International Journal of Combinatorial Optimization Problems and Informatics, 14(1), Jan-April 2023, 27-38. ISSN: 2007-1558.

www.editada.org

Exploring BERT-Based Pretrained Models for Polarity Analysis of Tweets in Spanish

Erick Barrios González[1], Mireya Tovar Vidal[1], José A. Reyes-Ortiz[2], Fernando Zacarias Flores[1], Pedro Bello López[1]

[1]Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, avenida San Claudio y 14 Sur, Ciudad Universitaria, Puebla 72570, Mexico

[2]Universidad Autónoma Metropolitana, Av. San Pablo Xalpa 180, Azcapotzalco, 02200 Mexico City, Mexico

erick.barrios@alumno.buap.mx, mireya.tovar@correo.buap.mx, jaro@azc.uam.mx, fzflores@yahoo.com.mx, pedro.bello@correo.buap.mx

Abstract. This paper reviews the implementation of three pre-trained models based on BERT ("bert-base-multilingual-cased", "IIC/beto-base-spanish-sqac" and "MarcBrun/ixambert-finetuned-squad-eu-en") to solve tasks 1.1 and 1.2 of "Workshop on Semantic Analysis at SEPLN 2020" (TASS 2020), these tasks consist of the polarity analysis of tweets in Spanish from different Spanish-speaking countries. The proposed models are evaluated individually by pre-processing and replacing synonyms. This research is carried out to find the points to improve in the polarity analysis of tweets (tweets), mainly in how the pre-trained models interpret words that are not in their vocabulary due to variations in the language, regional expressions, misspellings, and use of emojis.

Keywords: Polarity analysis, NLP, BERT.

Article Info
Received Sep 1, 2022
Accepted Jan 11, 2023.

1 Introduction

In 2021, Spanish was the second language with the most native speakers worldwide and the second most used on social networks such as Facebook, Instagram, LinkedIn, and Twitter [3]. Sentiment analysis, opinion extraction, or polarity analysis are subareas of natural language processing and are oriented to the study of feelings, emotions, and opinions expressed in a text; these texts focus on an object (a product, service, entity, organization, topic, or event).

Polarity analysis aims to detect whether a text expresses a positive, negative or neutral opinion and has become important because it is a tool that gives us more information about the opinion of many people. TASS ("Semantic Analysis at SEPLN") appeared in 2012, it was the first sentiment analysis task on Twitter for the Spanish language and from 2019 to 2020 it was part of IberLEF (Iberian Languages Evaluation Forum). In 2020, the name of TASS changed to "Workshop on Semantic Analysis at SEPLN" [19], encompassing more semantic processing tasks.

There are multiple approaches to solve the problems posed by sentiment analysis and polarity analysis as [22] shows, within these approaches is deep learning, which is a type of machine learning. Deep learning is currently the main strategy for solving sentiment analysis [22]. Deep learning consists of using neural networks (algorithms inspired by the functioning of the human brain) to learn tasks or learn large amounts of data [10].

Task 1.1 and task 1.2 of TASS 2020 address the analysis of polarity in tweets (tweets) in Spanish and will be the task to review in this article. These tasks' objective is to evaluate polarity classification systems of tweets written in Spanish (including its variants). Task 1.1 focuses on individually classifying each variation of Spanish according to the country where the tweets are from. For this task, the following variants are proposed by country: ES-Spain, PE-Peru, CR-Costa Rica, UR-Uruguay, MX-Mexico. TASS provides a training corpus and an assessment corpus, and any other corpus or linguistic resources besides those provided by TASS 2020 are allowed to be used. Task 1.2 focuses on classifying a corpus that contains together the different variants of Spanish mentioned in task 1.1, also the training corpus is the same used in task 1.1. The labels for the classification of the tweets are: "P": Positive, "N": Negative and "NEU": Neutral (Includes unclassified tweets).

This paper explores solutions for task 1.1 and task 1.2 of TASS 2020, implementing models (BERT) different from the models of related works. These models have been pre-trained with a more significant amount of information for the users. That will be shown in state of the art (contemplating a model specialized in multiple languages and a model specialized only in Spanish). An approach oriented to the pre-processing part of the tweets will be used (exploring the replacement of words that are not in the vocabulary of each model by synonyms). This paper is organized as follows: First, section 2 reviews the state of the art, section 3 shows the proposed solution and the implemented methodology, section 4 shows the experimental results and finally, in section 5 shows the conclusions.

2 State of the art

In 2020 of the 14 competitors in TASS 2020 [5], the best results are from the competitors Palomino and Ochoa [14], [8] and [4]. The authors of [14] propose a system to improve polarity classification in small data sets based on a high-performance language model (LM) called BERT (Bidirectional Encoder Representations from Transformers) (BERT variant to produce increased contextual data).

