ISSN 2007-9737

# Sentiment Analysis of Public Opinion in Spanish Speaking Countries Using BERT

Efrain Macedo González, Yulia Ledeneva*, Rene Arnulfo García

Universidad Autónoma del Estado de México, Santiago Tianguistenco, Mexico

efrainmacedo40@gmail.com, renearnulfo@hotmail.com, yledeneva@yahoo.com

**Abstract.** This study focuses on Sentiment Analysis (SA) specifically applied to the Spanish-speaking variant, using the pre-trained linguistic architecture Bidirectional Encoder Representations from Transformers (BERT). The effectiveness of the BERT architecture for detecting sentiment polarity in Spanish was explored. An experimental study was conducted using the TASS 2019 corpus, which included tweets from Spanish-speaking countries (Mexico, Costa Rica, Spain, Peru, and Uruguay). After cleaning the texts, a BERT model was refined to classify three sentiment polarities (positive, negative, and neutral). Traditional oversampling and synthetic data generation using ChatGPT were applied to correct imbalances in the classes analyzed. The model achieved 87% accuracy in the Mexican sample with balancing using synthetic data. However, the most innovative finding was that balancing with oversampling allowed 97% accuracy with balanced metrics, surpassing the generalizability of the previous methods. Oversampling balancing is the most robust strategy for understanding digital opinions. This approach allows machines to capture regional linguistic richness, facilitating informed strategic decision-making.

**Keywords.** Sentiment analysis, BERT, oversampling.

## 1 Introduction

In the digital era, interactions on online platforms and social networks have generated a substantial amount of data that reflects not only opinions but also emotions expressed during significant moments. From enthusiasm for a new product launch to indignation over global events and opinions expressed on social media.

This is where sentiment analysis (SA), also known as opinion mining or polarity analysis, plays a fundamental role. It is a key field within Natural Language Processing (NLP) focused on identifying sentiments, emotions, and opinions expressed in text [1]. This discipline has become an essential tool for understanding human behavior in relation to social events and relevant phenomena, as well as for market analysis and trend forecasting [1, 2].

Digital texts are generally classified into three main categories: positive, negative, and neutral [3].

Sentiment analysis provides valuable insights into public mood across different contexts, facilitating strategic and well-informed decision- making.

The development of sentiment analysis techniques in English has a well-established trajectory, with successful studies, particularly in political communication on Twitter, yielding relevant conclusions related to influence analysis [4].

Beyond English, these techniques have also been applied to other languages such as Portuguese, achieving satisfactory results [5].

Cervantes I. and Pastor C. [6] note that the development of these techniques applied to Spanish is still in a maturation phase.

The relevance of Spanish as a global language is undeniable. In 2021, Spanish was the second most spoken native language worldwide after Chinese and the second most used language on social networks such as Facebook, Instagram, LinkedIn, and Twitter [6].

Nearly 493 million people speak Spanish, and the global group of potential Spanish users (including native speakers, speakers with limited proficiency, and foreign language learners) exceeds 591 million, representing 7.5% of the world population. Projections indicate that by 2060, United States will be the second-largest Spanish-

speaking country, after Mexico, with 27.5% of its population being of Hispanic origin. Mexico is currently the largest Spanish-speaking country by number of speakers [6].

Given the vitality and global reach of Spanish, sentiment analysis in this language—and specifically in regional variants such as Mexican Spanish—is of great importance for capturing and understanding linguistic and cultural particularities expressed in text [7].

In this context, pre-trained language models based on the transformer architecture, such as BERT (Bidirectional Encoder Representations from Transformers), have revolutionized the field of NLP, achieving state-of-the-art results in multiple tasks. BERT distinguished by its bidirectional pre-training capability, allowing it to integrate both left and right context of a word. When fine-tuned for specific tasks, these models demonstrate a deep understanding of context and linguistic nuances [8,9], even in low-resource scenarios for languages other than English [10].

This article focuses on sentiment analysis in Mexican Spanish using BERT, aiming to explore the effectiveness of this architecture in polarity detection within the specific context of this linguistic variant. The study seeks to contribute to understanding how BERT-based models can adapt and provide accurate results in specific Spanish domains, such as Mexican Spanish, which exhibits its own lexical and pragmatic characteristics.

## 2 Related Work

Sentiment analysis has undergone significant evolution, transitioning from lexicon-based approaches and traditional machine learning methods to deep learning models and large language models (LLMs) [1].

Traditional and Machine Learning Approaches:

Initially, sentiment analysis was addressed through lexical resources that assign polarity values (positive, negative, neutral) or emotions to words [2]. These resources include opinion lexicons such as SenticNet, VADER, or adaptations of Affective Norms for English Words (ANEW); opinion lexicons such as Hu and Liu's Opinion Lexicon or the MPQA Subjectivity Lexicon by Wilson and Wiebe; domain-specific lexicons such as those proposed by Loughran and McDonald, the Stock Market Lexicon, and SentiWordNet; as well as the NRC Hashtag Emotion Lexicon [1,11].

