Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 291-303 recibido 05-12-2024 revisado 29-01-2025 aceptado 11-02-2025

Lexical Complexity Assessment of Spanish in Ecuadorian Public Documents

Evaluación de Complejidad Léxica del Español en Documentos Públicos Ecuatorianos

Jenny Ortiz-Zambrano,[1] **César Espin-Riofrio,**[1] **Arturo Montejo-Ráez**[2]
[1]Universidad de Guayaquil, Guayaquil, Ecuador
[2]Universidad de Jaén, Jaén, Spain
{jenny.ortizz, cesar.espinr}@ug.edu.ec, amontejo@ujaen.es

Abstract: This study presents a comprehensive assessment of lexical complexity (LC) in texts from Ecuadorian public institutions, with a particular focus on the development and application of advanced natural language processing (NLP) techniques. The analysis includes a comparative evaluation of several models and approaches applied to the GovAI*Ec* corpus, a recently developed collection of Ecuadorian government texts. The study examines the impact of incorporating linguistic features and varying the number of training epochs, providing an in-depth analysis of their contribution to model performance. Furthermore, a practical and accessible solution is proposed through a web platform designed to facilitate the understanding of complex words in public documents, which often hinder the successful execution of bureaucratic processes. This work aims to improve interactions with government systems by promoting more efficient and comprehensible communication. The best performance was achieved with bert-base-spanish-wwm-uncased, combining linguistic features and encodings, with a MAE = 0.1551. The results indicate that linguistic features are essential to improve performance, suggesting that hybrid approaches are more effective than those based solely on deep learning.
Keywords: Lexical Complexity Prediction, Linguistic Features, Public Documents, Deep Learning.

Resumen: Este estudio presenta una evaluación integral de la complejidad léxica (CL) en textos de instituciones públicas ecuatorianas, con un enfoque particular en el desarrollo y aplicación de técnicas avanzadas de procesamiento del lenguaje natural (PLN). El análisis incluye una evaluación comparativa de varios modelos y enfoques aplicados al corpus GovAI*Ec*, una colección recientemente desarrollada de textos gubernamentales ecuatorianos. El estudio examina el impacto de la incorporación de características lingüísticas y la variación del número de épocas de entrenamiento, proporcionando un análisis profundo de su contribución al rendimiento del modelo. Además, se propone una solución práctica y accesible a través de una plataforma web diseñada para facilitar la comprensión de palabras complejas en documentos públicos, que a menudo obstaculizan la ejecución exitosa de procesos burocráticos. Este trabajo tiene como objetivo mejorar las interacciones con los sistemas gubernamentales promoviendo una comunicación más eficiente y comprensible. El mejor rendimiento se alcanzó con bert-base-spanish-wwm-uncased, combinando características lingüísticas y codificaciones, con un MAE = 0.1551. Los resultados indican que las características lingüísticas son esenciales para mejorar el rendimiento, sugiriendo que los enfoques híbridos son más efectivos que los basados únicamente en aprendizaje profundo.
Palabras clave: Predicción de la Complejidad Léxica, Características Lingüísticas, Documentos Públicos, Aprendizaje Profundo.

1 Introduction

There are different groups of readers, such as language learners (Rets and Rogaten, 2021), people with cognitive disabilities (Licardo, Volčanjk, and Haramija, 2021), also those with low reading proficiency, who may face

ISSN 1135-5948 DOI 10.26342/2025-74-20

difficulties in understanding texts. It is important to emphasize that access to comprehensible texts is a fundamental right that has been increasingly recognized by international institutions and legislation (Bott et al., 2024).

On the other hand, many people encounter significant obstacles in understanding texts related to public administration (Yuan et al., 2023). These challenges may be due to difficulties in deciphering long sentences, technical jargon, uncommon terminology or complex linguistic structures. Such obstacles directly affect people from the above-mentioned target groups. Even highly educated people, such as university students in various fields of study, may be among those affected by reading difficulties (Alarcón, Moreno, and Martínez, 2020). Public institutions are not exempt from this reality. The content of texts intended for citizens often includes complex vocabulary, making interpretation difficult and hindering users' ability to initiate activities or administrative procedures (Roundy, Trussel, and Davenport, 2023).

Text simplification (TS) aims to reduce the complexity of a sentence without losing its meaning, thus making it easier to understand, especially for people with cognitive disabilities (Paetzold and Specia, 2017). Within TS, lexical simplification (LS) focuses on replacing complex words with simpler alternatives, thus limiting changes to the lexical level. LS tasks include complex word identification (CWI) and substitute generation (SG) (Tan et al., 2024). A related and similar task to CWI is Lexical Complexity Prediction (LCP) (Shardlow et al., 2021), which provides an estimate of the lexical difficulty level of each target word, rather than simply making a binary classification on whether a word should be replaced or not (Bott et al., 2024).

