Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 177-188 recibido 31-03-2022 revisado 03-06-2022 aceptado 06-06-2022

Transformers for Lexical Complexity Prediction in Spanish Language

Transformers para la Predicción de la Complejidad Léxica en Lengua Española

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

Abstract: In this article we have presented a contribution to the prediction of the complexity of simple words in the Spanish language whose foundation is based on the combination of a large number of features of different types. We obtained the results after run the fined models based on Transformers and executed on the pre-trained models BERT, XLM-RoBERTa, and RoBERTa-large-BNE in the different datasets in Spanish and executed on several regression algorithms. The evaluation of the results determined that a good performance was achieved with a Mean Absolute Error (MAE) = 0.1598 and Pearson = 0.9883 achieved with the training and evaluation of the Random Forest Regressor algorithm for the refined BERT model. As a possible alternative proposal to achieve a better prediction of lexical complexity, we are very interested in continuing to carry out experimentations with data sets for Spanish, testing state-of-the-art Transformer models.
Keywords: Lexical Complexity, Prediction, Encodings, Transformers.

Resumen: En este artículo hemos presentado una contribución a la predicción de la complejidad de palabras simples en lengua española cuyo fundamento se basa en la combinación de un gran número de características de distinta naturaleza. Obtuvimos los resultados después de ejecutar los modelos afinados basados en Transformers y ejecutados sobre los modelos pre-entrenados BERT, XLM-RoBERTa y RoBERTa-large-BNE en los diferentes conjuntos de datos en español y corridos con varios algoritmos de regresión. La evaluación de los resultados determinó que se logró un buen desempeño con un Error Absoluto Medio (MAE) = 0.1598 y Pearson = 0.9883 logrado con el entrenamiento y evaluación del algoritmo Random Forest Regressor para el modelo BERT afinado. Como posible propuesta alternativa para lograr una mejor predicción de la complejidad léxica, estamos muy interesados en seguir realizando experimentaciones con conjuntos de datos para español probando modelos de Transformer de última generación.
Palabras clave: Complejidad Léxica, Predicción, Incrustaciones de Palabra, Transformadores.

1 *Introduction*

A common assumption is that people who are familiar with the vocabulary of a text can often understand the meaning of the words, even if they have difficulty with grammatical structures (Uluslu, 2022). The task of detecting in the content of the documents the words that are difficult or complex to understand by the people of a given group is known as Complex Word Identification - CWI (Rico-Sulayes, 2020) and it is a task that constitutes a fundamental step in many applications related to natural language, such as Text Simplification. Automatic lexical simplification can then become an effective method of making the text accessible to different audiences (Uluslu, 2022).

Deep learning and its revolutionary tech-

ISSN 1135-5948. DOI 10.26342/2022-69-15

nology constitute the new state of the art in various Natural Language Processing (NLP) tasks (Singh and Mahmood, 2021), in which lexical complexity prediction (LCP) is no exception (Nandy et al., 2021). It is important to point out that, after the comparison and analysis of other approaches versus deep learning approaches, a viable path of possible solutions is generated for low-resource languages where deep models are not always available or work as well as those of deep learning in English language. Likewise, it should be taken into account that the computational requirements for the application of deep learning models turn out to be significantly higher compared to those used in traditional approaches (Bender et al., 2021).

The field of NLP has shown incredible progress in the last two years, this is particularly due to the Transformer architecture (Vaswani et al., 2017) that takes advantage of large amounts of unlabeled text corpus (Canete et al., 2020). Deep learning models show significant improvement over shallow machine learning models with the rise of transfer learning and pretrained language models. The deep learning pretrained language models, BERT and XLM-RoBERTa, are considered state-of-the-art in many NLP tasks (Yaseen et al., 2021).

We present our approach aimed at predicting the complexity score for single words in the Spanish language, since resources are scarce and are not as numerous as those available for the English language.

Our model leverages the combination of advanced NLP techniques of deep learning models based on Transformers: BERT (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2019), RoBERTa-large-BNE (Gutiérrez-Fandiño et al., 2021) and pre-trained word embeddings together with a set of textual complexity features made by hand (Hand-Crafted Features). For this, we use the corpus in Spanish CLexIS[2] corpus proposed by (Zambrano and Montejo-Ráez, 2021). Our challenge is achieving to improve the lexical complexity prediction implementing a fine-tuned model on a previously trained model, for which, we follow the research done by (Rojas and Alva-Manchego, 2021).

The models used achieve a good performance shown in the results with a MAE = 0.1592 and a Person correlation 0.9883 for the identification of simple complex words.

2 Related Work

In past decades, the application of very simple metrics such as calculating the number of syllables in words (Mc Laughlin, 1969) or verifying whether the word was part of a specific list to classify it as easy or complex (Dale and Chall, 1948) were the techniques that were applied in text legibility tasks.