However, the effectiveness of these lexicons is limited due to their difficulty in adapting to specific domains, detecting negation, sarcasm, or ambiguity, and handling the informal nature of language in social media [11]. In this regard, Molina and González [12] discuss how lexicons alone are not fully effective, since each domain has specific ways of expressing opinions, and they often fail to handle phenomena such as negation or sarcasm.

With the advancements of machine learning (ML), methods based on feature extraction and classifiers such as Support Vector Machines (SVM), Multinomial Naïve Bayes, decision trees, AdaBoost, and logistic regression emerged. These methods have been applied to sentiment classification tasks in Spanish, including tweet analysis, [11]. Previous research also explored SVM using features based on unigrams and favorability measures, or n-grams with different weighting schemes, demonstrating the usefulness of SVM for predicting sentiment on Twitter in a fast and accurate manner. Colas and Brazdil [13] also compared SVM with other classification algorithms in text classification tasks.

BERT in Sentiment Analysis for Spanish:

Despite Spanish being one of the most widely spoken languages in the world, the availability of specific resources for training and evaluating language models has been a challenge [14]. Nevertheless, pre-trained BERT models exclusively for Spanish have been developed, such as BETO (Bidirectional Encoder Representations from Transformers from scratch for Spanish) by Cañete et al. [14]. BETO has demonstrated superior performance in most Spanish NLP tasks, including sentiment analysis, compared to multilingual BERT models (mBERT).

Regarding the use of the BERT language model in sentiment analysis, S. Elmitwalli and J. Mehegan [15] presented an evaluation comparing several sentiment analysis techniques (lexicon-based, machine learning, Bi-LSTM, BERT, and GPT-3) on standard IMDB and Sentiment140 datasets using standard evaluation metrics such

as accuracy, F1-score, and precision. BERT achieved the highest F1-score (IMDB: 0.9380, Sentiment140: 0.8114), followed by GPT-3 (IMDB: 0.9119, Sentiment140: 0.7913) and Bi-LSTM (IMDB: 0.8971, Sentiment140: 0.7778).

Barrios González et al. [8] explored BERT-based pre-trained models such as bert-base-multilingual-cased, IIC/beto-base-spanish-sqac (a RoBERTa-based variant for Spanish), and MarcBrun/ixambert-finetuned-squad-eu-en for polarity analysis in Spanish tweets from various regional variants (Spain, Peru, Costa Rica, Uruguay, and Mexico). Their results indicated that for the Mexican corpus (MX), the bert-base-multilingual-cased model achieved a higher average F1-score, and that the Spanish-specialized model IIC/beto-base-spanish-sqac outperformed the results of the TASS 2019 dataset for Mexico after preprocessing.

In another study, K. I. Roumeliotis et al. [16] presented a comparison between GPT-4, BERT, and FinBERT using a dataset of 31,037 rows and six columns consisting of cryptocurrency news obtained from Kaggle. The dataset was subsequently preprocessed. For training BERT and FinBERT, 5,000 rows were randomly selected. In the case of GPT-4, the process involved model tuning, prediction, comparison with the dataset labels, and final integration of the results, reporting metrics such as accuracy, precision, recall, and F1-score. For BERT and FinBERT, the process was similar, except that instead of fine-tuning, the models were trained using a sample of the selected rows. The results showed that GPT-4 achieved the best performance with an accuracy of 86.7%, followed by FinBERT with 84.3% and BERT with 83.3%. This study demonstrated that GPT-4 is the most suitable model for accurately interpreting and categorizing sentiments extracted from cryptocurrency news articles.

The results also showed that using augmented data generated by ChatGPT significantly improved the performance of sentiment analysis models. In the PerSenT (Political Perspective Sentiment Corpus) dataset specialized dataset for perspective-conditioned sentiment analysis tasks, primarily designed to study how sentiment expressed in a text, changes depending on the political entity or group being referenced, the augmented models achieved better accuracy and F1-macro metrics compared to the reference models. For example, RoBERTa-small showed an improvement in F1 macro score from 36% to 40% and in accuracy from 38% to 41% when augmented data was used. For RoBERTa-base, accuracy increased from 39% to 46% and the F1-macro score from 38% to 43%. The XtremeDistil model, being the most resource-efficient, also demonstrated notable improvements, with an increase in accuracy from 43% to 46%. In the MultiEmo dataset, designed for emotion and sentiment analysis tasks in text, focusing on multi-class and multi-label scenarios, the results were equally positive, with improvements in both accuracy and F1 macro score across all models. For example, RoBERTa-small improved its accuracy from 78% to 85% and its F1 macro score from 78% to 84%. Gómez-Adorno et al. [7] conducted an extensive evaluation of machine learning and deep learning models on the SENT-COVID corpus of Spanish tweets from Mexico. Their study included BERT models pre-trained in Spanish and found that BETO-uncased achieved the highest accuracy (73.26%), highlighting its deep understanding of context and the nuances of Mexican Spanish discourse, acquired through extensive pre-training on a wide variety of text corpora.

