexican Spanish. In its 2021 edition, the Rest-Mex proposed two different tasks. Analysis of recommendation systems and sentiment analysis. For both tasks, data was collected from the TripAdvisor site.

For this Rest-Mex edition, we proposed three sub-tasks: Recommendation System, Sentiment Analysis on Mexican tourist texts, and as a novelty, the task of Determining the color of the Mexican Covid-19 epidemiological semaphore is added.

For this purpose, 3 data sets have been built. We collected **2,263** instances from 2,011 users who visited 18 touristic places in Nayarit, Mexico, for the recommendation system task. For the sentiment analysis task, the data is labeled to determine the polarity and origin of each opinion. For this, **43,150** opinions were collected from various tourist spots in Mexico. Finally, for determining the epidemiologic semaphore, **131,471** news items referring to covid were collected for all the states of the Mexican Republic, grouped into **2,656** weeks.

The remainder of this paper is organized as follows: Section 2 describes this forum's collection-building process and the evaluation metrics. Section 3 summarizes the solutions submitted for the tasks and shows the results obtained by the participants' systems and the analysis. Finally, Section 4 presents the conclusions obtained by this evaluation forum.

[1]Mexico is in the world's top ten and the second Iberoamerican country related to the arrival of international tourists.

2 Evaluation framework

This section outlines the construction of the three used corpora, highlighting particular properties, challenges, and novelties. It also presents the evaluation measures used for the tasks.

2.1 Recommendation System corpus

The first subtask consists of a classification task where the participating system can predict the degree of satisfaction a tourist may have when recommending a destination.

The collection consists of **2,263 instances** with 2,011 tourists and 18 touristic places from Nayarit, Mexico. This collection was obtained from the tourists who shared their satisfaction on TripAdvisor between 2010 and 2020. Each class of satisfaction is an integer between [1, 5], where {1: Very bad, 2: Bad, 3: Neutral, 4: Good,5: Very good}. Each instance consists of two parts:

1. **User information**:
   - Gender: The tourist's gender.
   - Place: The tourist place that the tourist recommends a visit.
   - Location: The place of origin of the tourist (the central, northeast, northwest, west, and southeast regions refer to the regions of Mexico).
   - Date: Date when the recommendation was issued.
   - Type: Type of trip that the tourist would do. The type would be in [Family, Friends, Alone, Couple, Business]
   - History: The history of the places the tourist has visited and his/her opinions on each of these places.
2. **Place information**: A brief text description of the place and a series of representative characteristics of the place as a type of tourism that can be done there (adventure, beach, relaxation, among others.), If it is a family atmosphere, private or public, it is free or paid, among others.

We use a 70/30 partition to divide into train and test. This means that we used

Class	Train instances	Test instances
1	45	20
2	53	24
3	167	72
4	457	196
5	860	369
Σ	1582	681

Table 1: Instances distribution for the recommendation system task.

1,582 labeled instances for the training partition while we used 681 unlabeled instances for the test partition.

Table 1 shows the distribution of the instances for the recommendation system task for the train and test partitions.

The class imbalance is clear since class 5 represents around 50 % of the total instances, making this task very difficult.

Formally the problem of this task is defined as:

"Given a TripAdvisor tourist and a Mexican tourist place, the goal is to automatically obtain the degree of satisfaction (between 1 and 5) the tourist may have when visiting that place."

2.2 Sentiment Analysis corpus

The second subtask is a classification task where the participating system can predict the polarity and the tourist attraction of an opinion issued by a tourist who traveled to the representative Mexican places. This collection was obtained from the tourists who shared their opinion on TripAdvisor between 2002 and 2021. Each opinion's polarity is an integer between [1, 5], where{1: Very bad, 2: Bad, 3: Neutral, 4: Good,5: Very good}. Also, the participants must determine the attractiveness of the opinion being issued. The possible classes are Attractive, Hotel and Restaurant.

The corpus consists of **43,150 opinions** shared by tourists. Like the recommendation task, we use a 70/30 partition to divide into train and test. This means that we used 30,212 labeled instances for the train partition, while we used 12,938 unlabeled instances for the test partition.

Table 2 shows the distribution of the instances for the sentiment analysis task for the train and test partitions for polarity and attraction.

As with the recommendation system subtask, the class imbalance is clear since class 5

Pol			Attr		
Class	Train	Test	Class	Train	Test
1	547	256	Attractive	5197	2216
2	730	315	Hotel	16565	7100
3	2121	884	Restaurant	8450	3622
4	5878	2423	-		
5	20936	9060	-		
Σ	30212	12938	-	30212	12938

Table 2: Instances distribution for polarity and attraction traits on sentiment analysis task.

and the class Hotel represents around 50 % of the total instances, making this a task with a significant degree of difficulty too.

Formally the problem of this task is defined as:

"Given an opinion about a Mexican tourist place, the goal is to determine the polarity, between 1 and 5, of the text and the visited attraction, which could be an attraction, a hotel, or a restaurant."