After, the systems based on the characterization of words (using contextual, lexical and semantic characteristics) and the application of a Random Forest classifier (Breiman, 2001) to determine whether a word is complex or not are presented. A total of 45 handwritten features were computed in these systems, and each word was modeled as a feature vector. Surface functions (three functions), dependency tree functions (eight functions), Corpus-based functions (fifteen functions), WordNet functions (eleven functions), and WordNet and corpus frequency functions (four functions) were applied. The best result obtained was a Precision value of 0.186, a Recall of 0.673, a G-score of 0.750 99 and an F-score of 0.292.

The investigations in the last years are directed to the Identification of Complex Words - CWI. The objective of these applications is to be able to predict the complexity of words based on the construction of their features, as exposed in the work carried out by (Shardlow, Cooper, and Zampieri, 2020) presenting their approach on a set of features of word embeddings from Glove, InferSent, and various linguistic features obtained as predictive sources of lexical complexity, such as word frequency, word length, or number of syllables. Then, they trained a linear regression model using different subsets of functions, obtaining as a result an MAE = of 0.0853.

(Shardlow et al., 2021) developed a system for predicting word complexity for the shared LCP task hosted on SemEval 2021 where task organizers distributed to participants the CompLex corpus (Shardlow, Cooper, and Zampieri, 2020) but in its augmented version. The task was located on the Lexical Semantics track, which consisted of predicting the complexity value of words in context.

(Ortiz-Zambrano and Montejo-Ráez, 2021) Carried out a machine learning approach that was based on 15 linguistic features obtained at the word level and their environment. Trained a supervised random forest regression algorithm on the set of features. Several runs were made with different values to observe the performance of the algorithm. The best results achieved were a MAE = 0.07347, MSE = 0.00938 and RMSE = 0.096871.

In our approach, we review the use of word embeddings from the pre-trained and fine-tuned models, and compare them to a broader list of linguistic features at the lexical level. Our objective is to provide an exhaustive evaluation that shows more clearly, the executions carried out on several different data sets in the Spanish language, how the lexical features together with the deep encodings contribute to the prediction of lexical complexity.

3 Materials and Method

This section briefly details about the pre-trained models and their application for the generation of encodings at both the sentence and word levels. Likewise, the data sets that have been used in the different experiments are presented. Finally, the different classification algorithms and the applied features are shown.

3.1 Dataset

The CLexIS^2 corpus was elaborated with the transcripts of the recorded classes of the professors of the degrees of Computer Systems Engineering and Software Engineering, two degrees that belong to the Faculty of Mathematical Sciences of the University of Guayaquil (Ecuador) (Zambrano and Montejo-Ráez, 2021).

CLexIS^2 has become a resource of great interest and importance, due to the fact that there are few resources in Spanish available for NLP researchers[1], and some of them do not usually contain annotations that facilitate the development of NLP models (Davidson et al., 2020). For its construction, the collection presented in the ALexS 2020 competition at IberLEF 2020 (Ortiz-Zambranoa and Montejo-Ráezb, 2020) was taken as a reference.

Annotated words as complex have an associated level of complexity, computed based on the number of annotators that agreed to consider it as a complex word. Therefore, the task we are facing here can be faced as a regression problem, so error metrics will be used to evaluate different systems.

Table 1 shows some statistics on different type of words present in the CLexIS^2 dataset.

3.2 Transformer based language models

The models were taken from the *Transformers*[2] library.

- The pre-trained BERT model that we chose was the one that the Spanish community uses mostly in research work to date, which is bert-base-uncased (BETO) (Canete et al., 2020).

  BERT-base model has the number of layers L=12, the hidden size H=768, the number of self-attention heads A=12,and Total Parameters=110M.

  BERT-large model has the number of layers=24, the hidden size=1024, the number of self-attention heads=16, and Total Parameters=335M (Conneau et al., 2019).

- The -RoBERTa model applied was xlm-roberta-base (Conneau et al., 2019).

  The XLM-RoBERTa-base model has the number of layers L=12, the hidden size H=768, the number of self-attention heads A=12,and Total Parameters=270M.

  XLM-RoBERTa-large model has the number of layers L=24, the hidden size H=1024, the number of self-attention heads A=16, and Total Parameters=550M (Conneau et al., 2019).

- The RoBERTa-large-BNE model used was PlanTL-GOB-ES/roberta-large-bne being the largest Spanish-specific model to date (Gutiérrez-Fandiño et al., 2021).

  XLM-RoBERTa-large is a transformer-based masked language model for the Spanish language. It is based on the RoBERTa large model[3].

  The RoBERTa-large-BNE model has the number of layers L=24, the hidden size H=1024, the number of self-attention heads A=16,and Total Parameters=335M(Conneau et al., 2019).

[1] CLexIS^2 - https://osf.io/kfpc9/?view$_o nly$ $= 18ae61a2049a48cb91c6773d53fb8ac2$

[2] https://huggingface.co/

[3] https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne

Number of	Count
Sum. of content words (verbs, adjectives and nouns)	153,885
Different content words	200,785
Rare words (low frequency in CREA corpus (Saggion et al., 2015))	143,464
Sentences	9,756
Complex sentences	4,101
Total words	300,420

Table 1: Volumetrics for CLexIS[2].