The study by C. Suhaeni and H. S. Yong [17] employed an approach to improve the efficiency and performance of sentiment analysis models through data augmentation using ChatGPT. The methodology focused on generating synthetic data to train smaller, less resource-intensive models, making them competitive with larger counterparts. The data augmentation process was carried out using OpenAI's GPT-3.5 model through its API, applying two main strategies: paraphrasing and inspirational generation. In paraphrasing, varied representations of original texts were generated to preserve contextual relevance. In inspirational generation, entirely new content was created based on the original topic while maintaining the sentiment, thereby expanding the data scope. Four specific prompts were used to cover these strategies, applied to samples from two sentiment analysis datasets: PerSenT and MultiEmo. The trained models included RoBERTa-small, RoBERTa-base, and XtremeDistil, fine-tuned with combinations of original and augmented data, and

their performance was evaluated in terms of precision and macro F1-score.

## 3 Proposed Methodology

Data preprocessing is proposed prior to feeding the algorithms. The corpora from Mexico, Spain, Peru, Costa Rica, and Uruguay are thoroughly cleaned to remove links, hashtags at the end of sentences, and punctuation, allowing the algorithms to better understand the text and improve prediction performance, as detailed below.

### 3.1 Preprocessing

For data cleaning, the following were removed:

- Links, mentions, and line breaks
- Characters that did not conform to UTF-8/ASCII
- Hashtags were cleaned, preserving the sentence and removing only “#” symbol
- Special characters such as "&" and "$" present in some words were filtered out
- Multiple consecutive spaces were removed.

Once the texts were cleaned, two new columns were created, for both the training and test sets to store the cleaned version of the tweet text. To store the length of the cleaned text, to verify whether excessive text was removed during cleaning or whether the tweet was almost eliminated.

It was observed that in each corpus there are cleaned tweets with fewer than 10 words, which is a result of the preprocessing steps described above. This indicates that some tweets originally contained only mentions, hashtags, and links, which were removed. However, no content was removed from the dataset.

Subsequently, an additional cleaning step was applied to both the training and test data by examining the tokenized version of the sentences. This process was carried out using the BERT tokenizer.

The categories were encoded numerically as shown in Tables 1, 2, 3, 4, and 5. Only three possible polarities were defined: Negative (0), Neutral (1), and Positive (2).

**Table 1.** Polarity Encoding of Texts in the Mexico Training Dataset

| Polarity | Number of texts |
|---|---|
| 0 | 505 |
| 2 | 312 |
| 1 | 172 |

**Table 2.** Polarity Encoding of Texts in the Spain Training Dataset

| Polarity | Number of texts |
|---|---|
| 0 | 474 |
| 2 | 354 |
| 1 | 297 |

**Table 3.** Polarity Encoding of Texts in the Peru Training Dataset

| Polarity | Number of texts |
|---|---|
| 0 | 522 |
| 2 | 228 |
| 1 | 216 |

**Table 4.** Polarity Encoding of Texts in the Costa Rica Training Dataset

| Polarity | Number of texts |
|---|---|
| 0 | 310 |
| 2 | 246 |
| 1 | 221 |

**Table 5.** Polarity Encoding of Texts in the Uruguay Training Dataset

| Polarity | Number of texts |
|---|---|
| 0 | 367 |
| 2 | 290 |
| 1 | 286 |

**Table 6.** Parameters Used in the BERT Model

| Parameters | |
|---|---|
| Activation Function | Softmax |
| Optimizer | Adam |
| Lost | CategoricalCrossEntropy |
| Epochs | 10 |
| Accuracy | CategoricalAccuracy |

**Table 7.** Classification Report of the Model for the Mexico Dataset

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Negative | 0.98 | 0.93 | 0.96 | 505 |
| Neutral | 0.86 | 0.95 | 0.60 | 172 |
| Positive | 0.95 | 0.96 | 0.95 | 312 |
| Accuracy | | | 0.95 | 989 |
| Macro avg | 0.93 | 0.95 | 0.94 | 989 |
| Weighted avg | 0.95 | 0.95 | 0.95 | 989 |

**Table 8.** Classification Report of the Model for the Spain Dataset

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Negative | 0.91 | 0.97 | 0.94 | 474 |
| Neutral | 0.93 | 0.89 | 0.91 | 297 |
| Positive | 0.96 | 0.92 | 0.94 | 354 |
| Accuracy | | | 0.93 | 1125 |
| Macro avg | 0.93 | 0.93 | 0.93 | 1125 |
| Weighted avg | 0.93 | 0.93 | 0.93 | 1125 |

**Table 9.** Classification Report of the Model for the Peru Dataset

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Negative | 0.69 | 0.92 | 0.79 | 522 |
| Neutral | 0.91 | 0.87 | 0.89 | 228 |
| Positive | 0.87 | 0.67 | 0.76 | 216 |
| Accuracy | | | 0.84 | 966 |
| Macro avg | 0.83 | 0.82 | 0.81 | 966 |
| Weighted avg | 0.85 | 0.84 | 0.84 | 966 |

### 3.2 Training – Validation

Before implementing the BERT model, a custom tokenization function was defined. The encode_plus method of the BERT tokenizer was called and applied to the training, validation, and test datasets.