While [4] (UMUTeam) implemented three executions. The first run (LF + WE) consisted of linguistic features trained with a multilayer perceptron in combination with word embeddings trained with a convolutional neural network (CNN); the second run (LF) used the linguistic features trained with Support Vector Machines (SVM), and the third run (LF + SE) used the combination of linguistic features with sentence embeddings.

In [8] (ELiRF-UPV) use Deep Averaging Networks (DAN) as a baseline (these models consist of applying forward networks on text representations based on average word embeddings) and TWilBERT (a BERT-based framework for training, evaluating and tuning models in the Twitter domain), which was the best implemented system for polarity analysis at TASS 2020. TWilBERT was trained with 94 million pairs of tweets (47M positive and 47M negative).

Reviewing subsequent works referring to this task, we find that in the work of [22] there is a review of the state of the art from 2009 to 2021 for sentiment analysis, showing that machine learning is currently the most used tool for these tasks.
In the work of [21], BERT is used for sentiment analysis of Google Play comments, comparing it with the Support Vector Machine (SVM) algorithm and the Naïve Bayes classifier, showing that the BERT model obtains better results.

In the work of [12] the BERT model is compared with the BILSTM (Bidirectional Long Short-Term Memory) models, for the extraction of keywords for sentiment analysis, the best results were obtained with BERT.

Also, in the work by [17], BERT is compared with a model based on LSTM (Long Short-Term Memory) neural networks, obtaining better results with BERT.

It can be observed in the work of [15] comparing several models based on BERT for the analysis of sentiments and emotions. It mainly compares BERTWEET (Model based on BERT specialized in the Twitter domain) and BETO (Model based on BERT specialized in the Spanish language).

Another comparison between models based on BERT is found in the work of [24], where a comparison of several pre-trained BERT models for Spanish is made, obtaining better results with "BERT-base-spanish-wwm-uncased".

Recently in 2022, in works such as that of [6] or [18], the relevance of sentiment analysis on twitter can be observed, and despite the fact that it is observed that deep learning techniques will have good results as in the work of [18], there are other alternatives such as in [6]. While in the work of [2] it is possible to observe the advantages that pre-trained models have in sentiment analysis tasks. In [7] we can observe recommendations of characteristics to be covered when using a Twitter data set to identify emotions (using a model based on BERT).

As a novelty, this article will review pre-trained models with a greater amount of information than those shown in the state of the art and that have not been used in the analysis of polarity of tweets, as well as exploring the replacement of words by synonyms to improve the results in the tweet polarity analysis task.

3 Proposed solution

For the detection of polarity in tweets, the following stages are proposed:
1) Pre-processing of the corpus.
2) Solution architecture for polarity detection in tweets.
3) Programming and training of deep learning neural networks.
4) Evaluation of the results.

Each of these stages is described below.

3.1 Corpus pre-processing

For the development, training and evaluation of the system, the corpus provided on the official event page [19] will be used.
As a proposal for this article, it is proposed to separate the symbols from the words, to later tokenize and introduce each tweet in the pre-trained BERT model. The goal is to get the pre-trained model to correctly identify as many words as possible.
Another important part of pre-processing is changing usernames like "@Erick" and replacing them with the word "user". Hashtags such as "#10best" are also replaced with the word "Trend" (as is done in other polarity analysis subtasks (Garcia J. et al., 2020)). Specific cases within a hashtag such as "#sarcasm", where the word "sarcasm" would indicate that what is textually is not really expressed, as mentioned in the work of [10] for the detection of sarcasm.

One of the main problems when using the tokenizer of a pre-trained model is the fact that some words will not be found within the model's vocabulary (due to misspellings or regionalisms of Spanish, words in other languages, etc.), for that reason, it is proposed to detect which are the words that are not in the vocabulary and find a word that has a similar meaning that is within the vocabulary.

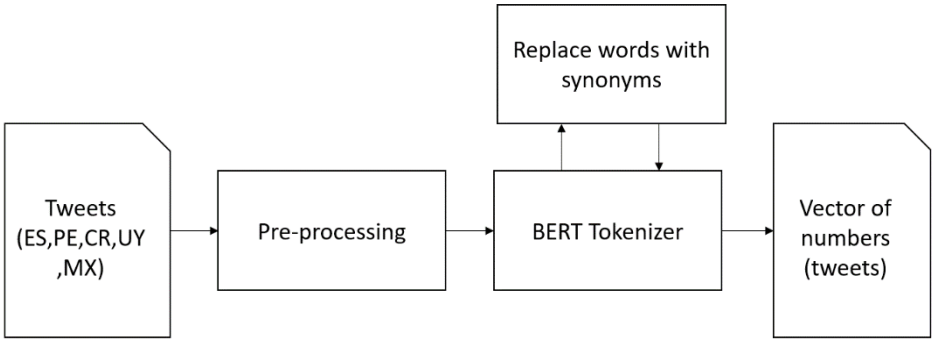

Fig. 1. Sequence of pre-processing of tweets.