3.3 Experiments design

Our purpose is to demonstrate how the combination of different types of features contribute to a better performance in predicting lexical complexity. We base our proposal on several of the works presented at the International Workshop on Semantic Evaluation - SemEval-2021 (Shardlow et al., 2021) where a total of 198 teams were presented, of which 54 teams officially sent their executions[4]; but the work that most attracted us due to its methodology was the experimentation carried out by (Zaharia, Cercel, and Dascalu, 2021) about *Combining Deep Learning and Hand-Crafted Features for Lexical Complexity Prediction.*

The figure 1 presents the workflow of the process executed to obtain of the Lexical Complexity Prediction. First, we chose the data sets for training were: the first data set was made up of the linguistic features made by hand - Hand-Crafted Features (HCF) and the second data set was made up of the Transformers encodings from the models: BERT in Spanish, multilingual XLM-RoBERTa and RoBERTa-grande-BNE. Next, we applied a fitted model on top of the previously trained model to demonstrate how running the fitted model on the previously trained model contributed to more accurate LCP results as see figure 1.

Finally, the different supervised learning algorithms were executed on the training data set to evaluate which of them achieved the best prediction score. Triple cross-validation was performed to ensure that the partitions contained independent data for training and testing. We have used some metrics that were applied to the results of the experiments presented in Sem-Eval 2021 (Shardlow et al., 2021), which are appropriate for evaluating continuous and classified data, such as: MAE, MSE, RMSE and Pearson's correlation.

3.3.1 Features

- **Hand-Crafted Features - HCF**

To obtain the morphological aspects of the text, we perform several experiments applying a total of 23 linguistic features and combine them with the word and sentence embeddings of previously trained deep learning models.

We have considered the 15 HCF proposed by (Ortiz-Zambrano and Montejo-Ráez, 2021) and added a sets of features computed from POS categories counts (Vettigli and Sorgente, 2021), (Liebeskind, Elkayam, and Liebeskind, 2021), giving a total of 23 Hand-Crafted Features. We used the Spacy library together with the model *es_core_news_sm* to extract these features. All these features were normalized with a z-score transformation before passing them to the learning algorithm.

1. *Absolute frequency*: the absolute frequency.
   The frequency of words is a measure that serves as an indicator of lexical complexity. If in common parlance a word occurs frequently, it is more likely to be recognized (Rayner and Duffy, 1986) and (Shardlow, Cooper, and Zampieri, 2020).

2. *Relative frequency*: the relative frequency of the target word.

3. *Word length*: the number of characters of the token. The length of the word was calculated in number of its characters. It is often the case that longer length words are more difficult to process and can therefore be considered *complex.* (Shardlow, 2013) (Shardlow, Cooper, and Zampieri, 2020) (Paetzold and Specia, 2016).

[4] https:// semeval.github.io/SemEval2021/tasks

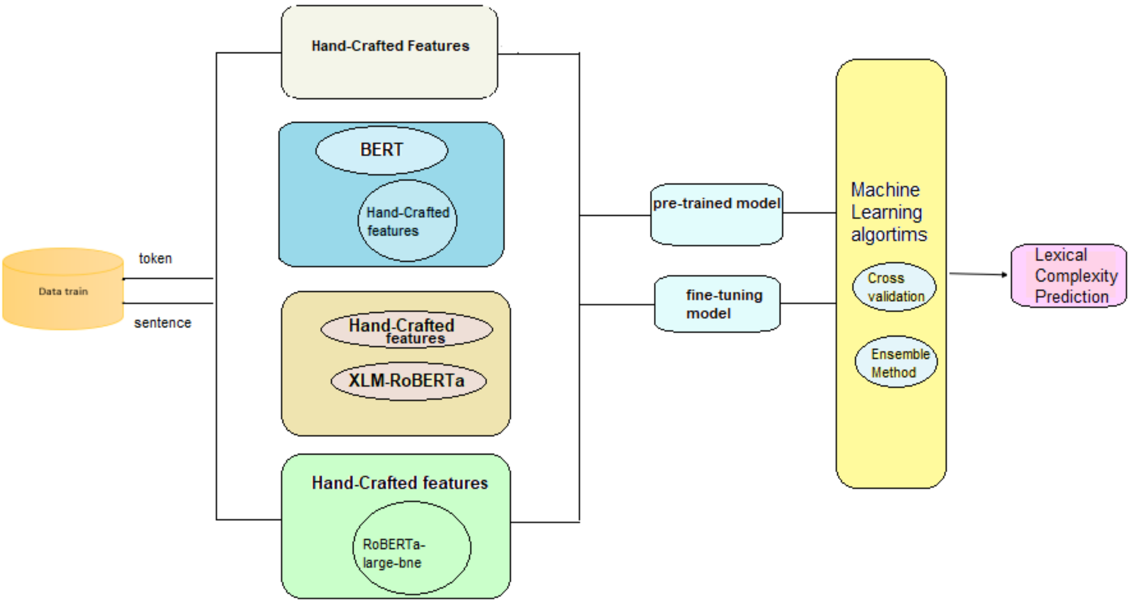

Figure 1: Representation of the workflow to obtain the Lexical Complexity Prediction.

