ISSN 1870-4069

Application of Condorcet's Jury Theorem for Enhancing Sentiment Analysis Performance Using BERT Transformers: A Case Study for Spanish

Gerardo Bárcena-Ruiz[1,2], Richard de Jesús Gil-Herrera[3]

[1] Universidad Americana de Europa,
Mexico

[2] Universidad Panamericana,
Mexico

[3] Universidad Internacional de la Rioja,
Spain

gbarcena@up.edu.mx, richard.dejesus@unir.net

Abstract. This paper examines the application of Condorcet's Jury Theorem (CJT) in the context of sentiment analysis using BERT models for Spanish language. While there are many BERT model variants, including vanilla BERT, DeBERTa, ALBERT or Longformer, this study focuses on BERT, RoBERTa and DistilBERT due to their superior accuracy and efficient retraining times. The objective of this research is to assess whether CJT can enhance sentiment analysis performance with BERT models. The experiments conducted explore various scenarios to evaluate the model's behavior and the effectiveness of a jury metamodel. The CJT approach can yield superior results, achieving an F1-Score of 0.994 compared to a single model's average F1-Score of 0.974, according to this study. Additionally, the study highlights the critical role of dataset language quality in training more effective models.

Keywords: Condorcet's Jury Theorem (CJT), BERT, transformer, performance, sentiment analysis, Spanish language.

1 Introduction

According to Cervantes Institute [1], Spanish is the second most spoken native language by number of speakers, following Mandarin Chinese. Additionally, in a worldwide count of total speakers (including native speakers, limited proficiency speakers and students of Spanish), it ranks forth, after English, Mandarin Chinese, and Hindi.

However, research on AI involving the Spanish language occupies a relatively small market share. For instance, studies on sentiment analysis in Spanish account for only 3.11 % to 5.26% of total research in this area [2]. Natural Language Processing (NLP) encompasses a range of critical tasks heavily influenced by the target language. Core examples include Automatic Translation, Text Summarization, and Text Generation,

pp. 5–14; rec. 2024-06-04; acc. 2024-08-17

all of which necessitate effective text comprehension for optimal performance [18]. Furthermore, NLP excels in specific applications such as topic modeling (extracting semantic information), news classification (leveraging news as a rich data source), sentiment analysis (categorizing text based on sentiment), and question answering (a challenging task with potential for intuitive knowledge acquisition) [19].

In recent years, Deep Learning has emerged as the dominant paradigm for text classification [19]. The evolution of text processing methodologies progressed from Recurrent Neural Networks (RNNs) to Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, which mitigated the vanishing gradient problem through more efficient backpropagation. However, these models faced limitations in parallelization and handling long sequences. The subsequent introduction of Transformers, equipped with the attention mechanism, revolutionized the field by addressing these shortcomings and enabling highly parallelizable text processing [18].

Transformers [20] are sequence-to-sequence models comprising encoder and decoder blocks, commonly employed in neural machine translation [21]. Bidirectional and Auto-Regressive Transformers (BART) and Text-to-Text Transfer Transformers (T5) exemplify this architecture. Disabling the encoder yields a sequence generation or language model, such as the renowned Generative Pretrained Transformer (GPT). Conversely, deactivating the decoder results in encoder-only or automatic coding Transformers, which generate input sequence representations. BERT is a prominent example of this latter category.

Given the potential of advanced text processing for the Spanish language, this study focuses on sentiment analysis utilizing BERT transformers. Based on [2], DistilBERT, BERT (vanilla), and RoBERTa were selected for their high F1-scores and efficient retraining times (F50 T50) among top-performing BERT models.

However, is it possible to enhance performance using the same models without altering their internal architecture? The principle of democracy might provide an answer to this question. If multiple entities (models) cast vote in a particular manner, the final verdict could represent the optimal value achievable. However, what is the probability to obtain correct answers? In the “Materials and methods” section, we discuss Condorcet’s Jury Theorem (CJT) and its associated probabilities, bet we can express in advance that the CJT framework employs a majority voting mechanism for jury decisions, anticipated to outperform individual judge assessments due to the collective expertise of participants in this domain.