In Figure 1 we can see that the tweets are preprocessed as explained above, then they go through the BERT tokenization tool, if the word that enters the tokenizer is in the model vocabulary, it is saved with its respective value. Otherwise, a synonym of that word found in the vocabulary is searched for replacement (with a script the synonym is obtained from the Wordreference [23]).
Regarding labeling, each class was converted to a positive integer, in this case 0 for "N", 1 for "NEU" and 2 for "P".

3.2 Solution architecture for polarity detection in tweets

For this stage, pre-trained models based on BERT are proposed to detect the polarity of sentiment in the tweets. For this task, it is proposed to use the following BERT-based pretrained models: "bert-base-multilingual-cased" [1], "IIC/beto-base-spanish-sqac" [9] and "MarcBrun/ixambert-finetuned-squad-eu-en" [13].

"Bert-base-multilingual-cased" is a model pretrained in 104 different languages with information from Wikipedia, it was pretrained with lowercase and tokenized texts using *WordPiece* [1], it also has a shared vocabulary size of 110,000 words.
"IIC/beto-base-spanish-sqac" It is a pre-trained model with a massive corpus of 570GB of clean and deduplicated text with 135 billion words extracted from web archives in Spanish between 2009 and 2019 [9] and later trained with the SQAC corpus (Spanish Question-Answering Corpus), this corpus contains 6,247 contexts and 18,817 questions with their answers, the source information

belongs to encyclopedic articles from Wikipedia in Spanish, Wikinoticias in Spanish, and texts from the corpus Spanish AnCora, which is a mixture of different sources of news and literature.

"IIC/beto-base-spanish-sqac" is based on RoBERTa-base (which is a BERT variation) and belongs to a family of Spanish language models, these models can be considered some of the biggest and best models around for Spanish [9].
"MarcBrun/ixambert-finetuned-squad-eu-en" is a multilingual model pretrained for English, Spanish and Basque. This model was trained with a corpus composed by documents from Wikipedia in English, Spanish and Basque, fine-tuned on SQuAD v1.1 and an experimental version of SQuAD1.1 in Basque (1/3 size of original SQuAD1.1). It can answer basic factual questions in English, Spanish and Basque.

Figure 2 shows the architecture that will be used to specialize the pre-trained BERT models in the polarity analysis task. Within this architecture, a dropout layer with a dropout rate of 0.5 and a layer that adds a linear transformation function (Linear layer) are added.

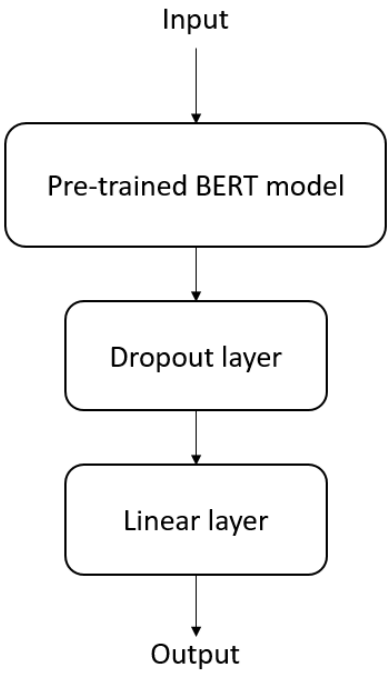

Fig. 2. Architecture for BERT pre-trained model specialization.

3.3 Programming and training of deep learning neural networks

The chosen pre-trained models (BERT) will be re-trained with the pre-processed corpus to specialize the model in detecting polarity in each tweet. For this, libraries such as Pytorch [16], Tensorflow [20], and Keras [11] in Python will be used.
The training corpus will comprise the corresponding training corpus of each country.

The 90% training corpus will be used to train the model, while the remaining 10% will be used for validation.
When the best parameters in training are found, the trained model will be evaluated against the evaluation corpus provided by TASS 2020 to compare the results with those of the TASS 2020 competitors.

3.4 Evaluation of the results

The chosen pre-trained models (BERT) will be re-trained with the pre-processed corpus. To evaluate the results it is necessary to establish the evaluation criteria. For the evaluation, the following metrics will be used: Precision, Recall, F_1 and Macro-F_1 (macro average F_1 score).

The systems will be classified according to the Macro-F_1 metric, for example to task 1.1 this means that the measures of recall, precision and F_1 will be calculated individually for each class ("P","N" and "NEU"), considering each variation of Spanish (ES-Spain, PE-Peru, CR-Costa Rica, UR-Uruguay, MX-Mexico) and then the average of all F_1 measurements will be obtained.
In the case of task 1.2, the system would be classified according to the average of the F_1 measure for the classes ("P","N", and "NEU").