2.3 Covid Semaphore Prediction

The last subtask is a classification task where the participating system can predict the future of the covid semaphore through the news. This collection was obtained from news websites that published reports regarding covid from June 2020 to December 2021. For this task, **131,471** news items referring to covid were collected for all the states of the Mexican Republic, grouped into **2,656** weeks. Like the previous tasks, a 70/30 partition was made for training and testing. Therefore, 94,540 news items distributed in 1912 weeks were selected for the training corpus. The test corpus consists of 36,931 news items distributed over 744 weeks.

Each week or instance consists of 4 labels. These labels correspond to the semaphore color of the instance after f weeks in the future. The possible colors to detect are: red, orange, yellow, and green, where red is the color that places the most restrictions on public activities and green is the color that corresponds to the best possible situation. The participants must predict the color of the semaphore for the weeks $f = \{0, 2, 4, 8\}$. For more information regarding the covid semaphore, you can consult (Alvarez-Carmona and Aranda, 2022), (Álvarez-Carmona et al., 2022b).

Table 3 shows the distribution of the instances for eaxh f value.

	$f = 0$		$f = 2$		$f = 4$		$f = 8$	
Class	Train	Test	Train	Test	Train	Test	Train	Test
Red	248	87	201	71	179	63	139	42
Orange	680	273	680	275	673	261	655	252
Yellow	545	216	554	221	568	227	615	232
Green	439	168	477	177	492	193	503	218
Σ	1912	744	1912	744	1912	744	1912	744

Table 3: Instances distribution for semaphore prediction.

Like the other tasks, for this corpus, it can be seen that the red class is the minority, which could be the most crucial class to predict, so this task has considerable complexity to solve.

Formally the problem of this task is defined as:

"Given the news set for a week f in a state of the Mexican Republic x, each system must return the color of the covid epidemiological semaphore for weeks f, $f+2$, $f+4$, and $f+8$ for the x state."

2.4 Performance measures

Systems are evaluated using standard evaluation metrics, including accuracy and F-measure, but MAE (mean absolute error) will rank the submissions for the recommendation system task. MAE are defined as equation 1.

$$MAE_{S_x} = \frac{1}{n}\sum_{i=1}^{n}|T(i) - S_x(i)| \qquad (1)$$

Where S_x is a participating system x, $T(i)$ is the result of the instance i according to the Ground Truth, and $S_x(i)$ is the output of the participant system x, for instance, i. Finally, n is the number of instances in the collection.

We proposed a measure to evaluate the sentiment analysis task for this edition. This measure is defined as shown in the equation 2.

$$measure_S = \frac{\frac{1}{1+MAE_p} + F_A}{2} \qquad (2)$$

Where F_A is the average among the micro F-measure for each class (hotel, restaurant, and attractive), and MAE evaluates the polarity.

Finally, for evaluating the semaphore task, we proposed a measure that gives more weight to well-ranked coming weeks to obtain a final result. This measure is defined in the equation 3.

$$measure_C = \frac{F_{w_0} + 2 * F_{w_2} + 4 * F_{w_4} + 8 * F_{w_8}}{15} \qquad (3)$$

Finally, it is essential to mention that the chosen baseline is the majority class for the three tasks.

3 Overview of the Submitted Approaches

This section presents the results obtained by the participants for the different tasks.

3.1 Recommendation system overview

For this study, three teams have submitted their solutions for the recommendation system task.

The authors of (Callejas-Hernández et al., 2022) noted that using simpler representations (BoW) independent of the language is well suited for the recommendation task. A similar simple approach is also applied in (Morales-Murillo, Pinto-Avendaño, and Rojas López, 2022). Finally, In (Veigas-Ramírez, Martínez-Davies, and Segura-Bedmar, 2022), a Bert representation is proposed.

Table 4 shows a summary of the results obtained by each team for the recommendation system task. The MAE was used to rank participants. The approach of the GPI-CIMAT team (Callejas-Hernández et al., 2022) obtained the best performance. Surprisingly, the simple approach overcomes the Bert-based approach. This would be the result of the small relativity database. Finally, it was expected that the F-measure of the baseline would not have good results; this is evident since all the experiments surpassed the baseline in this metric, although again, the result exceeded the baseline by 0.05.

Also, Table 4 shows the result of the team that obtained the best result in last year's edition. Since this is the only task of this

edition that is identical to that of the previous year, it is possible to make this comparison. It is possible to see that no team could beat the Alumni-MCE 2GEN team of the 2021 edition (Arreola et al., 2021).

Table 5 shows the best F-measure results by class in the recommendation task. These results show that the minority classes (1 and 2) were not well represented, which is why the best result of the 2021 edition is shown. From class 3, it is possible to see that a different team obtains a good result. It is possible to observe that the GPI CIMAT team obtains the best result for class 5, which explains its better MAE result.

Something remarkable is that all the systems exceeded the baseline (BL).