In this study, we apply CJT as follows: a) each BERT model independently classifies every text as either positive or negative, functioning as individual judges. b) The experiment utilizes different BERT models that have an F1-Score greater than 0.5, ensuring that each model has sufficient knowledge of the dataset’s subject matter to classify the text accurately.

As a result, when BERT models perform the sentiment analysis individually, they vote on whether each text is positive or negative. According to CJT, the collective accuracy of these models, acting as jury, could potentially be higher.

1.1 General Objective

General objective is to apply the Condorcet's Jury Theorem (CJT) for enhancing the performance of different BERT transformers, which have been pretrained on from different datasets.

1.2 Specific Objectives

Subsequently, the study aims to achieve several specific objectives: i) identifying three distinct BERT models for sentiment analysis tasks in Spanish. ii) Determining the availability of datasets in Spanish suitable for training, trying to get balanced subsets containing both positive and negative sentiments. iii) Establishing a comparison framework for evaluating performance of individual models versus the application of CJT.

2 Related Works

The objective of the authors in [3] and [6] was to compare FinBERT and FinDROBERTA with GPT-4 for text classification in the financial domain, using a specially developed market-based dataset for retraining the models.

Their primary source of financial information was Bloomberg Market Wraps (BMW, spanning from 2010 to 2024). The BMW repository is a daily-consolidated summary of financial news by human journalists. The dataset examples were organized as follows: 27% neutral, 31% negative, and 42% positive headlines classifications.

The accuracies (F-score) for each model were, before retraining, as follows: 0.47 for GPT-4 and 0.44 for DistilROBERTA and FinBERT. After supervised fine-tuning (SFT), the accuracies improved to 0.51 for SFT GPT-4, 0.49 for SFT DistilROBERTA, and 0.50 for SFT FinBERT. The authors organized the models into different arrays or bags as follows:

- Bagging 1: SFT GPT + SFT DistilROBERTA + SFT FinBERT.
- Bagging 2: SFT DistilROBERTA + SFT FinBERT.
- Bagging 3: All models, SFT + No SFT.

Each bag achieved the following F-scores: 0.51 for Bagging 1, 0.52 for Bagging 2 and Bagging 3. Some important conclusions highlighted by the authors include: the bag array facilitates the learning of more sophisticated patterns; it mitigates the bias introduced by humans; it enhances the objectivity of the model training process; and the jury votes outperformed the original SFT models, although not sufficient to validate the application of CJT. An increase in model parameters does not necessarily translate to performance improvements.

The objective of study in [7] is to classify cancer types using a dense artificial neural network. To perform the classifying task, the authors used 25,000 examples from Lung and Colon Cancer Histopathological Image Dataset (LC25000), organized into five balances classes: benign lung tissue, lung adenocarcinomas, lung squamous cell carcinomas, benign colon tissue, and colon adenocarcinomas.

In the proposed algorithm, classifiers were organized into a majority voting system based on CJT as follows: after training, the best performing model for each class was saved, and the output was expressed in an array that contains votes for each class. The classification accuracy for the five classes reached 99.88% with the CJT ensemble model.

Finally, the research team in [8] attempted to classify COVID-19 using Deep Neural Networks (DNN) models and CJT. The dataset used was the Curated Dataset for COVID-19 Posterior–Anterior Chest Radiography Images (X-rays). After data cleaning, the final dataset comprised the following classes' mixture: 18.15% normal, 16.95% COVID-19, 33.07% viral pneumonia, and 31.82% bacterial pneumonia.

During experimentation, the testing accuracy values for predefined models were as follows: InceptionV3 96.45%, InceptionResNetV2 97.42%, ResNet50V2 97.90%, DenseNet121 97.75%, DenseNet201 97.26%. The accuracy values for the proposed models were the following: DETL Ensemble Model 97.26%, Jury Ensemble Model 98.22%. The results show that the CJT-based model outperformed the individual models.

While previous study has not directly employed the tools focused on in this paper, the authors demonstrate the applicability of the CJT framework to classification tasks.

3 Materials and Methods

In the following sections, we provide an explanation of the different models and datasets that were utilized to perform the experimentation.

3.1 Models

In this study, we utilized three pretrained transformer models from the Hugging Face hub: a) BERT [9], b) RoBERTa [10] and c) DistilBERT [11]. All three models have demonstrated strong performance for tasks involving the Spanish language, as evidenced by previous research [2].