Integrating large language models (LLMs) into domain-specific applications is a key research challenge (Jebali et al., 2024). Deep learning approaches have become the state-of-the-art in numerous natural language processing (NLP) tasks, including lexical complexity prediction (LCP) (Singh and Mahmood, 2021), (Nandy et al., 2021), (North et al., 2024). Two of the main features that have driven this advancement are the self-attention mechanism, especially the Transformer architecture (Vaswani, 2017), and the adoption of unsupervised pre-training methods (Sarzynska-Wawer et al., 2021), (Howard and Ruder, 2018), (Devlin et al., 2019), which take advantage of large volumes of unlabeled text corpora. Transformer-based models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu, 2019), have set a precedent NLP. Building on these advances, latest language models (e.g., GPT-4) focus on complex task solving, with billions of parameters, have demonstrated state-of-the-art performance on a variety of tasks, as evidenced by recent findings in multiple NLP studies (Minaee et al., 2024), such as GPT-4- Turbo, a state-of-the-art model recognized for its outstanding ability to understand language and its capacity to generate high-quality text (Rampal et al., 2024).

This paper presents an approach to lexical complexity prediction applied to the spanish corpus called "GovAI*Ec*". Spanish, as the second most spoken language globally, has considerable relevance in both communication and culture (Bylund, Khafif, and Berghoff, 2024). Although spanish is one of the most widely spoken languages, it faces challenges in the field of natural language processing (NLP) due to the disparity in available resources compared to english. The quantity and diversity of datasets, tools, and pre-trained models for spanish are often limited in comparison to those available for English, making it more challenging to train or evaluate spanish language models (Cañete et al., 2023). This includes labeled text corpora and datasets to effectively train models. LLMs have shown significant achievements in NLP tasks. However, integrating linguistic features together with encodings could optimize their precision in predicting and identifying complex words (Ortiz-Zambrano, Espín-Riofrío, and Montejo-Ráez, 2024).

Original contribution of our work

1. Our study focuses on a relatively unexplored area. To date, this study represents the first systematic approach to address lexical complexity in texts from ecuadorian public institutions. While there are previous studies in the area of natural language processing applied to LLMs, this work stands out for its focus on a labeled corpus from Ecuadorian in-

stitutional texts.

2. Although deep learning models tend to adopt end-to-end approaches, by incorporating these linguistic features one can add meaningful and relevant information to the encodings of linguistic models (Ortiz-Zambrano, Espín-Riofrío, and Montejo-Ráez, 2024).

   The objective of this research work is *to contribute to the advancement of knowledge in the field of lexical simplification by evaluating the performance of algorithms based on fine-tuned Transformers, thus demonstrating the effectiveness of integrating linguistic features to improve the performance of models.*

3. An evaluation was carried out through a comparative analysis of several models and approaches applied to the GovAI*Ec* corpus (a new corpus of Ecuadorian state texts). A notable feature of this dataset is that it includes documents from various government institutions, representing a broad user base and areas where customer service is a critical factor. This underlines the potential of the proposals to optimize text management and comprehension in spanish language government contexts.

4. This study considers both homogeneous conditions and differentiated scenarios. In particular, runs were performed that add linguistic features to the corpus to assess their impact on performance, as well as runs without these features, allowing a thorough analysis of their contribution to the tasks evaluated.

5. In addition, an accessible and practical solution is proposed for users through a web platform GovAI*Ec*[1]. This tool aims to facilitate the understanding of complex words contained in public documents, which often hinder the successful execution of bureaucratic processes, thus contributing to a more efficient and understandable interaction with the government system.

The rest of the article is organized as follows:

Section 2 presents related work in the area of lexical simplification, focusing on systems based on lexical complexity metrics for spanish and linguistic models. Section 3 describes the methodology applied in this study. Section 4 describes the experiments performed and the results obtained, along with an analysis of these results. Section 5 discusses the implications of the findings. Section 6 summarizes the main conclusions of the research, while Section 7 offers perspectives on future directions following the same line of research.

2 Related work

In the past, innovative lexical simplification methods relied on complex systems composed of multiple components, each of which required deep technical knowledge and precise integration to achieve optimal performance (Aumiller and Gertz, 2023). However, recent advances in deep learning, especially with the emergence of large language models (LLMs), have significantly simplified this process. These rapidly tunable models have reinvigorated interest in lexical simplification (North et al., 2023b). Advanced models such as BERT (Devlin et al., 2018), RoBERTa (Liu, 2019), and GPT-3 (Ortiz-Zambrano, Espin-Riofrio, and Montejo-Ráez, 2023), (North et al., 2024) among others, have demonstrated remarkable ability to automatically generate, select, and rank candidate substitutions, far outperforming traditional approaches (North et al., 2024).

Since the introduction of the Transformer architecture in 2017, the field of natural language (Vaswani, 2017), which leverages large unlabeled corpora (Cañete et al., 2020). With the rise of transfer learning and pre-trained language models, deep learning-based solutions have significantly outperformed traditional shallow machine learning approaches. Models such as BERT and XLM-RoBERTa have established themselves as state-of-the-art benchmarks in a variety of natural language processing tasks (Yaseen et al., 2021).

The first BERT model pre-trained exclusively on spanish data was successfully developed, and both the model and the training corpus and evaluation benchmarks were made publicly available (Cañete et al., 2020). Later, a RoBERTa model trained on an even

[1] GovAI*Ec*
Available at `https://www.govaieasy.com/`

larger and more diverse set of spanish texts was presented. This model showed outstanding performance on various NLP tasks and benchmarks, setting a new standard in most of them, although it was not evaluated on the lexical complexity prediction (LCP) task (Gutiérrez-Fandiño et al., 2021a).