3.1.1 Perfect assemble for the recommendation system task

To analyze the complementarity of the predictions by the participants' systems, we built a theoretically perfect ensemble (PA) from their runs, as calculated in (Aragón et al., 2019). We considered that a test instance was correctly classified if at least one of the participating teams classified it correctly. Also, it is proposed to combine the participating systems to create a representation based on the outputs of each system. For this, they implemented a deep learning (DL) architecture like the one proposed in (Álvarez-Carmona et al., 2022a).

From these results, it is possible to observe that the perfect ensemble performance is considerably better than the participants' approaches, suggesting that the participants' systems complement each other. This phenomenon has already appeared in this type of task and is known as The Phenomenon of Completeness over Mexican Text Classification (Álvarez-Carmona et al., 2022a).

Figure 1 shows the number of instances correctly classified by s systems. That is, when $s = 0$, all the instances that were not classified well by any system are shown, while when $s = 9$, they are the instances that were classified well by all systems. The base 2 logarithm was applied to the number of instances to observe the graph better. The systems of this and the previous edition were taken for this exercise.

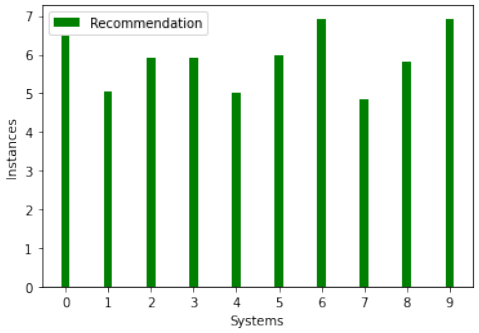


Figure 1: Instances that were correctly classified by the different numbers of possible systems for recommendation.

3.2 Sentiment analysis overview

For this study, 13 teams have submitted their solutions and descriptions for the sentiment analysis task.

For this edition, the transformers-based representation completely dominate the first places for the sentiment analysis task. The UMU team (García-Díaz et al., 2022), UC3M (Pérez Enríquez, Alonso-Mencía, and Segura-Bedmar, 2022), CIMAT MTY GTO (Gómez-Espinosa, Muñiz Sanchez, and López-Monroy, 2022), MCE (Mendoza, Ramos-Zavaleta, and Rodríguez, 2022), GPI CIMAT (Callejas-Hernández et al., 2022), CIMAT 2020 (Santibáñez Cortés et al., 2022), DCI UG (Barco, Rodríguez Rivera, and Hernández-Farías, 2022) and UCI-UC-CUJAE (Toledano-López et al., 2022) implemented solutions, mainly based on Bert. It is possible to observe that these types of methodologies are the ones that obtain the best results since the lowest result, of what transformers applied, is 0.84 when the best result is 0.89, that is, the results are very close to each other.

On the other hand, the rest of the works proposed more straightforward methods. ESCOM-IPN-LCD Team (Alcibar-Zubillaga et al., 2022) proposes a Logistic Regression classifier to train two models, one for polarity prediction and the other for the attraction type prediction. UPTC-UDLAP Team (Rico-Sulayes and Monsalve-Pulido, 2022) applies the Naive Bayes Multinomial algorithm to represent a supervised classification approach. The team uses unigrams, bigrams, and trigrams as features. The unsupervised classification was carried out by computing the total polarity of the opinions in an intensity spectrum according to the scale of the data using context embeddings. The SENA Team (Jurado-Buch, Bustio-Martínez, and Álvarez-Carmona, 2022) uses

Rank	Country	Institute	Team	MAE	Accuracy	F-measure
PA	-	-	-	0.28	86.58	0.66
DL	-	-	-	0.31	76.45	0.52
2021	Mex	CIMAT	Alumni-MCE 2GEN$_{Run1}$	0.31	77.28	0.50
1st	Mex	CIMAT	GPI-CIMAT	**0.69**	52.12	0.19
2nd	Mex	BUAP	LKEBUAP$_{Run_{k1-13-k2-18}}$	0.70	46.64	**0.22**
-	Mex	BUAP	LKEBUAP$_{Run_{k1-14-k2-18}}$	0.70	45.88	0.21
-	Mex/Che	CIMAT	GPI-CIMAT$_{Run_1}$	0.72	**53.66**	0.17
3rd	Esp	UC3M	UC3M-DEEPNLP$_{Run_1}$	0.72	48.89	**0.22**
BL			Majority Class	0.74	53.30	0.13
-	Esp	UC3M	UC3M-DEEPNLP$_{Run_2}$	0.75	52.71	0.13

Table 4: Performance for the Recommendation System task.

F-measure class	Best result	Team
1	0.32	Alumni-MCE 2GEN$_{Run1}$(2021)
2	0.24	Alumni-MCE 2GEN$_{Run1}$(2021)
3	0.14	UC3M-DEEPNLP$_{Run_1}$
4	0.37	LKEBUAP$_{Run_{k1-13-k2-18}}$
5	0.69	GPI-CIMAT

Table 5: Performance for the Recommendation System task per class.

a representation based on Topics extracted by LDA and classified with simple Deep Learning architecture. Finally, DevsEx-Machina (Rivas-Álvarez et al., 2022) proposes to extract all the terms in each class from one to four words (1...4-grams) as polarity characteristics. Also, they perform a chain of translations of the opinions, from Spanish to other languages and back to Spanish, to obtain meanings and synonymous terms.