For retraining purpose, we utilized Google Colaboratory Pro with T4 GPU [15], which fits to our computing needs. The parameters included a maximum length of 2048 for the tokenizer´s encoder, a batch size of thirty-two (32), and ten (10) epochs, according to which is reported in [16] this number of cycles produces a no overfitting model.

3.2 Datasets

Three datasets were employed for this experiment. The first dataset is about movie reviews expressed on the IMDB web site [12]. The remaining two datasets focus on tourism-related comments and were retrieved from Hugging Face hub: the Sepidmnorozy [13] and Alexcom [14] datasets. The IMDB and Alexcom datasets provided a balanced set of examples, equally divided between positive and negative sentiment classes. The Sepidmnorozy dataset, however, contained a slightly smaller sample size of 512 examples, maintaining a balanced distribution of positive and negative sentiment (256 each).

Table 1. Datasets, tokens characteristics.

Dataset	Subject	Min	Max	Avg.	Std. Dev.
IMDB	Movies	17	1251	238.319	179.927
Alexcom	Tourism	6	3153	82.531	123.599
Sepidmnorozy	Tourism	1	120	17.259	16.329

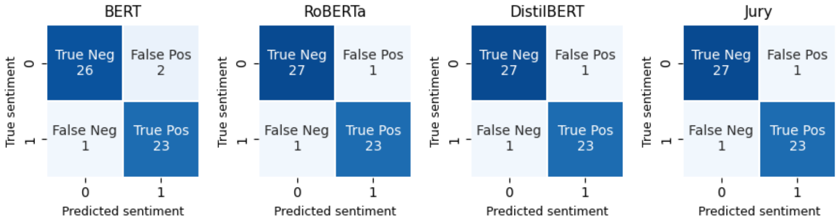

Fig. 1. Average confusion matrices for test #1.

Table 1 summarizes the tokens statistics for each dataset. As evident, the IMBD dataset exhibits a higher token count compared to the other two datasets.

3.3 Condorcet's Jury Theorem

Condorcet's Jury Theorem (CJT) [4] offers a theoretical framework [17] to address the issue of accuracy in collective decision-making tasks such as sentiment analysis. In the context of CJT, each participant can be conceptualized as a classifier, tasked with distinguishing positive from negative sentiment in text data. Each classifier must be independent, well-trained, and uniformly biased towards the correct alternative according to the following conditions:

- Independence: Each classifier must make its predictions independently.
- Identical Distribution: The performance of all classifiers should be statistically similar, implying they have the same underlying distribution of accuracy.
- Better than random: Each classifier's performance must exceed random guessing because they are well-trained.
- Uniform distribution among incorrect alternatives: All classifiers should exhibit the same probability, which is lower than random guessing, when they choose the incorrect alternative [3].

Building upon the CJT framework, we can extend its principles to the realm of ensemble classifiers. Here, CJT can be applied to a collection of classifiers, denoted as C_1, C_2, …, C_n, that satisfy the previously outlined conditions. These classifiers can be envisioned as casting votes (positive or negative sentiment) into a C_{bag}. Notably, under the CJT framework, the collective decision reached by the bag (through a majority vote) is predicted to exhibit higher accuracy compared to the individual decisions of any single classifier within the ensemble.

Furthermore, as the number of classifiers (n) approaches infinity, the probability of the C_{bag} converging on the correct sentiment classification also increases. Conversely, if the individual classifiers perform worse than random guessing, the collective decision

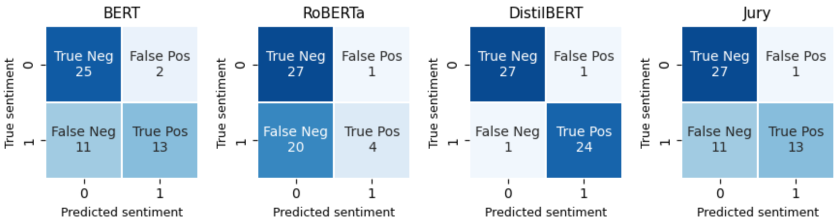

Fig. 2. Average confusion matrices for test #2.