Subsequently, the BERT model was imported from the pre-trained library provided by Hugging Face.

A custom function was created to encapsulate the pre-trained BERT model, to which an output layer with three neurons was attached. This layer is necessary to classify the three different classes in the dataset (the three sentiment polarities). The parameters used for the model were as follows:

After fine-tuning the model, the following results were obtained for each corpus separately:

As shown in Tables 1, 2, 3, 4, and 5, three classes are imbalanced. The precision percentages were satisfactory for this experiment; however, performance decreased for the minority classes. To address this issue, the classes in the training set were balanced using an oversampling technique called Random Over Sampler, to reduce bias toward the majority classes. This technique works by increasing the number of samples in the minority classes through random duplication of existing examples until the dataset becomes balanced. This approach helps machine learning models better learn the characteristics of minority classes, thereby improving their ability to make accurate predictions.

## 4 Analysis of Results with Oversampling in the BERT Model

The overall accuracy achieved was 96.66% for the Mexico dataset, 96.66% for Spain, 95.66% for Peru, 59.33% for Costa Rica, and 89.33% for Uruguay, suggesting robust performance in the sentiment classification task for some of the datasets.

### 4.1 Comparison of Evaluation Metrics for the Mexico Dataset

Average Precision Obtained is **97%.**

**Table 10.** Classification Report of the Model for the Costa Rica Dataset

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Negative | 0.93 | 0.94 | 0.94 | 310 |
| Neutral | 0.96 | 0.91 | 0.94 | 246 |
| Positive | 0.92 | 0.94 | 0.93 | 221 |
| Accuracy | | | 0.93 | 777 |
| Macro avg | 0.93 | 0.93 | 0.93 | 777 |
| Weighted avg | 0.93 | 0.93 | 0.93 | 777 |

**Table 11.** Classification Report of the Model for the Uruguay Dataset

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Negative | 0.73 | 0.90 | 0.81 | 367 |
| Neutral | 0.89 | 0.45 | 0.59 | 286 |
| Positive | 0.80 | 0.95 | 0.87 | 290 |
| Accuracy | | | 0.78 | 938 |
| Macro avg | 0.81 | 0.77 | 0.76 | 938 |
| Weighted avg | 0.80 | 0.78 | 0.76 | 938 |

**Table 12.** Classification Report of the Model for the Mexico Dataset with Oversampling

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Negative | 0.98 | 0.97 | 0.97 | 505 |
| Neutral | 0.95 | 0.98 | 0.97 | 505 |
| Positive | 0.97 | 0.98 | 0.97 | 505 |
| Accuracy | | | 0.97 | 1515 |
| Macro avg | **0.97** | 0.97 | 0.97 | 1515 |
| Weighted avg | 0.97 | 0.97 | 0.97 | 1515 |

In general terms, the BERT model trained with oversampling achieved a higher overall accuracy (96.66%) compared to the model trained without oversampling, which reached an accuracy of 93%.

Although both models exhibit high performance, a detailed class-wise analysis reveals important differences in how each version handles the imbalance present in the dataset.

**Negative class:**

Without oversampling, the negative class achieved a precision of 0.98 and a recall of 0.93, indicating that the model correctly identifies most negative tweets, although it still exhibits some confusion during prediction.

With oversampling, both precision and recall slightly increase to values between 0.98 and 0.97, demonstrating a more stable balance and a reduction in classification errors.

**Neutral class:**

In the model without oversampling, the neutral class showed a lower precision (0.86) compared to the other classes, reflecting a higher rate of false positives. Although recall is high (0.95), the disparity between these metrics suggests that the model tends to confuse this category with others.

With oversampling, both precision and recall increase considerably to values between 0.95 and 0.98, indicating that the model can better learn the patterns associated with this class and significantly reduce confusion.

**Positive class:**

The model without oversampling presents high values (precision of 0.95 and recall of 0.96); however, there is still slight variability compared to the negative class due to the smaller number of examples.

With oversampling, both metrics become more uniform, ranging between 0.97 and 0.98, demonstrating that class balancing contributes to a more stable and homogeneous classification across all categories.

These differences suggest that oversampling helps improve the balance between recall and precision for minority classes, particularly in the neutral category.

### 4.2 Comparison of Evaluation Metrics for the Spain Dataset

Average Precision Obtained is 96.66%.