In Equation 1 the precision metric is defined, where TP corresponds to the correct answers and FP corresponds to the false answers.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{1}$$

In Equation 2 the completeness metric is defined, where TP corresponds to the correct answers and FN corresponds to the missing answers.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{2}$$

Finally, in Equation 3 the metric of F_1 that integrates the precision and recall metrics is defined. For the evaluation, the gold file provided on the official page of the event will be used (TASS, 2020). To obtain the Macro-F_1 score, the F_1 obtained for each class ("P", "N", and "NEU") will be averaged.

$$\text{F1} = 2 * \frac{((\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall}))} \tag{3}$$

4 Results

Below are the results found in the review of the corpus, the evaluation of the models used and at the end a comparison with the results of the TASS 2020 competitors.

4.1 Corpus description

Finally, in Equation 3 it can be seen how to obtain the F_1 based on the precision and recall metrics. The training and development sets include a list of tweets with their corresponding code, the text of the tweet, and its classification. The test set includes a list of tweets with the tweet code and text. The test set includes the code of the tweet and its correct classification.

Within the corpus, no tweet contains more than 240 characters and contains informal language (misspellings, emojis, onomatopoeia are common).

The distribution of tweets by country and their ranking is shown in Table 1 (for the training corpus).

Table 1. Distribution of tweets by country (training corpus).

Set/Classification	N	NEU	P	Total
Spain (ES)	475	297	354	1126
Costa Rica (CR)	310	246	221	777
Peru (PE)	228	522	216	966
Uruguay (UY)	367	286	290	943
Mexico (MX)	505	172	313	990

4.2 Results task 1.1 “bert-base-multilingual-cased”

In Table 2 you can see the evaluation of the 'bert-base-multilingual-cased' model, trained for each variation of Spanish, using non-tokenized text and no type of pre-processing. It can be seen that the highest values of F_1 are in the "positive" class with the exception of the corpus of Mexico (MX) and Peru (PE).), In addition, MX corpus have the highest F_1 measurement on average

Table 2. Evaluation of the “bert-base-multilingual-cased” model without pre-processing.

Corpus	F_1 (N)	F_1 (NEU)	F_1 (P)	F_1 (Average)
ES	0.60742	0.66666	**0.69433**	0.65613
MX	**0.68192**	0.62415	0.68005	0.66204
PE	0.53063	**0.66912**	0.66009	0.61994
CR	0.60514	0.66309	**0.67334**	0.64719
UY	0.63769	0.65435	**0.67971**	0.65725
Average	0.61256	0.65547	**0.67750**	0.64851

Table 3 shows the evaluation of the "bert-base-multilingual-cased" model, trained for each variation of Spanish, pre-processing and tokenizing the text. As in Table 3, it can be seen that the highest values of F_1 are in the "positive" class, however, it can also be seen that the exception is the corpus of Mexico (MX).

Table 3. Evaluation of the “bert-base-multilingual-cased” model with pre-processing.

Corpus	F_1 (N)	F_1 (NEU)	F_1 (P)	F_1 (Average)
ES	0.75598	0.62877	**0.76793**	0.71756
MX	**0.80107**	0.53364	0.76642	0.70037
PE	0.72768	0.69387	**0.73666**	0.71940
CR	0.72684	0.65316	**0.72806**	0.70268
UY	0.74116	0.60194	**0.75645**	0.69985
Average	0.75054	0.62227	**0.75110**	0.70797

Table 4 shows the evaluation of the "bert-base-multilingual-cased" model, trained for each variation of Spanish, pre-processing, tokenizing the text and replacing words not found in the vocabulary with synonyms that were found. The highest values coincide with those of Table 3. In each respective table it can be seen that corpus that benefited from the pre-processing was Peru, while the corpus from Mexico was the one that took the worst advantage of the replacement of synonyms.

Table 4. Evaluation of the “bert-base-multilingual-cased” model with replacement by synonyms.

Corpus	F_1 (N)	F_1 (NEU)	F_1 (P)	F_1 (Average)
ES	0.75383	0.62006	**0.75959**	0.71116
MX	**0.78877**	0.49945	0.76642	0.68488
PE	0.72388	0.68686	**0.74358**	0.71810
CR	0.74032	0.65955	**0.74090**	0.71359
UY	0.74681	0.59330	**0.76260**	0.70090
Average	0.75072	0.61184	**0.75461**	0.70572

Finally, for this model it is important to mention that 363 words were successfully replaced (0.76% of the vocabulary), while 47,230 words could not be replaced (out of 47,593 words not found in the vocabulary).