To improve cognitive accessibility in university educational texts, a pre-trained multilingual BERT model was applied, designed to identify complex words using word-by-word generated vectors. Although the model's performance did not outperform other methods in the ALexS 2020 competition, the approach stood out for its supervised architecture (Alarcón, Moreno, and Martínez, 2020).

In 2021, a shared SemEval task (Shardlow et al., 2021) directly addressed the problem of lexical complexity prediction, building on previous related work (Shardlow, Cooper, and Zampieri, 2020b). Currently, the Multilingual Lexical Simplification Pipeline 2024 challenge presents a new proposal that combines elements of lexical complexity prediction (LCP) and lexical simplification (LS), extending the approach to the multilingual domain (Shardlow et al., 2024).

3 Methodology

The dataset used in our experiments is a newly created corpus called GovAI*Ec*, which is described in the following section. In our proposed approach, linguistic features extracted from the texts are combined with embeddings of the input text. The embeddings generated by transformer-based encoder models are concatenated with the linguistic features and passed to a final classification layer to predict the complexity score of the target term.

3.1 The GovAI*Ec* corpus

The GovAI*Ec*[2] corpus was created as part of an independent research effort. GovAI*Ec* is a contribution to research on the identification of complex words in state documents in Ecuador, specifically from public institutions with the largest number of users. GovAI*Ec* is a Lexical Complexity Corpus for spanish in Ecuadorian public documents that offers a collection of 1,500 texts obtained mainly from two sources: notifications and instructions for administrative procedures that users receive through email or find on the websites of public institutions. This corpus has a total of 7,813 identified complex words and a total of 12,095 annotations which were made by several annotators.

In this paper, we use the corpus to evaluate the performance of several large-scale language models (LLMs) in predicting lexical complexity. Words considered "complex" were selected based on criteria perceived by the annotators. Although some words may seem common in general contexts, they were labeled as complex in specific contexts where their use was difficult to understand for annotators with basic or intermediate educational levels. This corpus is essential for two fundamental reasons. First, it allows the identification of terms that are difficult for users who carry out administrative processes to understand. Second, it provides a valuable resource for the scientific community, allowing progress in research within the field of Lexical Simplification in the spanish language. The complexity metric presented in Table 1 corresponds to the scores provided by the annotators during the annotation process, which reflect the perceived degree of complexity for each word in its specific context. Note: It should be noted that words identified and labeled by the annotators as complex are in bold.

Table 1 presents several examples of the words identified and annotated as complex in the corpus during the GovAI*Ec* tagging process.

3.2 Linguistic Features (LF)

Regarding the data set has a combination of 23 LF for each sentence. The linguistic features we used as indicators to describe the linguistic properties of texts (Zeng et al., 2024).

To compute the LFs, we consider the 15 features proposed in previews works (Ortiz-Zambrano and Montejo-Ráez, 2021), (Zambrano and Montejo-Raéz, 2021) and add 8 features computed from the category counts of *POS* for a total of 23 LFs (Ortiz Zambrano, Espin-Riofrio, and Montejo Ráez, 2023). We indicate these features and some of the research papers that have also considered them:

1. The absolute frequency (Paetzold, 2021).

[2] GovAI*Ec* - Available at https://figshare.com/authors/JennyOrtiz/20731154

Words tagged by the annotators in the texts of the corpus *GovAIEc*

ID	Sentence	Complexity
CNE-3432	La Secretaría General del Consejo Nacional Electoral - CNE [..] [..] **remitir** a la Dirección Nacional de Organizaciones Políticas, que será la encargada de emitir el informe correspondiente, [..]	0.33
CNT-4334	La CNT EP, no cobrará ningún valor por las reparaciones de los daños producidos entre la central y la caja de **dispersión** inclusive si el daño se localiza entre la caja de **dispersión** y el aparato [..]	1.00
ATM-0097	De no haberse efectuado la aprehensión del o los vehículos [..] [..] el agente fiscal podrá solicitar al Juez de Tránsito disponga las [..] cautelares **pertinentes** para la práctica de las mencionadas [..].	1.00
SRI-7274	De acuerdo a lo señalado en el Código **Tributario** Artículo 153 (Plazos para el pago), el porcentaje para el pago de la primera cuota siempre será del 20% de la obligación tributaria, por lo que este [..]	0.33
AG-6613	Una vez identificado el bien se procede a la entrega del Parte de Retiro Temporal, con el que debe **acudir** a la Ventanilla # 38 [..] donde le entregan el número de expediente para que acuda [..]	0.33

Table 1: Examples of words tagged by the annotators in the texts of the *GovAIEc* corpus.