4. *Number of syllables*: the number of syllables. A good estimate of complexity is the number of syllables contained in a word (Shardlow, 2013) (Ronzano et al., 2016) (Shardlow, Cooper, and Zampieri, 2020) (Paetzold and Specia, 2016).

5. *Target word position* (token-position): the position of the target word in the sentence. Position of the word (WordPosition) (Shardlow, 2013) (Ronzano et al., 2016).

6. *Number of words in the sentence* : number of words in the sentence. Words in sentence (NumSentenceWords) (Shardlow, 2013) (Ronzano et al., 2016).

   Based on the work proposed by (Ronzano et al., 2016) in the exploring linguistic features for lexical complexity prediction.

7. *Part Of Speech (POS)*: the Part Of Speech category.

8. *Relative frequency of the previous token*: the relative frequency of the word before the token.

9. *Relative frequency of the word after the token*: the relative frequency of the word after the token.

10. *Length of previous word*: the number of characters in the word before the token.

11. *Length of the after word*: the number of characters in the word after the token.

12. *Lexical diversity - MTDL*: the lexical diversity of the target word in the sentence.

    Additionally, the following WordNet features were also considered for each target word, as in the works carried out by (Gooding and Kochmar, 2018):

13. *Number of synonyms.*

14. *Number of hyponyms.*

15. *Number of hyperonyms.*

We follow the recommendations of (Paetzold and Specia, 2016), (Ronzano et al., 2016), (Gooding and Kochmar, 2018), (Liebeskind, Elkayam, and Liebeskind, 2021), (Desai et al., 2021) with the aim of improving results, generating 8 new features originating from the POS, which were:

1. PROPN - Number of pronouns within the sentence.

2. AUX - Number of auxiliaries within the sentence.

3. VERB - Number of verbs within the sentence.

4. ADP - Number of adverbs within the sentence.

5. NOUN - Number of nouns within the sentence.

6. NN - Number of Nouns, singular or massive.

7. SYM - Number of symbols within the sentence.

8. NUM - Number of numbers within the sentence.

- **BERT vector:** The bert-base-uncased model from the Hugging Face transformer library (Wolf et al., 2020) was applied. We took all the 768-dimensional numerical representation produced by the pre-trained and fine-tuned BERT model (Devlin et al., 2018) and added the twenty-three Hand-Crafted Features obtaining a dataset with a total of 1559 linguistic features of different nature.

- **XLM-RoBERTa vector:** As in the case of the BERT model, we take all the 768-dimensional numerical representation produced by the pre-trained RoBERTa model (Conneau et al., 2019) in the different combinations of sentence and target word encodings, for both the pre-trained model and the model fine-tuned, reaching a total of 1559 linguistic characteristics of different nature.

- **RoBERTa-large-BNE vector:** Regarding this model, we take all the 1024-dimensional numerical representation produced by the pre-trained RoBERTa-large-model model (Gutiérrez-Fandiño et al., 2021), in the same way that they were applied in the previous models, the data sets were made up of for the different combinations of sentence and target word encodings, for both the pre-trained model and the fine-tuned model, reaching a total of 2071 linguistic characteristics of different nature.

3.3.2 Machine Learning Algorithms

Similar to the work done by (Zaharia, Cercel, and Dascalu, 2021) in the case of the algorithms, the training and evaluation of the different c ombinations o f t he s ets w as carried out with a total of eight supervised algorithms for the regression, these are:

1. AdaBoost - AB (Paetzold, 2021).

2. Desicion Tree - DT (Shardlow, Evans, and Zampieri, 2021).

3. Gradient Boosting - GB (Vettigli and Sorgente, 2021).

4. Stochastic Gradient - SG (Bottou, 2010).

5. Nearest Neighbors - KNN (Liebeskind, Elkayam, and Liebeskind, 2021).

6. Support Vector Machines - SVM (Liebeskind, Elkayam, and Liebeskind, 2021).

7. Passive Aggressive - PA (Crammer et al., 2006).

8. Random Forest - RF) (Zaharia, Cercel, and Dascalu, 2021), (Desai et al., 2021).

Several experiments were carried out for each of the datasets where different configurations were explored for each of the algorithms. We apply the default values for the algorithms except for the case for tree-based algorithms, achieving to determine the best hyper-parameters with the following number of nodes:

- AdaBoost with 100 nodes.
- Random forest with 241 nodes.
- Gradient Boosting algorithm with 350 nodes.

4 ***Results***

4.1 Features Sets

We build several datasets composed of the combination of the features described above to run them on the pre-trained models. The table 2 table presents the description of the abbreviations that will be used for a better understanding of the features applied to the data sets. The detail below:

- The *Hand-Crafted Features* with the features coming from the 768-dimensional vector of the initial [CLS] token as sentence embeddings ($BERT_{sent}$).
- *The Hand-Crafted Features* with the 768-dimensional vector corresponding to the target token as word embeddings ($BERT_{word}$).
- The *Hand-Crafted Features* with encodings of the [CLS] token and encodings of the target token.
- The encodings of the [CLS] token.
- The encodings of the target token.
- The encodings of the [CLS] token with the encodings of the target token.