4.3 Results task 1.1 "IIC/beto-base-spanish-sqac"

Table 5 shows the evaluation of the "IIC/roberta-base-spanish-sqac" model, trained for each variation of Spanish, using non-tokenized text and no type of pre-processing. The highest values of F_1 are in the "positive" class, except for the corpus of Peru (PE), like the previous model ("bert-base-multilingual-cased"), the corpus of Mexico (MX) obtains a better F_1 on average.

Table 5. Evaluation of the "IIC/roberta-base-spanish-sqac" model without pre-processing.

Corpus	F_1 (N)	F_1 (NEU)	F_1 (P)	F_1 (Average)
ES	0.60742	0.66666	**0.69433**	0.65613
MX	0.61256	0.65547	**0.67750**	0.66204
PE	0.53063	**0.66912**	0.66009	0.61994
CR	0.60514	0.66309	**0.67334**	0.64719
UY	0.63769	0.65435	**0.67971**	0.65725
Average	0.61256	0.65547	**0.67750**	0.64851

Table 6 shows the evaluation of the "IIC/roberta-base-spanish-sqac" model, trained for each variation of Spanish, pre-processing and tokenizing the text. Unlike Table 5, the highest values are concentrated in the negative class (N), and the corpus that best takes advantage of the pre-processing is Spain (ES).

Table 6. Evaluation of the "IIC/roberta-base-spanish-sqac" model with pre-processing.

Corpus	F_1 (N)	F_1 (NEU)	F_1 (P)	F_1 (Average)
ES	0.79306	0.64183	**0.80378**	0.74622
MX	**0.81599**	0.50328	0.80030	0.70652
PE	**0.76923**	0.70375	0.76517	0.74605
CR	**0.78663**	0.67762	0.77241	0.74555
UY	**0.78663**	0.67762	0.77241	0.74555
Average	**0.79030**	0.64082	0.78281	0.73798

Table 7 shows the evaluation of the "IIC/roberta-base-spanish-sqac" model, trained for each variation of Spanish, pre-processing, tokenizing the text and replacing words not found in the vocabulary with synonyms. This table shows that the replacement of synonyms improves the results in 3 of the 5 corpora analyzed (ES, MX and CR).

Table 7. Evaluation of the "IIC/roberta-base-spanish-sqac" model with replacement by synonyms.

Corpus	F_1 (N)	F_1 (NEU)	F_1 (P)	F_1 (Average)
ES	0.79448	0.64739	**0.80216**	0.74801
MX	**0.81599**	0.51627	0.79847	0.71024
PE	0.75899	0.70284	**0.76048**	0.74077
CR	**0.78933**	0.68959	0.77867	0.75253
UY	0.78006	0.63211	**0.80170**	0.73795
Average	0.78777	0.63764	**0.78829**	0.73790

Finally, for this model it is important to mention that 243 words were successfully replaced (0.28% of the vocabulary), while 85,679 words could not be replaced (out of 85,922 words not found in the vocabulary).

4.4 Results task 1.1 "MarcBrun/ixambert-finetuned-squad-eu-en"

Table 8 shows the evaluation of the "MarcBrun/ixambert-finetuned-squad-eu-en" model, trained for each variation of Spanish, pre-processing and tokenizing the text. The highest values are distributed between the "negative" and "positive" classes, while the corpus with the highest average F_1 is Costa Rica (CR).

Table 8. Evaluation of the "MarcBrun/ixambert-finetuned-squad-eu-en" model with pre-processing.

Corpus	F_1 (N)	F_1 (NEU)	F_1 (P)	F_1 (Average)
ES	0.75426	0.62657	**0.76447**	0.71510
MX	**0.80758**	0.54054	0.76419	0.70410
PE	0.72334	0.68991	**0.73479**	0.71601
CR	**0.75680**	0.67639	0.74501	0.72606
UY	0.76882	0.63644	**0.77649**	0.72725
Average	**0.76216**	0.63397	0.63397	0.75699

Table 9 shows the evaluation of the "MarcBrun/ixambert-finetuned-squad-eu-en" model, trained for each variation of Spanish, pre-processing, tokenizing the text and replacing words not found in the vocabulary with synonyms. The highest values are found in the "positive" class (with the exception of the Mexico corpus), while the corpus with the highest F_1 average is Spain (ES).

Table 9. Evaluation of the "MarcBrun/ixambert-finetuned-squad-eu-en" model with replacement by synonyms.

Corpus	F_1 (N)	F_1 (NEU)	F_1 (P)	F_1 (Average)
ES	0.76250	0.62265	**0.77748**	0.72087
MX	**0.80280**	0.51685	0.77319	0.69761
PE	0.72065	0.68904	**0.74422**	0.71797
CR	0.74573	0.66967	**0.74583**	0.72041
UY	0.75546	0.62784	**0.77201**	0.71843
Average	0.75742	0.62521	**0.76254**	0.71506

Finally, for this model it is important to mention that 238 words were successfully replaced (0.76% of the vocabulary), while 30,779 words could not be replaced (out of 31,017 words not found in the vocabulary).