2. The relative frequency of the target word.
3. The number of characters of the token (Paetzold, 2021).
4. The number of syllables (Shardlow, 2013), (Ronzano et al., 2016), (Shardlow, Cooper, and Zampieri, 2020a), (Paetzold and Specia, 2016).
5. The position of the target word in the sentence (Shardlow, 2013), (Ronzano et al., 2016).
6. Number of words in sentence (Shardlow, 2013), (Ronzano et al., 2016).
7. The Part Of Speech category (Ronzano et al., 2016).
8. The relative frequency of the word before the token (Paetzold, 2021).
9. The relative frequency of the word after the token (Paetzold, 2021).
10. The number of characters in the word before the token (Ronzano et al., 2016).
11. The number of characters in the word after the token (Ronzano et al., 2016).
12. Lexical diversity (Shiroyama, 2022).
13. The number of synonyms (Mosquera, 2021).
14. The number of hyponyms (Mosquera, 2021).
15. The number of hyperonyms (Mosquera, 2021).
16. The number of nouns, singular or massive.
17. The number of auxiliaries verbs.
18. The number of adverbs.
19. The number of symbols.
20. The number of numeric expressions.
21. The number of verbs.
22. The number of nouns.
23. The number of pronouns.

The last eight features (from 16 to 23) are traditional categories of the POS (Part Of Speech) applied in the investigations of (Ronzano et al., 2016), (Paetzold and Specia, 2016), (Desai et al., 2021).

3.3 Evaluation

In this study, to evaluate the performance of the models on the GovAI*Ec* corpus, we used models have been widely used to create state-of-the-art solutions for numerous tasks, they are known for their robustness and effectiveness in investigating lexical complexity in spanish texts, such as XLM-RoBERTa-base (Agerri and Agirre, 2023), XLM-RoBERTa-large (Ortiz Zambrano, Espin-Riofrio, and Montejo Ráez, 2023), RoBERTa-large-BNE (North et al., 2023a), and bert-base-spanish-wwm-uncased (Hossain et al., 2024). Runs were carried out with each of the models, applying 30, 50 and 70 epochs. Our strategy focuses on integrating the 23 linguistic features of the corpus together with the encodings generated by the pre-trained models. The goal is to evaluate whether this combination provides satisfactory answers to our research hypotheses. The integration of linguistic features involves concatenating them, after applying min-max scaling, with the embeddings resulting from the last encoding layer, and before reaching the classification header.

3.4 Combining LFs with embeddings

The linguistic vectors are then concatenated with the text embedding generated by the Transformer model. For BERT models, the embedding corresponding to the token [CLS] is used, while for RoBERTa models an aggregate embedding is used. Figure 1 graphically shows the flow of steps involved in the process. This visual representation details the execution of language models, such as BERT and RoBERTa. The graphic not only provides an overview of the execution process of these models, but also highlights the incorporation of linguistic features into the model. This approach seeks to optimize both accuracy and effectiveness in the specific task of predicting lexical complexity.

3.5 Training configuration

We tested with different hyperparameters and variants of Transformer models. Table 2 presents the hyperparameters that we apply in the executions of the models. The Google Colaboratory GPU was used for both training and evaluation of our system. We ran runs with different numbers of epochs (30, 50, 70) to observe the behavior of the model. This approach allowed us to identify that, with a higher number of epochs, the model achieved better performance, without falling into overfitting. We used an initial learning rate based on standard recommendations for large language models (LLMs) and dynamically adjusted this value during training. We selected the batch size based on hardware limitations and training stability, prioritizing a value that allowed efficient training without compromising model convergence. For other hyperparameters, such as the number of layers or attention heads, we use predefined configurations based on the selected model architecture (e.g. BERT, GPT, etc.), as these are typically optimized for general natural language processing tasks. The Transformer models were implemented using the versions available in the huggingface library[3].

Hyperparameter	Setup
Learning rate	$5 \cdot 10^{-5}$
per_device_train_batch_size	32
per_device_eval_batch_size	32
epochs	50
nn.Dropout	0.5
evaluation_strategy	50
logging_strategy	50
input size	768

Table 2: hyperparameter settings.

Table 3 provides a detailed description of the different models used in the study, including their key configurations such as the number of layers, hidden size, self-attention heads, and the total number of parameters. This table aims to improve the readability and understanding of each model's architecture, facilitating a clearer comparison of their specifications and helping to contextualize their performance in the evaluation.