It is interesting that despite being more straightforward, some of the results of the proposals that are not based on Transformers obtain close values. This seems ideal for environments with limited memory, time, or data.

Table 6 shows a summary of the results obtained by each team for the sentiment analysis task[2]. The UMU team obtained the best result, although the difference with UC3M is 0.002. Due to the closeness of the results, it is possible that there is no statistical significance between all the methods based on transformers.

Table 7 shows the best F-measure results by class in the sentiment analysis task. Interestingly, the UMU team does not get the best results for any polarity class. However, it is the best team for all three attraction classes. The DCI UG team obtains the best results for classes 1 and 2, which are the most difficult to classify due to their clear imbalance. UC3M obtains the best result for class 3, and finally, MCE, in its two attempts, obtains the best results for the majority of classes.

[2]For systems with *, the authors did not send the system's description.

3.2.1 Perfect assemble for the sentiment analysis task

As in the section 3.1.1, the complementarity of the systems was analyzed for the sentiment analysis task. We calculated the perfect assemble.

Table 6 also shows the perfect assemble (PA) result and the Deep Learning combination systems (DL).

As in the recommendation task, it is possible to observe that the perfect ensemble performance is considerably better than the UMU approach, suggesting that the participants' systems are complementary to each other again, with an error result very close to zero. The DL approach improves the best result obtained by the UMU team.

Figure 2 shows the number of instances correctly classified by s systems similar to the Figure 1. The color green is the polarity instances, whereas the color yellow represents the attraction instances.

3.2.2 Interesting opinions

PA approach got only six incorrect instances for the attractiveness detection task. The pattern of these instances is tourists talking about a hotel restaurant or vice versa, which confuses all systems. For example:

Este lugar era estupendo. Un montón de

Rank	Country	Institute	Team	Measure$_S$	MAE$_P$	F$_A$
PA	-	-	-	0.98	0.03	0.99
DL	-	-	-	0.91	0.19	0.99
1st	Esp	UMU	UMU-Team$_{Run_1}$	**0.89**	**0.25**	**0.99**
2nd	Esp	UC3M	UC3M$_{Run_1}$	**0.89**	0.26	0.98
3rd	Mex	CIMAT	CIMAT MTY-GTO$_{Run_1}$	0.88	0.26	0.98
HM	Mex	ITESM	MCE-Team$_{Run_2}$	0.88	0.26	0.98
-	Mex	ITESM	MCE-Team$_{Run_1}$	0.88	0.26	0.98
-	Esp	UMU	UMU-Team$_{Run_2}$	0.88	0.27	0.98
HM	Mex/Che	CIMAT	GPI-CIMAT$_{Run_1}$	0.88	0.26	0.98
HM	Mex	CIMAT	CIMAT2020$_{Beto_{Run_1}}$	0.88	0.27	0.97
HM	Mex	INAOE	DCI-UG$_{Run_1}$	0.87	0.26	0.96
HM	Cub/Bel	UCI	UCI-UC-CUJAE$_{Run_2}$	0.87	0.30	0.97
-	Cub/Bel	UCI	UCI-UC-CUJAE$_{Run_1}$	0.86	0.30	0.97
-	Mex	CIMAT	CIMAT2020$_{Run_2}$	0.86	0.31	0.97
-	Mex	INAOE	DCI-UG$_{Run_2}$	0.86	0.30	0.96
HM	Mex	IPN	ESCOM-IPN-IIA*$_{Run_2}$	0.85	0.32	0.96
-	Mex/Che	CIMAT	GPI-CIMAT$_{Run_1}$	0.84	0.28	0.91
HM	Mex	IPN	ESCOM-IPN-LCD$_{Run_2}$	0.84	0.35	0.94
-	Mex	IPN	ESCOM-IPN-IIA*$_{Run_1}$	0.83	0.34	0.92
HM	Mex/Col	UDLAP	UPTC-UDLAP$_{Run_1}$	0.82	0.44	0.96
HM	Col/Mex	SENA	SENA Team	0.80	0.47	0.92
HM	Mex	UAEM	DevsExMachina$_{Run_2}$	0.70	0.63	0.79
-	Mex	UAEM	DevsExMachina$_{Run_1}$	0.66	0.97	0.82
-	Mex	IPN	ESCOM-IPN-LCD$_{Run_1}$	0.59	0.85	0.65
-	Mex/Col	UDLAP	UPTC-UDLAP$_{Run_2}$	0.54	0.54	0.43
BL	-	-	*Majority class*	0.45	0.47	0.23

Table 6: Performance for the Sentiment Analysis task.