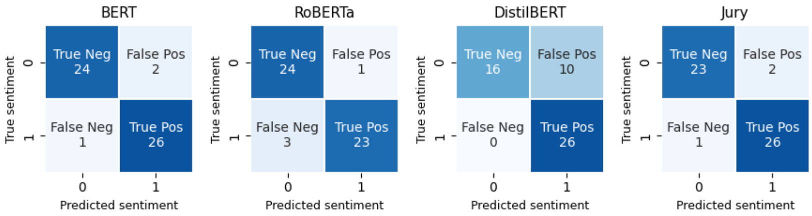

Fig. 3. Average confusion matrices for test #3.

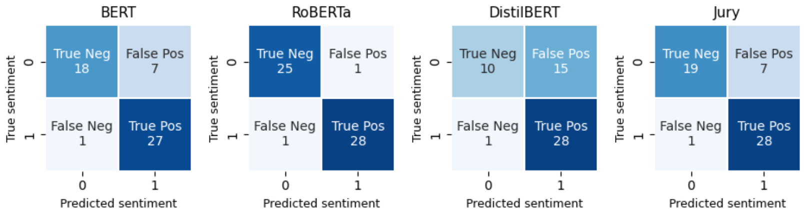

Fig. 4. Average confusion matrices for test #4.

of the C_{bag} is more likely to converge on the incorrect sentiment [3, 5]. The proof for this theorem is given in [4] and [17].

3.4 Experiment

Five ensemble models were created by combining three pretrained BERT models with available datasets as follows: BERT-IMDB, BERT-Alexcom, RoBERTa-IMDB, RoBERTa-Sepidmnorozy and DistilBERT-IMDB.

Four separate experiments were conducted, each consisting of ten rounds. During each round, 50 examples were randomly sampled from each dataset for classification (with 52 examples used from the Sepidmnorozy dataset). The summarized results of all rounds are shown in Tables 2 and Fig. 1 to Fig. 4.

4 Experimental Results and Discussion

The first experiment employed the IMDB dataset for both training and testing the ensemble models. Here, the proposed Jury (or CJT) approach achieved excellent performance, surpassing both the simple average of individual model votes and the

Table 2. Experiment statistics.

Testing Dataset		**BERT**	**RoBERTa**	**DistilBERT**	**Avg.**	**Jury**
IMDB	*Training Dataset*	*IMDB*	*IMDB*	*IMDB*		
Test	F1 Avg.	0.950	0.986	0.986	0.974	0.994
#1	F1 StdD.	0.026	0.020	0.020	0.015	0.009
	Times > Avg.	1	8	8		10
IMDB	*Training Dataset*	*Alexcom*	*Sepidmnorozy*	*IMDB*		
Test	F1 Avg.	0.754	0.596	0.990	0.780	0.788
#2	F1 StdD.	0.044	0.065	0.013	0.034	0.063
	Times > Avg.	2	0	10		5
Alexcom	*Training Dataset*	*Alexcom*	*Sepidmnorozy*	*IMDB*		
Test	F1 Avg.	0.974	0.934	0.814	0.907	0.970
#3	F1 StdD.	0.025	0.037	0.066	0.033	0.027
	Times > Avg.	10	6	1		10
Sepidmnorozy	*Training Dataset*	*Alexcom*	*Sepidmnorozy*	*IMDB*		
Test	F1 Avg.	0.862	0.994	0.707	0.854	0.879
#4	F1 StdD.	0.035	0.009	0.056	0.028	0.048
	Times > Avg.	4	10	0		5

average F1-score in 10 out of 10 rounds, according to Table 2 – Test #1. Notably, the confusion matrices exhibited perfect balance, as seen in Fig. 1.

In the second experiment, IMDB was again used for testing, while the training data comprised a combination of BERT-Alexcom, RoBERTa-Sepidmnorozy, and Distil-BERT-IMDB ensembles. While BERT, RoBERTa, and the Jury (CJT) model all exhibited lower performance in this scenario, the Jury (CJT) approach still maintained a higher accuracy than the average vote and achieved a superior F1-score in 5 out of 10 rounds, as seen in Table 2 – Test #2. Interestingly, both BERT and RoBERTa models displayed a significant difficulty in identifying negative sentiment, as evident from the values in the Fig. 2, because of their high rate of False-Negative values.