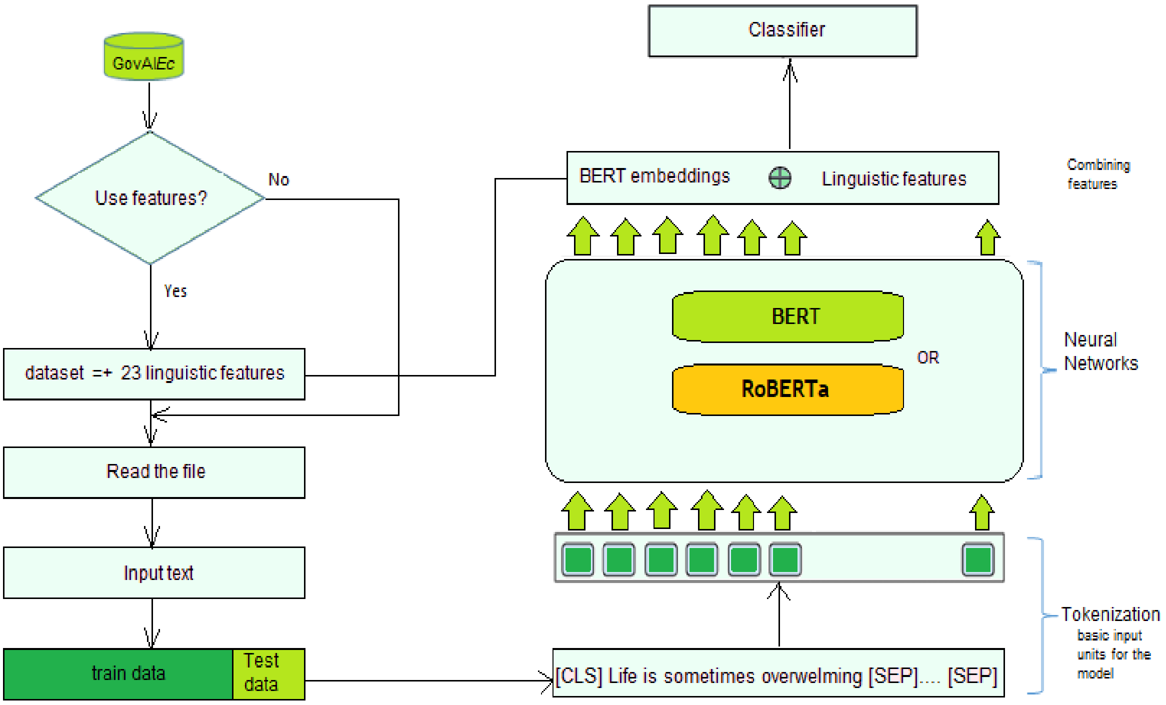
4 Experiments

In this section we describe the architecture used, the hyperparameters tuned, the optimization strategies applied, and the approach implemented that support the execution and training of the models. To improve the identification of complex words, we implement a strategy that combines linguistic features (such as word frequency, sentence length, and syntactic density) with embeddings generated by language models (LLM). This combination allows capturing both the explicit information provided by the linguistic features and the semantic context captured by the embeddings. The results show that this strategy significantly improves the model's ability to identify complex words, as it takes advantage of the complementarity of both sources of information.

The computed linguist features were normalized with a Min-Max transformation before being passed to the learning algorithm. The goal of this transformation is to ensure

[3] https://huggingface.co/docs/transformers/index

Figure 1: Linguistic feature integration.


GovAIEc
Use features?
No
Yes
dataset =+ 23 linguistic features
Read the file
Input text
train data
Test data
Classifier
BERT embeddings ⊕ Linguistic features
Combining features
BERT
OR
RoBERTa
Neural Networks
Tokenization
basic input units for the model
[CLS] Life is sometimes overwelming [SEP].... [SEP]

Model	Layers	Hidden Size	Self-Attn. Heads	Total Params
BERT-base (BETO) (Devlin et al., 2019)	12	768	12	110M
BERT-large (Devlin et al., 2019)	24	1024	16	335M
RoBERTa-large-BNE (Gutiérrez-Fandiño et al., 2021b)	-	-	-	-
XLM-RoBERTa-base (Conneau et al., 2020)	12	768	12	125M
XLM-RoBERTa-large (Conneau et al., 2020)	24	1024	16	335M

Table 3: Model Descriptions and Parameters.

that all features are normalized to the same scale in the range [0, 1]. For this, first, the standardized value X_{std} is computed for these linguistic vectors in the training set:

$$X_{std} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1}$$

We split the data into two sets: 80% was used for model training and the remaining 20% was used for evaluation. The batch size during model training was 32 samples.

4.1 Using encoders for lexical complexity prediction

Fine-tuning was performed following the process shown below (see also Figure 1):

1. The sequence input was augmented with the target term for which the complexity estimate is determined. This term is placed after a [SEP] token.

2. Once the sequence input passes through the encoder (BERT or XLM-RoBERTa), the sentence embedding was concatenated with a Max-Min normalized version of the linguistic features.

3. The resulting vector was fed into the classification layer. This layer is a normal dense layer for BERT. In the case of XLM-RoBERTa, the classification header is composed of a pair of dense layers prepended with a dropout layer and an activation layer *tanh* after the first dense layer.

Therefore, the entire network can be fine-tuned even if linguistic features are injected. The best performance of the model was fine-tuned with a batch size of 32 for 50 epochs, a learning rate of $5 \cdot 10^{-5}$ and the AdamW optimization algorithm. The results obtained are shown in Table 4.

4.2 Results

Table 4 presents the performance of various configurations of the deep language models with or without LF. Our goal in the evaluation stage is to comprehensively analyze the model's performance in terms of accuracy (MAE), sensitivity to large errors (MSE), and a combination of both (RMSE). Furthermore, these are standard metrics in regression tasks and have been widely used in the related literature (North, Zampieri, and Shardlow, 2023), which facilitates the comparison of our results with other studies.

In this research, it has been shown how the inclusion of linguistic features and the increase of epochs substantially improves the LCP, even for smaller models such as XLM-RoBERTa-base. The addition of LF helped to significantly reduce errors in different metrics, showing that these features provide complementary information to the model. Using linguistic features and encodings from pre-trained models seemed to consistently improve the results. The performance gain seems to be related to the model along with the number of epochs. Based on running different models and combinations of data, with a MAE = 0.1551 on the spanish GovAIEc corpus, the best performance was obtained using the fine-tuned version of the spanish BERT model (bert-base-spanish-wwm-uncased) along with linguistic features. With a MAE = 0.1575, fine-tuning BERT was again the best choice, but only if linguistic features are introduced into the model, although we can notice that the number of epochs is lower this time (epoch=20).

