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.

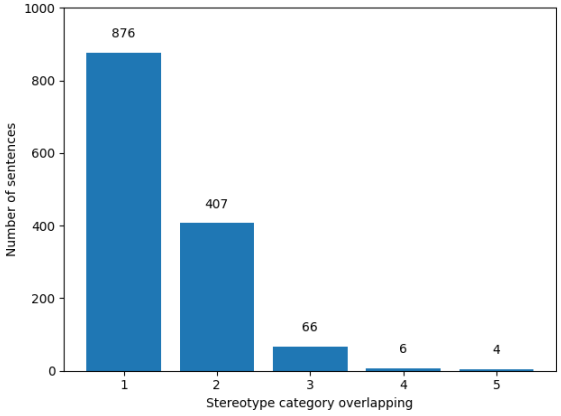

Figure 1: Multi-label distribution.

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-

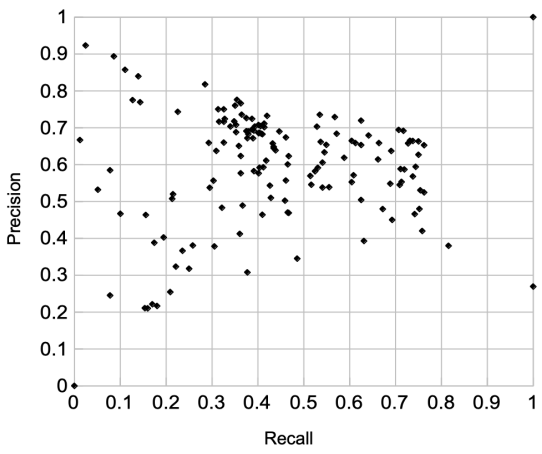

Figure 2: Precision vs. Recall in subtask 1.

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.

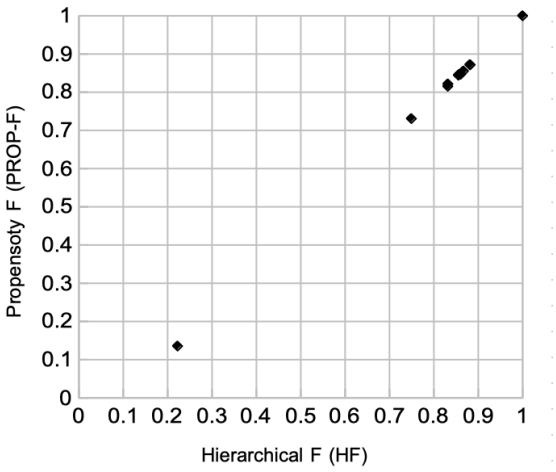

Figure 3: Hierarchical F-measure vs. Propensity F-measure in subtask 2.

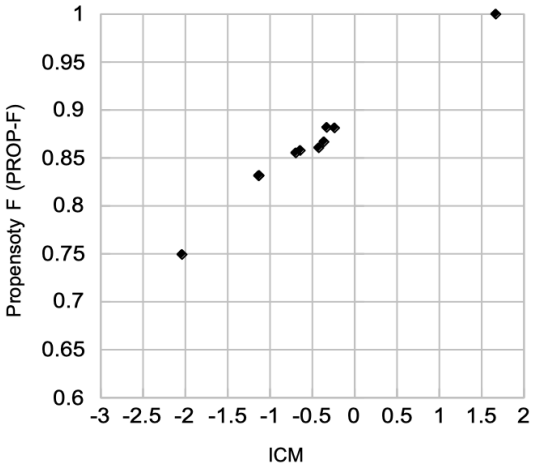

Figure 4: ICM vs. Propensity F-measure in subtask 2.

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).

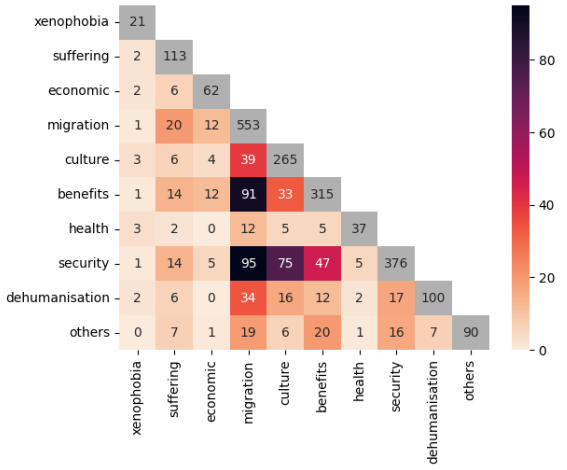

Figure 5: Heatmap representation of the sentence-level co-ocurrence of stereotype categories with the occurrence count of each category coloured in gray.