Features identifier	Description
HCF	Hand-Crafted (linguistic) Features.
BERT$_{sent}$	Sentence encodings from BERT models.
BERT$_{word}$	Token encodings from BERT models.
XLMR$_{sent}$	Sentence encodings from XLM-RoBERTa model.
XLMR$_{word}$	Token encodings from RoBERTa model.
RBNE$_{sent}$	Sentence encodings from RoBERTa-large-BNE model.
RBNE$_{word}$	Token encodings from RoBERTa-large-BNE model.

Table 2: Description of the Feature sets.

For the evaluation of the trained and fine-tuned models, those that were widely applied to LCP for the shared LCP task hosted in SemEval 2021: Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE) and Pearson correlation (Shardlow et al., 2021).

4.2 BERT model *pre-trained*

The table 3 shows the eight best performances corresponding to different combinations of features described in section 4.1 executed with BERT pre-trained.

As we can see in the three best results in predicting lexical complexity were achieved by the ABR and SVR algorithms. The best performance was achieved by the ABR - AdaBoost algorithm presenting the best prediction for the Spanish language with a MAE = 0.1632 and a Pearson = 0.999 in the execution with the data set made up of the combination of the features generated at the sentence level and at the word level - BERT$_{sent}\oplus$ BERT$_{word}$.

4.3 BERT model *fine-tuned*

We have applied the fine-tuned BERT model on top of the pre-trained BERT model for the purpose of the results. The table 4 shows the eight best executions, positioning RFR - Random Forest Regressor algorithm and the GBR - Gradient Boosting Regressor algorithm in the first places.

The best performance was obtained with the dataset composed of the combination of the features with target word encodings together with sentence encodings from BERT fine-tuned. The same combination of features achieved the best performance in the pre-trained model, but with lower results.

It should be noted that the RFR algorithm does not appear within the top eight places in the execution of the pre-trained model, but it achieves its best result when the model is refined, placing first and third within the three best executions tuned. RFR presented the best prediction for the Spanish language with a MAE = 0.1592 and a Pearson = 0.988 combining BERT$_{sent}\oplus$ BERT$_{word}$.

4.4 XLM-RoBERTa model *pre-trained*

Similar to the BERT model, the top eight sites were taken from all the runs that were done on the different data sets. The results of the best place for the pre-trained XLM-RoBERTa model were achieved by the ABR - AdaBoots algorithm with a MAE = 0.1623 and a Pearson = 0.9973 result of the combination of the features with target word and sentence encodings together with the HCF - XLMR$_{sent}\oplus$ XLMR$_{word}\oplus$ HCF, as can be seen in the table 5. It can be clearly shown that the pre-trained XLM-RoBERTa model has a better performance compared to the pre-trained BERT model, achieving a better prediction of Lexical Complexity.

4.5 XLM-RoBERTa model *fine-tuned*

We also highlight that in the execution of the XLM-RoBERTa tuned model, it achieved a significant improvement compared to the results of the pre-trained model, reaching a MAE = 0.1601 and a Pearson = 0.998 as result of the combination of the features with target word encodings together HCF - XLMR$_{word}\oplus$ HCF. See table 6.

Comparing the results of the BERT and XLM-RoBERTa both tuned models, BERT tuned is so far the one that has an important performance achieved by a MAE = 0.1592 with the execution of the RFR algorithm combining BERT$_{sent}\oplus$ BERT$_{word}$.

BERT ***model pre-trained*** **with CLexIS^2**					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
BERT$_{sent}$⊕ **BERT**$_{word}$	**ABR**	**0,1632**	**0,0502**	**0,2343**	**0,9999**
BERT$_{word}$	SVR	0,1634	0,0432	0,2074	0,9023
BERT$_{word}$	ABR	0,1643	0,0512	0,2332	0,9947
BERT$_{word}$	GBR	0,1653	0,0494	0,0447	0,6977
BERT$_{word}$⊕ **HCF**	GBR	0,1655	0,0454	0,2088	0,7040
BERT$_{sent}$⊕ **BERT**$_{word}$⊕ **HCF**	GBR	0,1659	0,0418	0,2074	0,7147
BERT$_{sent}$⊕ **BERT**$_{word}$	GBR	0,1694	0,0444	0,2039	0,7167
BERT$_{sent}$⊕ **BERT**$_{word}$⊕ **HCF**	ABR	0,1699	0,0554	0,2334	0,9939

Table 3: Results of the model BERT pre-trained with features of different nature.