Table 4 presents the results of running deep learning models for prediction tasks, evaluated using the MAE. The table is organized by the number of training epochs for each model.

4.3 Comparison

Using LF together with BERT (bert-base-spanish-wwm-uncased) achieved the best results with a MAE = 0.1551 and RMSE = 0.229, showing that this model, in combination with linguistic features, is the most robust in terms of accuracy and stability. XLM-RoBERTa-base and RoBERTa-large-BNE show less consistent behavior, presenting larger fluctuations in MAE and RMSE, which could be related to a lower ability to take advantage of the introduced linguistic features or to overfitting derived from their larger size. For all models, an increase in the number of epochs is not always a guarantee of an improvement in the metrics. For the case of bert-base-spanish-wwm-uncased with the LF it reaches its best performance in terms of MAE and RMSE at 30-50 epochs, while 70 epochs leads to a deterioration in performance.

In Figure 2 we observed that models with a higher number of epochs (50 and 70) do not always achieve significant improvement, implying that the models could be reaching their optimization limit with fewer epochs.

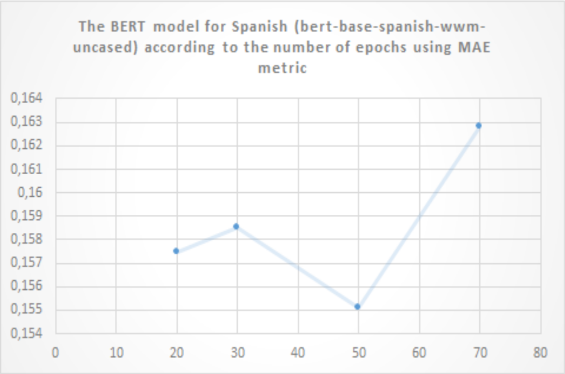


Figure 2: Model performance when integrating linguistic features and increasing the number of training epochs.

5 *Discussion*

This study is compared to previous work that applies a similar approach based on pre-trained models, such as BERT, but in a different domain. A detailed comparison with similar research conducted in other contexts is then presented. Although the method used

# Epochs	Model	MAE	MSE	RMSE
50	**bert-base-spanish-wwm-uncased + LF**	0.1551	0.0584	**0.242**
20	bert-base-spanish-wwm-uncased + LF	0.1575	0.0565	0.238
30	bert-base-spanish-wwm-uncased + LF	0.1585	0.0526	0.229
50	bert-base-spanish-wwm-uncased	0.1604	0.0602	0.245
30	bert-base-spanish-wwm-uncased	0.1627	0.0598	0.245
70	bert-base-spanish-wwm-uncased	0.1628	0.0631	0.251
30	XLM-RoBERTa-base + LF	0.1694	0.0585	0.242
20	RoBERTa-large-BNE + LF	0.1992	0.0582	0.241
50	XLM-RoBERTa-large + LF	0.2033	0.0543	0.233
70	XLM-RoBERTa-large + LF	0.2047	0.0695	0.264
70	RoBERTa-large-BNE + LF	0.2058	0.0518	0.228
50	XLM-RoBERTa-base + LF	0.2062	0.0531	0.230
70	XLM-RoBERTa-base + LF	0.2072	0.0532	0.231
30	RoBERTa-large-BNE + LF	0.2095	0.0548	0.234
30	XLM-RoBERTa-large + LF	0.2099	0.0545	0.233
20	XLM-RoBERTa-base + LF	0.2099	0.0549	0.234
20	XLM-RoBERTa-large + LF	0.2130	0.0553	0.235
50	XLM-RoBERTa-base	0.2136	0.0560	0.237
50	RoBERTa-large-BNE + LF	0.2137	0.0556	0.236
50	XLM-RoBERTa-large	0.2145	0.0577	0.240
30	XLM-RoBERTa-large	0.2159	0.0567	0.238
30	RoBERTa-large-BNE	0.2160	0.0566	0.238

Table 4: Final Results of Language Models for Lexical Complexity Assessment in the GovAI*Ec* Corpus.

in both cases is the same, the differences in the results can be explained by factors such as the inherent difficulty of the domain of Ecuadorian government texts or variations in the quality and characteristics of the data sets used for training. Despite these differences, the behaviors observed in both studies are comparable, suggesting that although the context and data vary, the model maintains its ability to address lexical complexity tasks in different domains with consistent results.

In the context of this study, the BERT model stands out as the best model compared to the other approaches evaluated, both in terms of prediction accuracy and adaptability to different domains. With an MAE = 0.1323 on the academic domain corpus, fine-tuned BERT showed superior performance, especially when trained specifically on a corpus with academic domain texts. This capability suggests that BERT has managed to effectively capture linguistic particularities and lexical complexity patterns in specialized texts, highlighting its generalization capacity for lexical complexity prediction tasks in an academic context (Ortiz-Zambrano, Espín-Riofrío, and Montejo-Ráez, 2024).