F-measure class	Best result	Team
1	0.61	DCI-UG$_{Run_1}$
2	0.37	DCI-UG$_{Run_1}$
3	0.50	UC3M]
4	0.48	MCE-Team$_{Run_1}$
5	0.88	MCE-Team$_{Run_2}$
Attractive	0.99	UMU-Team$_{Run_1}$
Hotel	0.99	UMU-Team$_{Run_1}$
Restaurant	0.98	UMU-Team$_{Run_1}$

Table 7: Performance for the Sentiment Analysis task per class.

opciones y gran comida fresca. El desayuno buffet era grandes mucha fruta fresca.

This instance is a hotel; however, the opinion refers to the food.

Another example is:

El hotel está increíble, pero resaltó el excelente servicio en insu sky bar, muchas gracias al capitán Iván, y a su staff Heriberto, Gabriel, Luis, Isidoro y Hugo, las bebidas de Gerardo y Alexis increíbles y la cocina un placer ! En el área de alberca al señor Wenceslao! Muchas gracias por todo !

Which, although it is inside a hotel, is a restaurant.

None of these instances have the attractive label.

3.3 Semaphore covid prediction results

For this last task, six systems were received from 4 teams.

MCE team(Ramos-Zavaleta and Rodríguez, 2022) presents an approach based on features extracted directly from the news and the other applying transfer learning. First, they propose a system based on CorEx topics, and as a second attempt, they propose a system based on Bert.

Arandanito team (Carmona-Sánchez, Carmona, and Álvarez-Carmona, 2022) proposes a method based on topic extraction. This topic-based representation is applied to a series of linear regressions, which serve as input for simple deep learning architecture.

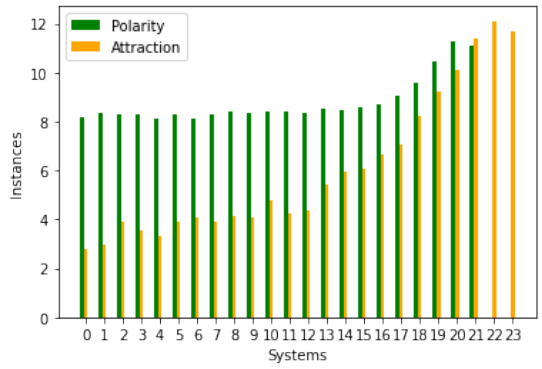


Figure 2: Instances that were correctly classified by the different numbers of possible systems for sentiment analysis.

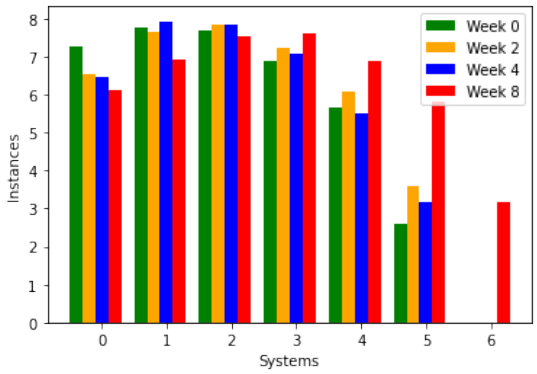


Figure 3: Instances that were correctly classified by the different numbers of possible systems for semaphore prediction.

The last Team (Romero-Cantón et al., 2022) proposes an approach based on weighing representative words as features extracted directly from the text news. Those words were weighed by using the Mutual information (MI) measure.

Table 8 shows the results of the participating teams. It can be seen that both MCE and Arandanito have a very close results to each other. Curiously, both approaches are topic-based.

It can be seen that the best results are obtained for week 2 in the future. That is, two weeks after the news was published. However, the results of weeks 4 and 8, considering the imbalance and that there are four classes, are competitive.

Like the other tasks, all the participants managed to pass the Baseline (BL).

Table 9 shows the best results for each class for each evaluated week.

For week 0, it is possible to see that MCE obtains all the best results. However, from week 2, it can be seen that Arandanito obtains the best result for the Red class. This is the most challenging class because it is the minority class. For all other classes, MCE gets the best result.

3.3.1 Perfect assemble for the semaphore prediction task

Table 8 also shows the perfect assembly and the combination of the systems, like the other two tasks.

The perfect ensemble is much higher than the best of the individual results, which indicates that these systems also complement each other.

On the other hand, the combination of the two systems once again enhances the individual best value.

Figure 3 shows the number of instances correctly classified by s systems similar to the Figure 1. The color green is Week 0, yellow for 2, blue for 4, and red for 8.

For more details of the results of both tasks, it is possible to go to the following web page: `https://sites.google.com/cicese.edu.mx/rest-mex-2022/results`.

4 *Conclusions*

This paper described the design and results of the Rest-Mex shared task collocated with IberLef 2022. Rest-Mex stands for *Recommendation system, Sentiment analysis and covid semaphore prediction in Spanish tourists text for Mexican places*. For the three tasks, 18 teams participated. Mainly, the members of these teams come from institutes in countries such as Mexico, Spain, Cuba, Colombia, Belgium, and Switzerland. Thirty-five different systems were received to be evaluated to solve each of the three tasks proposed in the Rest-Mex 2021.