BERT ***model fine-tuned*** **with CLexIS^2**					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
BERT$_{sent}$⊕ **BERT**$_{word}$	**RFR**	**0,1592**	**0,0379**	**0,1982**	**0,9883**
BERT$_{sent}$⊕ **BERT**$_{word}$⊕ **HCF**	GBR	0,1600	0,0367	0,1979	0,8202
BERT$_{sent}$⊕ **BERT**$_{word}$⊕ **HCF**	RFR	0,1610	0,0401	0,1988	0,9987
BERT$_{sent}$⊕ **BERT**$_{word}$⊕ **HCF**	ABR	0,1610	0,0506	0,2242	0,9998
BERT$_{sent}$⊕ **BERT**$_{word}$⊕ **HCF**	ABR	0,1621	0,0487	0,2300	0,9999
BERT$_{word}$⊕ **HCF**	GBR	0,1622	0,0430	0,1984	0,8983
BERT$_{word}$	SVR	0,1622	0,0429	0,2018	0,9183
BERT$_{word}$	GBR	0,1632	0,0472	0,0429	0,7083

Table 4: Results of the model BERT tuned with features of different nature.

4.6 RoBERTa-large-BNE model *pre-trained*

The novelty of this research is to have incorporated the executions with the pre-trained model RoBERTa-large-BNE and its adjusted model. The eight best results are displayed in the table 7. The best position were achieved by the ABR-AdaBoost algorithm with a MAE = 0.1609 and a Person = 0.6754 combining the sentence and word encodings together with the HCF - RBNE$_{sent}$⊕ RBNE$_{word}$⊕ HCF.

It should be noted that the pre-trained model RoBERTa-large-BNE is the one that achieves a better prediction for lexical complexity in the Spanish language compared to the pre-trained models BERT and XLM-RoBERTa. See table 9.

4.7 RoBERTa-large-BNE model *fine-tuned*

Executing the RoBERTa-large-BNE tuned model, the results are encouraging, there is an improvement compared to the results of the pre-trained model. The table 8 displays the first places reached by the GBR-Gradient Boosting Regressor and SVR-Super Vector Regressor algorithms. It presents a low improvement, achieving in its performance a MAE = 0.1609 and a Pearson = 0.6754 combining the sentence and word encodings together with the HCF - RLBNE$_{sent}$⊕ RLBNE$_{word}$⊕ HCF, and the second and third places prove it in comparison with the pre-trained model.

It should be noted that the tuned model BERT is the one that achieves a better prediction for lexical complexity in the Spanish language compared to the tuned models XLM-RoBERTa and RBNE. See table 10.

It can be seen that the fined models based on Transformers make an important contribution to the Prediction of Lexical Complexity in the Spanish language. The table 11 presents the best five best results of all the experiments carried out with the models, both pre-trained and fined. It is important to mention that the Hand-Crafted Features, being such simple features because they are only based on the frequency of the words and several manual calculations, have been shown to contribute to improving the level of prediction of the complexity of the words.

5 *Discussion*

We have applied the BERT, RoBERTa, and RoBERTa-large-BNE models for our research in predicting lexical complexity in Spanish. We have closely followed the methodology applied in several of the works presented in the LCP task of the SemEval 2021 International Conference (Shardlow et al., 2021) which has allowed us to achieve very important results that demonstrate a relevant contribution in

XLM-RoBERTA ***model pre-trained*** **with CLexIS**^{2}					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HCF**	ABR	0,1623	0,0527	0,2270	0,9973
XLMR$_{word}\oplus$ **HCF**	ABR	0,1623	0,0513	0,2273	0,9973
XLMR$_{sent}\oplus$ **HCF**	ABR	0,1630	0,0524	0,2293	0,9973
XLMR$_{word}\oplus$ **HCF**	GBR	0,1653	0,0433	0,2073	0,4848
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HCF**	GBR	0,1658	0,0434	0,2074	0,4874
XLMR$_{sent}\oplus$ **HCF**	GBR	0,1663	0,0433	0,2082	0,4807
XLMR$_{word}\oplus$ **HCF**	SVR	0,1680	0,0483	0,2194	0,3095
XLMR$_{word}\oplus$ **HCF**	RFR	0,1690	0,0445	0,2093	0,9803

Table 5: Results of the model XLMR pre-trained with features of different nature.

XLM-RoBERTA ***model fine-tuned*** **with CLexIS**^{2}					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
XLMR$_{word}\oplus$ **HCF**	ABR	0,1601	0,0501	0,2251	0,9987
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HCF**	ABR	0,1620	0,0526	0,2268	0,9987
XLMR$_{sent}\oplus$ **HCF**	ABR	0,1620	0,0519	0,2287	0,9979
XLMR$_{sent}\oplus$ **HCF**	GBR	0,1630	0,0420	0,2062	0,4790
XLMR$_{word}\oplus$ **HCF**	GBR	0,1638	0,0429	0,2034	0,4800
XLMR$_{sent}\oplus$ **XMLR**$_{word}\oplus$ **HCF**	GBR	0,1652	0,0430	0,2069	0,4930
XLMR$_{word}\oplus$ **HCF**	SVR	0,1660	0,0482	0,2172	0,3083
XLMR$_{word}\oplus$ **HCF**	RFR	0,1669	0,0427	0,2013	0,9849

Table 6: Results of the model XLMR tuned with features of different nature.