**Negative class:**

With oversampling: precision 0.96, recall 0.96, F1-score 0.96.

**Table 13.** Classification Report of the Model for the Spain Dataset with Oversampling

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Negative | 0.96 | 0.96 | 0.96 | 471 |
| Neutral | 0.94 | 0.95 | 0.95 | 471 |
| Positive | 0.96 | 0.95 | 0.95 | 471 |
| Accuracy | | | 0.95 | 1413 |
| Macro avg | 0.95 | 0.95 | 0.95 | 1413 |
| Weighted avg | 0.95 | 0.95 | 0.95 | 1413 |

**Table 14.** Classification Report of the Model for the Peru Dataset with Oversampling

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Negative | 0.96 | 0.98 | 0.97 | 520 |
| Neutral | 0.99 | 0.95 | 0.97 | 520 |
| Positive | 0.92 | 0.99 | 0.95 | 520 |
| Accuracy | | | 0.95 | 1560 |
| Macro avg | 0.95 | 0.95 | 0.95 | 1560 |
| Weighted avg | 0.95 | 0.95 | 0.95 | 1560 |

**Table 15.** Classification Report of the Model for the Costa Rica Dataset with Oversampling

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Negative | 0.93 | 0.94 | 0.94 | 309 |
| Neutral | 0.96 | 0.91 | 0.94 | 309 |
| Positive | 0.92 | 0.94 | 0.93 | 309 |
| Accuracy | | | 0.93 | 772 |
| Macro avg | 0.93 | 0.93 | 0.93 | 760 |
| Weighted avg | 0.93 | 0.93 | 0.93 | 760 |

Without oversampling: precision 0.91, recall 0.97, F1-score 0.94.

It can be observed that, although the model identifies negative examples in both cases, precision improves substantially when oversampling is applied, reducing the number of false positives.

**Neutral class:**

With oversampling: precision 0.94, recall 0.95, F1-score 0.94.

Without oversampling: precision 0.93, recall 0.89, F1-score 0.94.

The neutral class benefits from oversampling, as it achieves a solid balance between precision and recall. Without oversampling, the model tends to confuse neutral tweets with other categories, decreasing classification accuracy.

**Positive class:**

With oversampling: precision 0.96, recall 0.95, F1-score 0.96.

Without oversampling: precision 0.96, recall 0.92, F1-score 0.94.

For this class, the use of oversampling results in a more robust balance across evaluation metrics.

### 4.3 Comparison of Evaluation Metrics for the Peru Dataset

Average Precision Obtained is 95.66%.

In this experiment, a notable increase in precision can be observed. With oversampling, the model achieved an overall accuracy of 95.66%, with balanced values across all evaluation metrics (precision, recall, and F1-score).

**Negative class:**

With oversampling: precision 0.96, recall 0.98, F1-score 0.97.

Without oversampling: precision 0.69, recall 0.92, F1-score 0.79.

As this class is one of the minority classes, its performance decreases significantly without oversampling. Recall improves considerably when oversampling is applied, indicating that the model is able to capture a higher proportion of negative examples without sacrificing precision.

**Neutral class:**

With oversampling: precision 0.99, recall 0.95, F1-score 0.97.

Without oversampling: precision 0.91, recall 0.87, F1-score 0.89.

**Table 16.** Classification Report of the Model for the Uruguay Dataset with Oversampling

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Negative | 0.94 | 0.90 | 0.92 | 364 |
| Neutral | 0.94 | 0.79 | 0.86 | 364 |
| Positive | 0.80 | 0.98 | 0.88 | 364 |
| Accuracy | | | 0.89 | 1092 |
| Macro avg | 0.89 | 0.89 | 0.89 | 1092 |
| Weighted avg | 0.90 | 0.89 | 0.89 | 1092 |

In this case, the model trained with oversampling retrieves more examples, and its precision increases notably, achieving a more stable balance between precision and recall.

**Positive class:**

With oversampling: precision 0.92, recall 0.99, F1-score 0.86.

Without oversampling: precision 0.87, recall 0.67, F1-score 0.76.

The positive class shows a considerable improvement in both precision and recall, indicating that a higher proportion of positive examples is correctly classified when oversampling is applied.

### 4.4 Comparison of Evaluation Metrics for the Costa Rica Dataset

Average Precision Obtained is 93.66%.

For this case, a notable increase in overall accuracy can be observed. With oversampling, the model achieved an overall accuracy of 93%, with balanced values across all evaluation metrics.

**Negative class:**

With oversampling: precision 0.93, recall 0.94, F1-score 0.94.

Without oversampling: precision 0.71, recall 0.31, F1-score 0.43.

As this is the majority class, its performance was the highest without oversampling; however, it exhibits deficiencies in recall, as less than half of the instances are correctly classified. Recall improves considerably with oversampling, indicating that the model can capture a larger proportion of negative examples without sacrificing precision.