4.5 Results task 1.2

Table 10 shows the results obtained with the pretrained model "bert-base-multilingual-cased" in task 1.2. In Table 10 it can be seen that the results obtained from the "NEU" label are those that obtain the highest result.

Table 10. Results with "bert-base-multilingual-cased" in task 1.2 with pre-processing.

Corpus	F_1 (N)	F_1 (NEU)	F_1 (P)	F_1 (Average)
Task 1.2	0.58369	**0.70534**	0.67264	0.65389

Table 11 shows the results obtained with the pretrained model "IIC/roberta-base-spanish-sqac" in task 1.2. The average F_1 is lower than that shown in Table 10 and higher than that of Table 12

Table 11. Results with "IIC/roberta-base-spanish-sqac" in task 1.2 with pre-processing.

Corpus	F_1 (N)	F_1 (NEU)	F_1 (P)	F_1 (Average)
Task 1.2	0.51079	**0.68880**	0.62586	0.60848

Table 12 shows the results obtained with the pretrained model "MarcBrun/ixambert-finetuned-squad-eu-en" in task 1.2.

Table 12. Results with "MarcBrun/ixambert-finetuned-squad-eu-en" in task 1.2 with pre-processing.

Corpus	F_1 (N)	F_1 (NEU)	F_1 (P)	F_1 (Average)
Task 1.2	0.49771	**0.68155**	0.62674	0.60200

In Tables 10, 11 and 12 it can be seen that they have in common that the highest result is obtained with the "NEU" label and the second best result is obtained with the label "P". It is important to highlight that these results were obtained only using pre-processing and without the replacement of synonyms. Unlike the experimentation in task 1.1 where the "IIC/roberta-base-spanish-sqac" model obtained the best results, the "bert-base-multilingual-cased" model obtained the highest F_1, this could be due to the fact that when training and evaluating the different corpora together, this model assigns values to each word in such a way that they do not decline to such a specific meaning, thanks to the fact that the model was trained with a fairly neutral Spanish [1].

4.6 Comparison of TASS 2020 results for task 1.1

In Table 13 we can see the best results of the TASS 2020 competitors in task 1.1. It can be seen that the best results are obtained in the corpus of Mexico (MX) and the corpus of Uruguay (UY) by both competitors.

Table 13. Best results (F_1) of TASS 2020 competitors.

Corpus	*TASS* 2020 (F_1)	Competitors
ES	0.61008	ELiRF-UPV [8]
MX	0.63621	ELiRF-UPV [8]
PE	0.60315	ELiRF-UPV [8]
CR	0.61148	ELiRF-UPV [8]
UY	0.62808	Palomino-Ochoa [14]

Below in Table 14 you can see the comparison of all the implemented proposals (for "bert-base-multilingual-cased") with the best results of TASS 2020, as can be seen, an improvement has been achieved with pre-processing proposed, and in some cases the F_1 is improved with the replacement of synonyms (after pre-processing).

Table 14. Comparison "bert-base-multilingual-cased" with better results (F_1) from TASS 2020.

Corpus	TASS 2020 (F_1)	BERT (Multi) No pre-processing (F_1)	BERT (Multi) With pre-processing (F_1)	BERT (Multi) Replacing synonyms (F_1)
ES	0.61008	0.65613	**0.71756**	0.71116
MX	0.63621	0.66204	**0.70037**	0.68488
PE	0.60315	0.61994	**0.71940**	0.71810
CR	0.61148	0.64719	0.70268	**0.71359**
UY	0.62808	0.65725	0.69985	**0.70090**

Table 15. Comparison "IIC/roberta-base-spanish-sqac" with better results (F_1) of TASS 2020.

Corpus	TASS 2020 (F_1)	IIC/roberta No pre-processing (F_1)	IIC/roberta With pre-processing (F_1)	IIC/roberta Replacing synonyms (F_1)
ES	0.61008	0.63109	0.74622	**0.74801**
MX	0.63621	0.66419	0.70652	**0.71024**
PE	0.60315	0.61011	**0.74605**	0.74077
CR	0.61148	0.63243	0.74555	**0.75253**
UY	0.62808	0.63873	**0.74555**	0.73795

Table 15 shows the comparison of all the implemented systems (with the "IIC/roberta-base-spanish-sqac" model) with the best results of TASS 2020, as well as the previous model ("bert-base-multilingual-cased") an improvement is obtained in the F_1 with the proposed pre-processing, and in some cases the F_1 is improved with the replacement of synonyms (after the pre-processing).