For the recommendation task, a tourist's satisfaction with a recommendation of a tourist place for the state of Nayarit in Mexico was evaluated. The best MAE result obtained was that of (Callejas-Hernández et al., 2022), which belongs to the CIMAT of Mexico. This team proposed a simple system based on BoW. Although all the participating systems outperformed the Baseline, no system was able to obtain better results than the 2021 edition. This result indicates that this task still has many challenges to be solved.

Rank	Country	Institute	Team	Measure$_C$	F$_{w_0}$	F$_{w_2}$	F$_{w_4}$	F$_{w_8}$
PA	-	-	-	0.84	0.92	0.88	0.87	0.81
DL	-	-	-	0.67	0.74	0.76	0.71	0.63
1st	Mex	ITESM	MCE-Team$_{Run_2}$	**0.49**	**0.56**	0.52	0.46	**0.48**
2nd	Mex	BUAP	Arandanito	0.48	0.33	**0.56**	**0.51**	0.46
-	Mex	ITESM	MCE-Team$_{Run_1}$	0.32	0.33	0.34	0.32	0.32
-	Mex	UNAM	*ML-Team$_{Run_2}$	0.24	0.25	0.27	0.23	0.24
	Mex	UNAM	*ML-Team$_{Run_1}$	0.22	0.20	0.22	0.22	0.23
HM	Mex	UAN	The Last	0.17	0.18	0.18	0.18	0.16
BL			Majority Class	0.12	0.13	0.13	0.12	0.12

Table 8: Performance for the semaphore prediction task.

F-measure class	Best result	Team	F-measure class	Best result	Team
Red$_{w_0}$	0.38	MCE-Team$_{Run_2}$	Red$_{w_2}$	0.37	Arandanito
Orange$_{w_0}$	0.66	MCE-Team$_{Run_2}$	Orange$_{w_2}$	0.68	MCE-Team$_{Run_2}$
Yellow$_{w_0}$	0.45	MCE-Team$_{Run_2}$	Yellow$_{w_2}$	0.52	MCE-Team$_{Run_2}$
Green$_{w_0}$	0.73	MCE-Team$_{Run_2}$	Green$_{w_2}$	0.74	MCE-Team$_{Run_2}$
Red$_{w_4}$	0.39	Arandanito	Red$_{w_8}$	0.2	Arandanito
Orange$_{w_4}$	0.66	MCE-Team$_{Run_2}$	Orange$_{w_8}$	0.65	MCE-Team$_{Run_2}$
Yellow$_{w_4}$	0.47	MCE-Team$_{Run_2}$	Yellow$_{w_8}$	0.55	MCE-Team$_{Run_2}$
Green$_{w_4}$	0.71	MCE-Team$_{Run_2}$	Green$_{w_8}$	0.73	MCE-Team$_{Run_2}$

Table 9: Performance for the Semaphore Prediction task per class.

The sentiment analysis task aimed to identify the polarity and precedence of an opinion made about a Mexican tourist destination. The polarity was evaluated with MAE while the origin with F-Measure. The team that got the best performance was (García-Díaz et al., 2022). This team represents the University of Murcia in Spain. They proposed a method based on Bert. Other teams that also implemented Bert obtained results very close to first place. This is further evidence of the importance of transformers in textual classification tasks. Also, the results indicate that distinguishing between opinions of hotels, restaurants, and attractions is a task that can have very high results, close to 100 %.

The task of determining the semaphore covid was a novelty introduced for this year's edition. Based on the news regarding covid, this task consists of determining the color of the epidemiological semaphore for weeks 0, 2, 4, and 8 in the future based on the news publications. The best result obtained for this task can be seen in (Ramos-Zavaleta and Rodríguez, 2022). This team comes from ITESM of Mexico. Their solution is based mainly on extracting topics, although they obtains his best result with Bert. Best results are achieved when ranked 2 weeks into the future; however, results for 4 weeks also seem competitive. The results at 8 weeks suffer a drop in the classification.

Finally, it is shown that there is significant complementarity between the participating systems. In other evaluation forums, attempts have been made to mix the participating systems to obtain better results (Álvarez-Carmona et al., 2018), taking the proposal to use a simple deep learning architecture (Álvarez-Carmona et al., 2022a), it was possible to improve the best results of the three tasks. However, the perfect theoretical ensemble is still above the results obtained.

Acknowledgements

Our special thanks go to all of Rest-Mex's participants, the organizers, and their institutions.