**Neutral class:**

With oversampling: precision 0.96, recall 0.91, F1-score 0.94.

Without oversampling: precision 0.48, recall 0.75, F1-score 0.59.

In this case, the model trained with oversampling retrieves more examples, showing a substantial increase in both precision and recall.

**Positive class:**

With oversampling: precision 0.92, recall 0.94, F1-score 0.93.

Without oversampling: precision 0.59, recall 0.69, F1-score 0.63.

For this minority class, performance was low without oversampling; however, all evaluation metrics increased notably after applying class balancing.

### 4.5 Comparison of Evaluation Metrics for the Uruguay Dataset

Average Precision Obtained is 89.33%.

In this experiment, the BERT model achieved a similar level of overall accuracy in both scenarios (62%), although with substantial differences in how it behaves across classes. This indicates that, while oversampling affects the distribution of correct and incorrect predictions, it does not always guarantee an improvement in the overall accuracy metric.

**Negative class:**

With oversampling: very high precision (0.89) but a recall of only 0.09, indicating that the model predicts negative cases only in very specific instances, leaving the majority unclassified.

Without oversampling: precision of 0.81 and recall of 0.31, with an F1-score of 0.45. Although precision is slightly lower, recall improves, meaning that the model identifies a larger number of negative instances in this scenario.

**Neutral class:**

With oversampling: precision 0.45, recall 0.91, and F1-score 0.60, showing that the model

retrieves almost all neutral examples, albeit at the cost of a high number of false positives.

Without oversampling: precision 0.62 and recall 0.30, with an F1-score of 0.41. In this case, the opposite effect is observed: higher precision but lower coverage, highlighting that oversampling enhances the retrieval capability for this class.

**Positive class:**

With oversampling: precision 0.52, recall 0.59, and F1-score 0.55, reflecting a modest balance.

Without oversampling: precision 0.43 and very high recall (0.98), with an F1-score of 0.59. In this scenario, the model without oversampling classifies almost all positive instances, albeit with a large number of false positives.

## 5 Analysis of Results with Oversampling Using Synthetic Data in the Mexico Dataset

Subsequently, an additional experiment was conducted focusing only on the best-performing result obtained for Mexico (97%) using the Random Over Sampler oversampling technique. In this experiment, an alternative approach was implemented to address class imbalance in the dataset by using synthetic data generated with ChatGPT. Unlike classical oversampling, which consists of duplicating existing instances, this method incorporates newly generated artificial samples, enriching the representation of minority classes and providing greater diversity to the training set.

### 5.1 Overall Performance

The BERT model achieved an overall accuracy of 87%, with macro and weighted average values of 0.87 for precision, 0.86 for recall, and 0.86 for F1-score, demonstrating a highly balanced performance across the different classes. These results are strong, although they do not surpass those obtained in previous experiments using traditional oversampling techniques.

**Negative class:**

Precision: 0.75, Recall: 0.93, F1-score: 0.83

The model exhibits excellent recall, successfully identifying most negative examples, although with slightly lower precision due to the presence of false positives.

**Table 17.** Classification Report of the Model with Oversampling Using Synthetic Data on the Mexico Dataset

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Negative | 0.75 | 0.93 | 0.83 | 494 |
| Neutral | 0.95 | 0.86 | 0.90 | 590 |
| Positive | 0.91 | 0.79 | 0.84 | 456 |
| Accuracy | | | 0.86 | 1540 |
| Macro avg | 0.87 | 0.86 | 0.86 | 1540 |
| Weighted avg | 0.87 | 0.86 | 0.86 | 1540 |

**Neutral class:**

Precision: 0.95, Recall: 0.86, F1-score: 0.90

This class shows the best balance among evaluation metrics, indicating that synthetic data generation significantly improved the model's ability to recognize neutral examples, which were previously difficult to identify.

**Positive class:**

Precision: 0.91, Recall: 0.79, F1-score: 0.84

Although recall is slightly lower than in other classes, the model maintains robust performance, with high precision reducing the number of false positives in this category.

### 5.2 Impact on Model Generalization

The use of oversampling improves the model's ability to correctly identify all classes without biasing predictions toward the majority class. However, it may also increase the risk of overfitting, as the model learns from an artificially expanded dataset, which could affect its performance on unseen data.

To determine whether the model trained with oversampling offers better generalization, it would be useful to conduct additional experiments using an external validation set or data collected during a different time.

## 6 Conclusions and Future Work

The experiments conducted with the BERT model on the TASS 2019 corpus highlight the importance of data balancing strategies in sentiment analysis tasks on social media. Across the different approaches evaluated, clear trends were observed that allow several relevant conclusions to be drawn.

### 6.1 Impact of Data Imbalance

The results obtained without applying balancing techniques showed significantly lower performance, with drastic drops in overall accuracy metrics and the F1 score.

In this scenario, BERT showed a bias toward the majority classes, even almost completely ignoring the minority classes, reflecting the model's sensitivity to imbalance.