Table 16. Comparación "MarcBrun/ixambert-finetuned-squad-eu-en" con mejores resultados (F_1) de TASS 2020.

Corpus	TASS 2020 (F_1)	ixambert-finetuned-squad-eu-en With pre-processing (F_1)	ixambert-finetuned-squad-eu-en Replacing synonyms (F_1)
ES	0.61008	0.71510	**0.72087**
MX	0.63621	**0.70410**	0.69761
PE	0.60315	0.71601	**0.71797**
CR	0.61148	**0.72606**	0.72041
UY	0.62808	**0.72725**	0.71843

Table 16 shows the comparison of all the implemented systems (with the "MarcBrun/ixambert-finetuned-squad-eu-en" model) with the best results of TASS 2020. It can be seen that their results, despite being very similar to those of "bert-base-multilingual-cased", "MarcBrun/ixambert-finetuned-squad-eu-en" exceeds the F_1 score in some variations of Spanish, such as that of Uruguay (UY). It also appears that in most cases the model does not benefit from synonym replacement.

Table 17. Comparison of the best results of each proposed model, with the best results (F_1) of TASS 2020.

Corpus	TASS 2020 (F_1)	IIC/Roberta (F_1)	BERT (Multi) (F_1)	Ixambert (F_1)
ES	0.61008	**0.74801**	0.71756	0.72087
MX	0.63621	**0.71024**	0.70037	0.70410
PE	0.60315	**0.74605**	0.71940	0.71797
CR	0.61148	**0.75253**	0.71359	0.72606
UY	0.62808	**0.74555**	0.70090	0.72725

Finally, in Table 17 we can see a comparison of the best results obtained with each proposed model, the "IIC/roberta-base-spanish-sqac" model obtained the best results in these tests, because it is specialized for texts in Spanish. , with "bert-base-multilingual-cased" were found more in his vocabulary, however, many of these words are not Spanish words (because it was disturbed for multiple languages), something similar happens with "MarcBrun/ ixambert- finetuned-squad-eu-en" which is specialized in English, Spanish and Basque.

On the other hand, the improvement between TWilBERT (the best BERT model in TASS 2020) and the models proposed in this article is due to the amount and variety of training information, being that TWilBERT was trained mainly with tweets and "IIC /roberta-base-spanish-sqac" was pre-trained with a massive corpus of 570GB [9].

4.7 Comparison of TASS 2020 results for task 1.2

To compare the results obtained with the results of the TASS 2020 competitors, Table 18 shows the best results of task 1.2, it can be seen that the best result is obtained by the competitor Palomino-Ochoa [14] with 0.49796 of F_1.

Table 18. TASS 2020 Best Competitor for Task 1.2.

Competitor	F_1	Presicion	Recall
Palomino-Ochoa [14]	**0.49796**	**0.48685**	**0.50959**
ELiRF-UPV [8]	0.35787	0.35866	0.35709

Finally, comparing the results obtained with the results of the TASS 2020 competitors, we can observe in Table 19 that the results of all the proposed models exceed the best result of TASS 2020 in task 1.2, observing that the best result is obtained by the model "IIC/roberta-base-spanish-sqac", which is a model specialized in the Spanish language, being that both in task 1.1 and in task 1.2 the best performance was obtained by the "IIC/roberta-base-spanish-sqac" model.

Table 19. Comparison results with TASS 2020.

Corpus	TASS 2020 (F_1)	IIC/Roberta (F_1)	BERT (Multi) (F_1)	Ixambert (F_1)
Task 1.2	0.49796	**0.65389**	0.60848	0.60200

4.8 Discussion of results

In the results it can be seen that the "IIC/roberta-base-spanish-sqac" model obtained the best results, this is mainly due to the fact that it is a model specialized in the Spanish language, since the training corpus and the test are entirely in Spanish. Another important detail to highlight is that it is the model with the most words not found (85,922), which can mean a great opportunity for improvement when expanding the model's vocabulary.

5 Conclusions

In order to solve the polarity analysis in Spanish tweets, pre-trained models based on BERT ("IIC/beto-base-spanish-sqac", "bert-base-multilingual-cased" and "MarcBrun/ ixambert- finetuned-squad-eu-en"). The results obtained managed to exceed the score of the TASS 2020 participants, this is attributed to the fact that the models used were trained with a much larger data set than the best model of the participants (TWilBERT).

It was observed that "IIC/beto-base-spanish-sqac", obtained better results than "bert-base-multilingual-cased" and "MarcBrun/ ixambert-finetuned-squad-eu-en", because most of the Words found in the model vocabulary are within a Spanish context, since "bert-base-multilingual-cased" and "MarcBrun/ixambert-finetuned-squad-eu-en" are specialized in multiple languages.
The proposed pre-processing significantly improves all tests from different regions, and synonym replacement only gives hundredths improvements in some regions because the synonym of a word is not always the same in different regions.
As future work, it is proposed to improve the process of replacing words not found since the number of words replaced was less than 1% in all models, and synonyms differ in all regions. Also, implement other architectures to improve the results in model training, such as recurrent neural networks or convolutional neural networks.