RoBERTa-large-BNE ***model pre-trained*** **with CLexIS**^{2}					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	ABR	0,1609	0,0421	0,2047	0,6754
RBNE$_{sent}\oplus$ **RBNE**$_{word}$	ABR	0,1675	0,0556	0,2347	0,9952
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	GBR	0,1691	0,0434	0,2073	0,6607
RBNE$_{word}$	ABR	0,1693	0,0563	0,2360	0,9948
RBNE$_{word}$	GBR	0,1696	0,0447	0,2101	0,6400
RBNE$_{sent}\oplus$ **RBNE**$_{word}$	GBR	0,1698	0,0447	0,2102	0,6450
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	SVR	0,1708	0,0507	0,2224	0,2363
RBNE$_{sent}\oplus$ **HCF**	SVR	0,1708	0,0507	0,2224	0,0857

Table 7: Results of the model RBNE pre-trained with features of different nature.

RoBERTa-large-BNE ***model fine-tuned*** **with CLexIS**^{2}					
Features	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	GBR	0,1609	0.0421	0.2047	0.6754
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	SVR	0,1630	0,0435	0,2070	0,4883
RBNE$_{word}\oplus$ **HCF**	SVR	0,1666	0,0466	0,2136	0,4220
RBNE$_{word}$	ABR	0,1677	0.0551	0,2336	0,9952
RBNE$_{sent}\oplus$ **RBNE**$_{word}$	SVR	0,1684	0.0472	0.2152	0.4425
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	GBR	0,1686	0,0432	0,2067	0,6854
RBNE$_{word}$	SVR	0,1686	0,0468	0,2146	0,5021
RBNE$_{sent}\oplus$ **RBNE**$_{word}\oplus$ **HCF**	ABR	0,1689	0,0558	0,2351	0,9951

Table 8: Results of the model RBNE tuned with features of different nature.

The ***Spanish Language Models*** pre-trained **Best Result**			
Model	**Features**	**Alg**	**MAE**
RBNE	RBNE$_{sent}\oplus$ RLBNE$_{word}\oplus$ HCF	ABR	0.1609
XLMR	XLMR$_{sent}\oplus$ XLMR$_{word}\oplus$ HCF	ABR	0.1623
BERT	BERT$_{sent}\oplus$ BERT$_{word}$	ABR	0.1632

Table 9: Best results models pre-trained.

The *Spanish Language Models* fine-tuned			
Best Result			
Model	**Features**	**Alg**	**MAE**
BERT	$\text{BERT}_{sent} \oplus \text{BERT}_{word}$	RFR	0,1592
XLMR	$\text{XLMR}_{word} \oplus$ HCF	ABR	0,1601
RBNE	$\text{RBNE}_{sent} \oplus \text{RBNE}_{word} \oplus$ HCF	GBR	0,1609

Table 10: Best results models fine-tuned.

Summary of best results on the CLexIS^2 corpus						
Model	**Features**	**Alg**	**MAE**	**MSE**	**RMSE**	**Pearson**
$\textbf{BERT}_{fine-tuned}$	$\textbf{BERT}_{sent} \oplus \textbf{BERT}_{word}$	RFR	0,1592	0,0379	0,1982	0,9883
$\textbf{BERT}_{fine-tuned}$	$\textbf{BERT}_{sent} \oplus \textbf{BERT}_{word} \oplus$ **HCF**	GBR	0,1600	0,0367	0,1979	0,8202
$\textbf{XLMR}_{fine-tuned}$	$\textbf{XLMR}_{word} \oplus$ **HCF**	ABR	0,1601	0,0501	0,2251	0,9987
$\textbf{RBNE}_{fine-tuned}$	$\textbf{RBNE}_{sent} \oplus \textbf{RBNE}_{word}$	GBR	0,1609	0,0421	0,2047	0,6754
$\textbf{BERT}_{fine-tuned}$	$\textbf{BERT}_{sent} \oplus \textbf{BERT}_{word} \oplus$ **HCF**	RFR	0,1610	0,0401	0,1988	0,9987

Table 11: Summary of best results on the CLexIS^2 dataset.

the area of Lexical Simplification for Spanish.

We observe that according to the results of the final evaluation, especially in terms of fine-tuning, the Spanish language fined models made an important contribution to the prediction of lexical complexity by outperforming the proposal presented after the execution of the manual features-HCF. In the case of the RoBERTa-large-BNE model, we have found a performance that exceeds the rest of the models after the execution of the pre-trained model and even remains within the three best executions in the results of the tuned models, such as the proposals presented by (Gutiérrez-Fandiño et al., 2021)

6 Conclusions and Further Work

In this article, we have presented a contribution to predict the complexity of simple words in the Spanish language, combining a large number of features of different types. We consider that, after the multiple experimentations that we carried out, it allowed us to know the maximum performance for the different combinations of the data sets by applying the regression algorithms.