### 6.2 Classical Oversampling

The most promising approach was the use of traditional oversampling techniques, which notably improved the model's overall performance, achieving accuracy between 89% and 97%, depending on the data subset used. This method achieved an overall accuracy of 97%, with balanced metrics across all classes and clearly superior performance compared to the previous methods.

### 6.3 Synthetic Data Generation with ChatGPT

This method achieved an overall accuracy of 87%, with balanced metrics across all classes and acceptable performance.

The explanation for this result lies in the fact that, unlike classical oversampling, synthetic generation introduces a great deal of linguistic and semantic variability, incorporating new texts into the corpus.

### 6.4 Key Findings

Class imbalance is a determining factor in sentiment analysis tasks: without balancing strategies, even advanced models such as BERT may fail to correctly classify minority categories.

While traditional oversampling is an effective solution in some cases, the incorporation of synthetic data generation techniques based on large language models represents a more robust and innovative alternative.

These findings open a research line focused on combining deep learning with generative data augmentation, which could be generalized to other natural language processing (NLP) tasks beyond sentiment analysis.

### 6.5 Future Work

Based on the results obtained in this study, several research directions are identified for future work aimed at further advancing sentiment analysis in social media using advanced language models:

#### 6.5.1 Exploration of Advanced Synthetic Data Generation Techniques

Although example generation using ChatGPT showed a positive impact on reducing class imbalance, it would be relevant to evaluate other data augmentation approaches, such as back-translation, automatic paraphrasing, embedding-based lexical substitution, or adversarial data augmentation techniques.

Additionally, the effectiveness of hybrid methods combining classical oversampling with synthetic generation should be analyzed to balance diversity and control within the corpus.

#### 6.5.2 Application for Class-Weighted Learning Techniques

A complementary alternative to oversampling is to adjust the loss function using class-specific weights, allowing the model to penalize errors in minority categories more heavily. This approach could improve the recognition of underrepresented sentiments without directly modifying the dataset.

#### 6.5.3 Advanced Semantic and Functional Classification

Methods will be adopted to address complex semantic and linguistic challenges, following the recommendations of Patra et al. [18] and Kolesnikova [19].

Lexical Function Detection (LF): In the context of lexical functions, future work will focus on incorporating syntactic or dependency-based features alongside contextual embeddings to enhance the model's ability to capture subtle semantic–syntactic interactions in collocations [19].

Hybrid Approaches: Hybrid approaches combining transformer-based models with rule-based methods or knowledge graphs will be explored to leverage the strengths of both symbolic and statistical paradigms in lexical function detection [19].

Emotion Classification: An algorithm will be proposed to automatically classify tweets with higher accuracy for use in data preprocessing systems. Additionally, tweets containing multiple emotions will be investigated, as the work by Patra et al. [18] considered only tweets expressing a single emotion.

Stylistic Classification: The development of classification models for different genres and text domains will also be explored [20].

#### 6.5.4 Evaluation with More Recent Language Models

Although BERT demonstrated solid performance, more recent architectures such as RoBERTa, XLM-RoBERTa, DeBERTa, or instruction-based generative models (GPT-4, LLaMA, Falcon, among others) may provide additional improvements in both accuracy and generalization.

A systematic comparison of these models on the TASS 2019 corpus (and other Spanish-language datasets) would represent a valuable contribution to the field of NLP for the Spanish language.

Use of Distilled Versions and Transformer Alternatives:

To reduce training time and improve efficiency, future work will explore the use of distilled versions of BERT (such as DistilBERT) [20]. Additionally, alternative transformer architectures such as RoBERTa and GPT will be investigated to assess their impact on text classification quality [20].

Fine-Tuning Strategies: Advanced fine-tuning strategies will be implemented to stabilize training and optimize performance. These include learning rate scheduling techniques to mitigate fluctuations in the loss function during training [20].

#### 6.5.5 Adaptation to Dynamic Social Media Contexts

Given that language on platforms such as Twitter evolves constantly (new slang, sarcasm, emoji usage, irony, etc.), it would be relevant to explore continual learning approaches that allow models to be updated without forgetting previously acquired knowledge.

#### 6.5.6 Transfer to Other Domains and Multilingual Corpora

Finally, an important research direction involves evaluating the transfer of this approach to other domains (product reviews, forum comments, service reviews) and to multilingual corpora, to validate the robustness of the method in broader and comparative scenarios.