The results obtained highlight the influence of domain and linguistic characteristics on the performance of BERT-based models when tackling specific prediction tasks in spanish. In the academic domain corpus, the BERT model achieved a MAE = 0.1323, suggesting a high capacity of the model to capture the linguistic complexities of specialized texts. In contrast, when additional linguistic features were integrated into the BERT model trained with general spanish data, the MAE increased to 0.1361 (Ortiz-Zambrano, Espín-Riofrío, and Montejo-Ráez, 2024). On the other hand, when adjusting the BERT model with a dataset from public institutions, the performance decreased slightly, with a MAE = 0.1551, probably due to the greater stylistic and semantic variability in these texts.

When comparing the results obtained with the XLM-RoBERTa-large + LF model trained on a legal corpus, which achieved a MAE = 0.1338 after 50 epochs (Ortiz Zambrano, Espin-Riofrio, and Montejo Ráez, 2023), with the values previously reported for the BERT model tuned in other domains, notable differences emerge that highlight the impact of the corpus domain and the integration of linguistic features. For example, in the academic domain, BERT obtained a MAE = 0.1323, slightly higher than the performance of the model tuned on the legal corpus. However, the same BERT model tuned on data from public institutions achieved a MAE = 0.1361, which puts it at a disadvantage compared to XLM-RoBERTa-large + LF in the legal domain.

The results obtained indicate that, although the model shows positive performance in the lexical simplification of texts from Ecuadorian public institutions, it is essential to determine the appropriate number of training epochs to avoid overfitting. This adjustment must be done cautiously, since an excessive number of epochs can lead to the model overfitting the training data, losing its ability to generalize to new texts. Thus, it becomes essential to find a balance that allows the model to learn effectively without compromising its ability to adapt to unseen contexts.

The model's performance varies with the corpus domain and configurations, yet the integration of linguistic features proves highly effective for tasks requiring in-depth language analysis. This approach enhances accuracy and capability in identifying complex words, demonstrating its strategic value for natural language processing systems. Overall, BERT remains the best model in this study due to its ability to efficiently adapt to different domains and datasets, and its superior performance in terms of prediction accuracy, especially in the government domain (table 4). Although other models such as XLM-RoBERTa and RoBERTa show competitive performance, BERT has managed to stand out, highlighting its effectiveness as a reference model for lexical complexity prediction tasks in specific natural language contexts.

6 Conclusions

A series of experiments have been carried out to test the convenience of combining Transformers model encodings with linguistic features traditionally used in lexical complexity. Our approach takes advantage of the combination of advanced NLP techniques by applying the variants of Transformers-based deep learning models: BERT, RoBERTa. The dataset is composed of features of different nature: linguistic and encodings.

Several combinations of features were tested to measure the actual contribution of each potential set of features.

We can summarize the main contributions as follow:

1. A comprehensive set of experiments was performed to test the suitability of combining transformer encodings with lexical features traditionally used in complex word identification.
2. Some relevant findings were extracted: (a) fine-tuning with the word to be analyzed in the input sequence as a separate fragment leads to better results, (b) Overall, fine-tuning the monolingual versions of BERT yielded better results compared to the other models.

It is important to highlight that for a correct model performance the number of epochs directly influences how the model learns and generalizes, and even determining the optimal number of epochs is important, as the fine-tuning process leads to superior performance, mitigating the risk of overfitting as it becomes essential to obtain good results (Jebali et al., 2024). However, deep learning models operate as a black box, so understanding how linguistic features complement deep features requires working on the explainability of the deep model itself, as Transformers can encode information related to syntax, dependencies, grammar, gender, negation, semantics, among others, within its layers. The wealth of knowledge present in transformer-based models can help extract complementary clues to contextual complexity (Paetzold, 2021).

7 Feature work

It is clear that further studies are needed on integrating linguistic features into deep learning models. We plan to investigate which linguistic features bring additional information to the network by performing ablation tests, introducing these features incrementally into the newly implemented end-to-end approach. New features, selection strategies, and transformation methods could also be explored. We believe that optimizing the network parameters by taking these external features into account could significantly improve its performance.

References

Agerri, R. and E. Agirre. 2023. Lessons learned from the evaluation of spanish language models.

Alarcón, R., L. Moreno, and P. Martínez. 2020. Hulat-alexs cwi task-cwi for language and learning disabilities applied to university educational texts. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain*.

Aumiller, D. and M. Gertz. 2023. Unihd at tsar-2022 shared task: Is compute all we need for lexical simplification. *arXiv preprint arXiv:2301.01764*.

Bott, S., H. Saggion, N. P. Rojas, M. S. Salazar, and S. C. Ramirez. 2024. Multils-sp/ca: Lexical complexity prediction and lexical simplification resources for catalan and spanish.

Bylund, E., Z. Khafif, and R. Berghoff. 2024. Linguistic and geographic diversity in research on second language acquisition and multilingualism: An analysis of selected journals. *Applied Linguistics*, 45(2):308–329.

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*, 2020:2020.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2023. Spanish pre-trained bert model and evaluation data.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Desai, A. T., K. North, M. Zampieri, and C. Homan. 2021. LCP-RIT at SemEval-2021 task 1: Exploring linguistic features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 548–553, Online, August. Association for Computational Linguistics.