Based on the implementation of the models in this work, it has been concluded that it is better to use specialized models in the Spanish language for this task. Due to the limited number of specialized models in this language, it is proposed as a job if it is possible to collect texts from other social networks such as Facebook, WhatsApp, Youtube, Instagram, and Telegram, among others, to obtain a broader vocabulary.

References

1. Devlin, J., Wei Chang, M., Lee, K., y Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Association for Computational Linguistics, Minneapolis, Minnesota, Vol. 1, (2018) 4171–4186.
2. Estienne, L., Vera, M., Rey Vega L.: Cross-domain Sentiment Classification in Spanish," 2022 IEEE Biennial Congress of Argentina (ARGENCON), 2022, pp. 1-7.
3. Fernández, D.: El español: una lengua viva. Informe 2021", Instituto Cervantes, (2021), 5-6.
4. García, J., Almela, A., Valencia, R.: UMUTeam at TASS 2020: Combining Linguistic Features and Machine-learning Models for Sentiment Classification, Proceedings172 of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, 2020, 179-186.
5. García, M., Díaz, M., Plaza, M., Montejo, A., Jiménez, S., Martínez, E., Aguilar, A., Sobrevilla, M., Chiruzzo, L., Moctezuma, D. : Overview of TASS 2020: Introducing Emotion Detection, Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, 2020, 163-170.
6. García, P., Sánchez, I., Pontiel, D., González, J.L.: A novel flexible feature extraction algorithm for Spanish tweet sentiment analysis based on the context of words, Expert Systems with Applications, Vol. 212, 2023. 118817.
7. González, I.: Procesamiento del lenguaje natural con BERT: Análisis de sentimientos en tuits, Universidad Carlos III de Madrid, (2020).
8. González, J., Moncho, J., A., Hurtado, L.: ELiRF-UPV at TASS 2020: TWilBERT for Sentiment Analysis and Emotion Detection in Spanish Tweets, Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, 2020 ,179-186.
9. Gutiérrez, A., Armengol, J., Pámies, M., Llop, J., Silveira, J., Carrino, P., Gonzalez, A., Armentano, C., Rodríguez, C., Villegas, M.: MarIA: Spanish Language Models, CoRR, vol. 68, 2021, 39-60.
10. Hernández, J.C.: Aprendizaje Profundo y Neuroevolución para el Análisis de Sentimientos en Tweets Escritos en Español Mexicano, [Tesis de maestria, Universidad Veracruzana]. (2021).
11. Keras, https://keras.io/getting_started/
12. Mazo, J.D.: Detección de palabras clave en el análisis de sentimiento de tweets usando técnicas de ML", [Trabajo de grado especialización, Universidad de Antioquia], (2021).
13. Otegi, A., Agirre, A., Campos, J., Soroa, A., Agirre, E.: Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque, European Language Resources Association, vol. Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, 436-442.
14. Palomino, D., Ochoa, J.: Palomino-Ochoa at TASS 2020: Transformer-based Data Augmentation for Overcoming Few-Shot Learning, Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, 2020, 171-178.
15. Pérez, J., Manuel, Giudici, Carlos, y Luque, F.: Pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks, arXiv, vol. 1, 2021, 1-4.
16. Pytorch, https://pytorch.org/features/
17. Scola, E. y Segura, I.: Sarcasm Detection with BERT, Procesamiento del Lenguaje Natural, Madrid, Legan´es, Spain, vol. 67, 2021, 13-25.
18. Sunitha, D., Patra, R.K., Babu, N.V., Suresh, A., Gupta S.: Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in India and European countries, Pattern Recognition Letters, vol. 158, 2022, 164-170.
19. Workshop on Semantic Analysis at SEPLN, http://tass.sepln.org/2020/
20. TensorFlow, https://www.tensorflow.org/?hl=es-419
21. López, C., Gonzales, S., Orlando: Análisis de sentimiento de comentarios en español en Google Play Store usando BERT, scielocl, vol. 29, 2021, 557 - 563.
22. Valladares, J.G. Análisis de sentimientos para textos cortos en español, una revisión del estado del arte, Universidda Politécnica Salesiana Sede Quito, 2022, 4-14.
23. Wordreference, https://www.wordreference.com/sinonimos/
24. Zárate, G.H.: Análisis de sentimientos en información de medios periodísticos y redes sociales mediante redes neuronales recurrentes", [Tesis para obtener el título profesional de Ingeniero, Pontificia Universidad Católica del Perú], (2021).