## References

1. **Reducindo, J., Calero, H., Fernández, C., Ramos, E. (2024).** Análisis de sentimientos utilizando ChatGPT: Una revisión sistemática de la literatura. Revista Científica Emprendimiento Científico Tecnológico. doi: 10.54798/RWFA3855.
2. **Miranda, C.H., Sanchez-Torres, G., Salcedo, D. (2023).** Exploring the Evolution of Sentiment in Spanish Pandemic Tweets: A Data Analysis Based on a Fine-Tuned BERT Architecture. Data, Vol. 8, No. 6, pp. 96. doi: 10.3390/data8060096.
3. **Liu, B. (2022).** Sentiment Analysis and Opinion Mining. 1st ed. Springer Nature.
4. **Caton, S., Hall, M., Weinhardt, C. (2015).** How Do Politicians Use Facebook? An Applied Social Observatory. Big Data & Society, Vol. 2, No. 2. doi: 10.1177/2053951715612822.
5. **Prata, D.N., Soares, K.P., Silva, M.A., Trevisan, D.Q., Letouze, P. (2016).** Social Data Analysis of Brazilian's Mood from Twitter. International Journal of Social Science and Humanity, Vol. 6, No. 3, pp. 179–183. doi: 10.7763/IJSSH.2016.V6.640.

6. **Cervantes, I., Pastor, C., Fernández, D. (2021).** El español: una lengua viva. Instituto Cervantes.
7. **Gomez-Adorno, H., Bel-Enguix, G., Sierra, G., Barajas, J.C., Álvarez, W. (2024).** Machine Learning and Deep Learning Sentiment Analysis Models: Case Study on the SENT-COVID Corpus of Tweets in Mexican Spanish. Informatics, Vol. 11, No. 2, pp. 24. doi: 10.3390/informatics11020024.
8. **Barrios González, E., Tovar Vidal, M., Reyes-Ortiz, J.A., Zacarias Flores, F., Bello López, P. (2023).** Exploring BERT-Based Pretrained Models for Polarity Analysis of Tweets in Spanish. International Journal of Combinatorial Optimization Problems and Informatics, Vol. 14, No. 1, pp. 27–38. doi: 10.61467/2007.1558.2023.v14i1.336.
9. **Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018).** BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.
10. **Zeng, L. (2024).** Leveraging Large Language Models for Code-Mixed Data Augmentation in Sentiment Analysis. arXiv preprint arXiv:2411.00691. doi: 10.48550/arXiv.2411.00691.
11. **Bordoloi, M., Biswas, S.K. (2023).** Sentiment Analysis: A Survey on Design Framework, Applications and Future Scopes. Artificial Intelligence Review, Vol. 56, No. 11, pp. 12505–12560. doi: 10.1007/s10462-023-10442-2.
12. **Molina-González, M.D., Martínez-Cámara, E., Martín-Valdivia, M.T., Perea-Ortega, J.M. (2013).** Semantic Orientation for Polarity Classification in Spanish Reviews. Expert Systems with Applications, Vol. 40, No. 18, pp. 7250–7257. doi: 10.1016/J.ESWA.2013.06.076.
13. **Colas, F., Brazdil, P. (2006).** Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. In Artificial Intelligence in Theory and Practice, Springer, Vol. 217, pp. 169–178. doi: 10.1007/978-0-387-34747-9_18.
14. **Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., Pérez, J. (2023).** Spanish Pre-trained BERT Model and Evaluation Data. arXiv preprint arXiv:2308.02976.
15. **Elmitwalli, S., Mehegan, J. (2024).** Sentiment Analysis of COP9-Related Tweets: A Comparative Study of Pre-Trained Models and Traditional Techniques. Frontiers in Big Data, Vol. 7. doi: 10.3389/fdata.2024.1357926.
16. **Roumeliotis, K.I., Tselikas, N.D., Nasiopoulos, D.K. (2024).** LLMs and NLP Models in Cryptocurrency Sentiment Analysis: A Comparative Classification Study. Big Data and Cognitive Computing, Vol. 8, No. 6, pp. 63. doi: 10.3390/bdcc8060063.
17. **Suhaeni, C., Yong, H.S. (2023).** Mitigating Class Imbalance in Sentiment Analysis Through GPT-3-Generated Synthetic Sentences. Applied Sciences, Vol. 13, No. 17, pp. 9766. doi: 10.3390/app13179766.
18. **Patra, S., Satpathy, R.N., Panigrahi, C.R., Nanda, S. (2025).** Human Emotion and Sentiment Analysis Using Machine Learning. Computación y Sistemas, Vol. 29, No. 2, pp. 765–772. doi: 10.13053/CyS-29-2-5041.
19. **Kolesnikova, O. (2025).** Lexical Function Detection in Spanish Collocations Using Transformer Architecture. Computación y Sistemas, Vol. 29, No. 2, pp. 847–856. doi: 10.13053/CyS-29-2-5620.
20. **Solovyev, V.D., Ten, A.M., Solnyshkina, M.I., Andreeva, M.I. (2025).** Bert for Classification of Russian Functional Styles. Computación y Sistemas, Vol. 29, No. 2, pp. 633–641. doi: 10.13053/CyS-29-2-5712.

*Article received on 20/02/2026; accepted on 09/04/2026.*
**Corresponding authors is Yulia Ledeneva.*