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*.

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

Gutiérrez-Fandiño, A., J. Armengol-Estapé, A. Gonzalez-Agirre, and M. Villegas. 2021a. Spanish legalese language model and corpora. *arXiv preprint arXiv:2110.12201*.

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. 2021b. Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.

Hossain, M. S., A. I. Paran, S. H. Shohan, J. Hossain, and M. M. Hoque. 2024. SemanticCUETSync at SemEval-2024 task 1: Finetuning sentence transformer to find semantic textual relatedness. In A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá, editors, *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1222–1228, Mexico City, Mexico, June. Association for Computational Linguistics.

Howard, J. and S. Ruder. 2018. Universal language model fine-tuning for text classification.

Jebali, M. S., A. Valanzano, M. Murugesan, G. Veneri, and G. D. Magistris. 2024. Leveraging the regularizing effect of mixing industrial and open source data to prevent overfitting of LLM fine tuning. In *International Joint Conference on Artificial Intelligence 2024 Workshop on AI Governance: Alignment, Morality, and Law*.

Licardo, M., N. Volčanjk, and D. Haramija. 2021. Differences in communication skills among elementary students with mild intellectual disabilities after using easy-to-read texts. *The new educational review*, 64:236–246.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Minaee, S., T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Mosquera, A. 2021. Alejandro mosquera at semeval-2021 task 1: Exploring sentence and word features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 554–559.

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.

North, K., A. Dmonte, T. Ranasinghe, M. Shardlow, and M. Zampieri. 2023a. ALEXSIS+: Improving substitute generation and selection for lexical simplification with information retrieval. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, and T. Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 404–413, Toronto, Canada, July. Association for Computational Linguistics.

North, K., T. Ranasinghe, M. Shardlow, and M. Zampieri. 2023b. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.

North, K., T. Ranasinghe, M. Shardlow, and M. Zampieri. 2024. Deep learning approaches to lexical simplification: A survey. *Journal of Intelligent Information Systems*, pages 1–24.

North, K., M. Zampieri, and M. Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

Ortiz-Zambrano, J., C. Espin-Riofrio, and A. Montejo-Ráez. 2023. Sinai participation in simpletext task 2 at clef 2023: Gpt-3 in lexical complexity prediction for general audience.

Ortiz-Zambrano, J. A., C. H. Espín-Riofrío, and A. Montejo-Ráez. 2024. Deep encodings vs. linguistic features in lexical complexity prediction. *Neural Computing and Applications*, pages 1–17.

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 Zambrano, J. A., C. Espin-Riofrio, and A. Montejo Ráez. 2023. Legalec: A new corpus for complex word identification research in law studies in ecuatorian spanish.

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 International Workshop on Semantic Evaluation (SemEval-2016)*, pages 969–974.

Paetzold, G. H. and L. Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

Rampal, N., K. Wang, M. Burigana, L. Hou, J. Al-Johani, A. Sackmann, H. S. Murayshid, W. A. AlSumari, A. M. AlAbdulkarim, N. E. Alhazmi, et al. 2024. Single and multi-hop question-answering datasets for reticular chemistry with gpt-4-turbo. *Journal of Chemical Theory and Computation*, 20(20):9128–9137.

Rets, I. and J. Rogaten. 2021. To simplify or not? facilitating english l2 users' comprehension and processing of open educational resources in english using text simplification. *Journal of Computer Assisted Learning*, 37(3):705–717.

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.

Roundy, P. T., J. M. Trussel, and S. A. Davenport. 2023. The text complexity of local government annual reports. *Local Government Studies*, 49(5):1135–1156.

Sarzynska-Wawer, J., A. Wawer, A. Pawlak, J. Szymanowska, I. Stefaniak, M. Jarkiewicz, and L. Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

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., F. Alva-Manchego, R. T. Batista-Navarro, S. Bott, S. Calderon-Ramirez, R. Cardon, T. François, A. Hayakawa, A. Horbach, and A. Huelsing. 2024. The bea 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589.

Shardlow, M., M. Cooper, and M. Zampieri. 2020a. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France, May. European Language Resources Association.

Shardlow, M., M. Cooper, and M. Zampieri. 2020b. 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*.

Shiroyama, T. 2022. Comparing lexical complexity using two different ve modes: a pilot study. *Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022*, page 358.

Singh, S. and A. Mahmood. 2021. The NLP cookbook: Modern recipes for transformer based deep learning architectures. *IEEE Access*, 9:68675–68702.

Tan, K., K. Luo, Y. Lan, Z. Yuan, and J. Shu. 2024. An llm-enhanced adversarial editing system for lexical simplification.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

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.

Yuan, Y.-P., Y. K. Dwivedi, G. W.-H. Tan, T.-H. Cham, K.-B. Ooi, E. C.-X. Aw, and W. Currie. 2023. Government digital transformation: Understanding the role of government social media. *Government Information Quarterly*, 40(1):101775.

Zambrano, J. A. O. and A. Montejo-Raéz. 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.

Zeng, J., X. Tong, X. Yu, W. Xiao, and Q. Huang. 2024. Interpretara: Enhancing hybrid automatic readability assessment with linguistic feature interpreter and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19497–19505.