References

Alcibar-Zubillaga, J., Y. De-Luna Ocampo, I. Pacheco-Castillo, K. Ramirez-Mendez, J. P. M. Sainz-Takata, and O. Juárez Gambino. 2022. Participation of escom's data science group at rest-mex 2022: Sentiment analysis task. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Álvarez-Carmona, M. Á., R. Aranda, S. Arce-Cardenas, D. Fajardo-Delgado,

R. Guerrero-Rodríguez, A. P. López-Monroy, J. Martínez-Miranda, H. Pérez-Espinosa, and A. Y. Rodríguez-González. 2021. Overview of rest-mex at iberlef 2021: recommendation system for text mexican tourism. *Procesamiento del Lenguaje Natural*.

Álvarez-Carmona, M. Á., R. Aranda, R. Guerrero-Rodríguez, A. Y. Rodríguez-González, and A. P. López-Monroy. 2022a. A combination of sentiment analysis systems for the study of online travel reviews: Many heads are better than one. *Computación y Sistemas*, 26(2).

Álvarez-Carmona, M. A., R. Aranda, A. Y. Rodríguez-González, L. Pellegrin, and H. Carlos. 2022b. Classifying the mexican epidemiological semaphore colour from the covid-19 text spanish news. *Journal of Information Science*.

Álvarez-Carmona, M. Á., E. Guzmán-Falcón, M. Montes-y Gómez, H. J. Escalante, L. Villasenor-Pineda, V. Reyes-Meza, and A. Rico-Sulayes. 2018. Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), seville, spain*, volume 6.

Álvarez-Carmona, M. Á., E. Villatoro-Tello, L. Villaseñor-Pineda, and M. Montes-y Gómez. 2022. Classifying the social media author profile through a multimodal representation. In *Intelligent Technologies: Concepts, Applications, and Future Directions*. Springer, pages 57–81.

Alvarez-Carmona, M. Á. and R. Aranda. 2022. Determinación automática del color del semáforo mexicano del covid-19 a partir de las noticias.

Aragón, M. E., M. A. Álvarez-Carmona, M. Montes-y Gómez, H. J. Escalante, L. V. Pineda, and D. Moctezuma. 2019. Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In *IberLEF@ SEPLN*, pages 478–494.

Arce-Cardenas, S., D. Fajardo-Delgado, M. Á. Álvarez-Carmona, and J. P. Ramírez-Silva. 2021. A tourist recommendation system: a study case in mexico. In *Mexican International Conference on Artificial Intelligence*, pages 184–195. Springer.

Arreola, J., L. Garcia, J. Ramos-Zavaleta, and A. Rodríguez. 2021. An embeddings based recommendation system for mexican tourism. submission to the rest-mex shared task at iberlef 2021. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2021), CEUR WS Proceedings*.

Barco, G. M., G. E. Rodríguez Rivera, and D.-I. Hernández-Farías. 2022. Sentiment analysis in spanish reviews: Dataket submission on rest-mex 2022. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings*.

Callejas-Hernández, C., E. Rivadeneira-Pérez, F. Sánchez-Vega, A. P. López-Monroy, and E. Villatoro-Tello. 2022. The winning approach for the recommendation systems shared task @rest_mex 2022. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings*.

Carmona-Sánchez, G., A. Carmona, and M. A. Álvarez-Carmona. 2022. Combining linear regressions to determine the future of the covid in mexico from the news. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings*.

García-Díaz, J. A., M. A. Rodríguez-García, F. García-Sánchez, and R. Valencia-García. 2022. Umuteam at rest-mex 2022: Polarity prediction using knowledge integration of linguistic features and sentence embeddings based on transformers. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings*.

Gómez-Espinosa, V., V. Muñiz Sanchez, and A. P. López-Monroy. 2022. Automl and ensemble transformers for sentiment analysis in mexican tourism texts. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings*.

Guerrero-Rodriguez, R., M. Á. Álvarez-Carmona, R. Aranda, and A. P. López-Monroy. 2021. Studying online travel reviews related to tourist attractions using nlp methods: the case of guanajuato, mexico. *Current issues in tourism*, pages 1–16.