In our experiments, we obtained the results after the execution of several previously trained transformer-based models on several datasets in Spanish, combining features of different nature. The application of the fine-tuned models to generate features (embeddings) achieved a better performance of explored machine learning algorithms, which led to a MAE = 0.1598 and a Pearson of 0.9883 achieved with the evaluation and training of the Random Forest Regressor algorithm for the tuned model BERT.

Additional features can boost pre-trained models to levels of performance close to those of fine-tuned models alone, so it could be a feasible approach when there are not enough computational resources for such a downstream training.

As a possible alternative proposal to achieve a better prediction of lexical complexity, we are very interested in continuing to carry out experimentations on data sets for Spanish, testing state-of-the-art Transformer models. To this end, extrinsic evaluation will be overcome, comparing the best systems on this specific task with the possibilities of integrating external features like the ones proposed in this work.

Acknowledgements

We appreciate Jostin Daniel Escobar Suarez, Joel Stalin Sorroza Contreras, Diego Gabriel Bernal Yucailla y Diana Geovanna Aroca Pincay graduate of the Computer Systems Engineering degree from the University of Guayaquil, for their valuable contribution to the development of our work.

This work is partially funded by grants P20_00956 (PAIDI 2020) and 1380939 (FEDER Andalucía 2014-2020) from the Andalusian Regional Government.

References

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer, pages 177–186.

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

Canete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Crammer, K., O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Dale, E. and J. S. Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Davidson, S., A. Yamada, P. F. Mira, A. Carando, C. H. S. Gutierrez, and K. Sagae. 2020. Developing nlp tools with a new corpus of learner spanish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7238–7243.

Desai, A., K. North, M. Zampieri, and C. M. Homan. 2021. Lcp-rit at semeval-2021 task 1: Exploring linguistic features for lexical complexity prediction. *arXiv preprint arXiv:2105.08780*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Gooding, S. and E. Kochmar. 2018. Camb at cwi shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, and M. Villegas. 2021. Spanish language models. *arXiv preprint arXiv:2107.07253*.

Liebeskind, C., O. Elkayam, and S. Liebeskind. 2021. Jct at semeval-2021 task 1: Context-aware representation for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 138–143.

Liu, X., P. He, W. Chen, and J. Gao. 2019. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *arXiv preprint arXiv:1904.09482*.

Mc Laughlin, G. H. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Nandy, A., S. Adak, T. Halder, and S. M. Pokala. 2021. cs60075_team2 at semeval-2021 task 1: Lexical complexity prediction using transformer-based language models pre-trained on various text corpora. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 678–682.

Ortiz-Zambrano, J. A. and A. Montejo-Ráez. 2021. Complex words identification using word-level features for semeval-2020 task 1. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 126–129.

Ortiz-Zambranoa, J. A. and A. Montejo-Ráezb. 2020. Overview of alexs 2020: First workshop on lexical analysis at sepln. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*.

Paetzold, G. 2021. Utfpr at semeval-2021 task 1: Complexity prediction by combining bert vectors and classic features. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 617–622.

Paetzold, G. and L. Specia. 2016. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th Interna-*

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

Rayner, K. and S. A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201.

Rico-Sulayes, A. 2020. General lexicon-based complex word identification extended with stem n-grams and morphological engines. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain.*

Rojas, K. R. and F. Alva-Manchego. 2021. Iapucp at semeval-2021 task 1: Stacking fine-tuned transformers is almost all you need for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 144–149.

Ronzano, F., L. E. Anke, H. Saggion, et al. 2016. Taln at semeval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1011–1016.

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Shardlow, M. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109.

Shardlow, M., M. Cooper, and M. Zampieri. 2020. Complex: A new corpus for lexical complexity prediction from likert scale data. *arXiv preprint arXiv:2003.07008.*

Shardlow, M., R. Evans, G. H. Paetzold, and M. Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. *arXiv preprint arXiv:2106.00473.*

Shardlow, M., R. Evans, and M. Zampieri. 2021. Predicting lexical complexity in english texts. *arXiv preprint arXiv:2102.08773.*

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

Uluslu, A. Y. 2022. Automatic lexical simplification for turkish. *arXiv preprint arXiv:2201.05878.*

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vettigli, G. and A. Sorgente. 2021. Compna at semeval-2021 task 1: Prediction of lexical complexity analyzing heterogeneous features. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 560–564.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Yaseen, T. B., Q. Ismail, S. Al-Omari, E. Al-Sobh, and M. Abdullah. 2021. Just-blue at semeval-2021 task 1: Predicting lexical complexity using bert and roberta pre-trained language models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 661–666.

Zaharia, G.-E., D.-C. Cercel, and M. Dascalu. 2021. Upb at semeval-2021 task 1: Combining deep learning and hand-crafted features for lexical complexity prediction. *arXiv preprint arXiv:2104.06983.*

Zambrano, J. A. O. and A. Montejo-Ráez. 2021. Clexis2: A new corpus for complex word identification research in computing studies. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1075–1083.