Jurado-Buch, J. D., L. Bustio-Martínez, and M. A. Álvarez-Carmona. 2022. The role of the topics for the sentiment analysis task on a mexican tourist collection. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Mendoza, D., J. Ramos-Zavaleta, and A. Rodríguez. 2022. A transfer learning model for polarity in touristic reviews in spanish from tripadvisor. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Morales-Murillo, V. G., D. Pinto-Avendaño, and F. Rojas López. 2022. A hybrid recommender model based on information retrieval for mexican tourism text in rest-mex 2022. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Pérez Enríquez, M., J. Alonso-Mencía, and I. Segura-Bedmar. 2022. Transformers approach for sentiment analysis: Classification of mexican tourists reviews from tripadvisor. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Ramos-Zavaleta, J. and A. Rodríguez. 2022. A mexico's covid traffic light color prediction system based on mexican news. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Rico-Sulayes, A. and J. Monsalve-Pulido. 2022. A proposal and comparison of supervised and unsupervised classification techniques for sentiment analysis in tourism data. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Rivas-Álvarez, J. C., R. A. García-Hernández, S. I. Medina-Martínez, A. M. Martínez-Ortiz, N. Hernández-Castañeda, J. E. Ruiz-Melo, A. Hernández-Castañeda, and Y. Nikolaevna-Ledeneva. 2022. Devs-ex-machina at rest-mex 2022 opinion mining of the mexican tourism sector through sets of normalized n-grams. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Romero-Cantón, A., A. Diaz-Pacheco, R. Aranda, and P. Ramírez-Silva. 2022. Mexican epidemiological semaphore color prediction by means of mutual information features. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Santibáñez Cortés, E., A. Carrillo-Cabrera, Y. A. Castillo-Castillo, D. A. Moctezuma-Ochoa, and V. H. Muñíz Sánchez. 2022. Bert model and data augmentation for sentiment analysis in tourism reviews for mexican spanish language. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Toledano-López, O. G., J. Madera, H. González, A. Simón-Cuevas, T. Demeester, and E. Mannens. 2022. Fine-tuning mt5-based transformer via cma-es for sentiment analysis. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Veigas-Ramírez, S., D. Martínez-Davies, and I. Segura-Bedmar. 2022. Recommendation system rest-mex 2022 for mexican tourism using natural language processing. In *Proceedings of the Third Workshop for Iberian Languages Evaluation Forum (IberLEF 2022), CEUR WS Proceedings.*

Información General

Información para los Autores

Formato de los Trabajos

- La longitud máxima admitida para las contribuciones será de 10 páginas DIN A4 (210 x 297 mm.), incluidas referencias y figuras.
- Los artículos pueden estar escritos en inglés o español. El título, resumen y palabras clave deben escribirse en ambas lenguas.
- El formato será en Word ó LaTeX

Envío de los Trabajos

- El envío de los trabajos se realizará electrónicamente a través de la página web de la Sociedad Española para el Procesamiento del Lenguaje Natural (http://www.sepln.org)
- Para los trabajos con formato LaTeX se mandará el archivo PDF junto a todos los fuentes necesarios para compilación LaTex
- Para los trabajos con formato Word se mandará el archivo PDF junto al DOC o RTF
- Para más información http://www.sepln.org/la-revista/informacion-para-autores

Información Adicional

Funciones del Consejo de Redacción

Las funciones del Consejo de Redacción o Editorial de la revista SEPLN son las siguientes:

- Controlar la selección y tomar las decisiones en la publicación de los contenidos que han de conformar cada número de la revista
- Política editorial
- Preparación de cada número
- Relación con los evaluadores y autores
- Relación con el comité científico

El consejo de redacción está formado por los siguientes miembros

L. Alfonso Ureña López (Director)
Universidad de Jaén
laurena@ujaen.es
Patricio Martínez Barco (Secretario)
Universidad de Alicante
patricio@dlsi.ua.es
Manuel Palomar Sanz
Universidad de Alicante
mpalomar@dlsi.ua.es
Felisa Verdejo Maíllo
UNED
felisa@lsi.uned.es

Funciones del Consejo Asesor

Las funciones del Consejo Asesor o Científico de la revista SEPLN son las siguientes:

- Marcar, orientar y redireccionar la política científica de la revista y las líneas de investigación a potenciar
- Representación
- Impulso a la difusión internacional
- Capacidad de atracción de autores
- Evaluación
- Composición
- Prestigio
- Alta especialización
- Internacionalidad

El Consejo Asesor está formado por los siguientes miembros:

Xabier Arregi	Universidad del País Vasco (España)
Manuel de Buenaga	Universidad de Alcalá (España)
José Camacho Collados	Cardiff University (Reino Unido)
Sylviane Cardey-Greenfield	Centre de recherche en linguistique et traitement automatique des langues (Francia)
Irene Castellón	Universidad de Barcelona (España)
Arantza Díaz de Ilarraza	Universidad del País Vasco (España)
Antonio Ferrández	Universidad de Alicante (España)
Koldo Gojenola	Universidad del País Vasco (España)
Xavier Gómez Guinovart	Universidad de Vigo (España)
José Miguel Goñi	Universidad Politécnica de Madrid (España)
Inma Hernaez	Universidad del País Vasco (España)

Cartas al director

Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, Edificio A3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Más información

Para más información sobre la Sociedad Española del Procesamiento del Lenguaje Natural puede consultar la página web http://www.sepln.org.
Si desea inscribirse como socio de la Sociedad Española del Procesamiento del Lenguaje Natural puede realizarlo a través del formulario web que se encuentra en esta dirección http://www.sepln.org/sepln/la-sociedad
Los números anteriores de la revista se encuentran disponibles en la revista electrónica: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/archive
Las funciones del Consejo de Redacción están disponibles en Internet a través de http://www.sepln.org/la-revista/consejo-de-redaccion

Las funciones del Consejo Asesor están disponibles Internet a través de la página http://www.sepln.org/la-revista/consejo-asesor
La inscripción como nuevo socio de la SEPLN se puede realizar a través de la página http://www.sepln.org/sepln/inscripcion-para-nuevos-socios