The third and fourth experiments utilized Alexcom and Sepidmnorozy datasets for testing, respectively. These experiments yielded improved F1-scores. In third test, BERT achieved an F1-score of 0.974 and RoBERTa obtained a score of 0.934. Next, in fourth experiment, BERT obtained an F1-score of 0.862 and RoBERTa achieved a score of 0.994, according to Table 2. Additionally, both models BERT and RoBERTa obtained more balanced confusion matrices, as seen in Fig. 3 and Fig. 4. The Jury (CJT) model consistently achieved superior results compared to the simple average vote across all experiments.

As expected, model performance deteriorated when classifying text from a dataset that differed from the training dataset. This is evident in Test #2, where BERT achieved an F1-score of 0.754 and RoBERTa obtained a score of 0.596, as presented in Table 2. These values are lower compared to the scenario where training and testing occur on the same dataset. However, DistilBERT exhibited a contrasting behavior in Test #3. It attained an F1-score of 0.814, surpassing its performance in the fourth experiment (F1-score of 0.707).

This anomaly can be attributed to the training data used. As shown in Table 1, the IMDB dataset employed for DistilBERT's training purpose, likely contained more informative examples as evidenced by the higher average number of tokens.

5 Conclusion

In conclusion, our findings demonstrate the effectiveness of the Jury method in sentiment analysis tasks. When appropriately finetuned, the Jury approach consistently outperformed individual models. However, this success hinges on the models' exposure to domain-specific vocabulary during fine-tuning. For instance, BERT and RoBERTa exhibited strong performance when the training and testing phases involved tourism data, likely due to their familiarity with sentiment-laden tourism vocabulary. Interestingly, the DistilBERT model, fine-tuned on the movie dataset (IMDB), also achieved acceptable performance on the tourism datasets.

This can potentially be attributed to the richer vocabulary present in the IMDB dataset compared to the tourism datasets (as observed in Table 1), which provided a stronger foundation for sentiment understanding in Spanish language.

Our primary conclusion is that Condorcet's Jury Theorem can be successfully applied to sentiment analysis tasks, particularly when the constituent models possess domain-specific knowledge. This highlights the importance of tailoring models to the specific text domain being classified.

Looking towards future work, incorporating models trained on high-quality language data into the Jury framework represents a promising avenue for further improvement. This approach has the potential to deliver robust sentiment analysis across a wide range of domains.

References

1. Centro Virtual Cervantes: El español en el mundo, anuario del instituto Cervantes 2023. Spain: Instituto Cervantes (2023) https://cvc.cervantes.es/lengua/anuario/anuario_23/informes_ic/p01.htm
2. Bárcena-Ruiz, G., de-Jesús-Gil, R.: BERT transformers performance comparison for sentiment analysis: A case study in spanish. In: Rocha, Á., Adeli, H., Dzemyda, G., Moreira, F., Poniszewska-Marańda, A. (eds) Good Practices and New Perspectives in Information Systems and Technologies, WorldCIST 2024, Lecture Notes in Networks and Systems, Springer, vol. 989, pp. 152–164 (2024) doi: 10.1007/978-3-031-60227-6_13

3. Benhamou, E.: Small triumphs over large: Instances where BERT-based fine-tuned models surpass GPT-4 in classification tasks (2024)
4. Sancho, Á. R.: On the probability of the Condorcet jury theorem or the miracle of aggregation. Mathematical Social Sciences, vol. 119, pp. 41–55 (2022) doi: 10.1016/j.mathsocsci. 2022.06.002
5. Stanford Encyclopedia of Philosophy: Jury theorems. USA: Stanford University (2021) https://plato.stanford.edu/entries/jury-theorems/#CondJuryTheo
6. Lefort, B., Benhamou, E., Ohana, J. J., Guez, B., Saltiel, D., Challet, D.: When small wins big: Classification tasks where compact models outperform original GPT-4. SSRN 4780454 (2024) doi: 10.2139/ssrn.4780454
7. Srivastava, G., Chauhan, A., Pradhan, N.: CJT-DEO: Condorcet's jury theorem and differential evolution optimization based ensemble of deep neural networks for pulmonary and colorectal cancer classification. Applied Soft Computing, vol. 132, p. 109872 (2023) doi: 10.1016/j.asoc.2022.109872
8. Srivastava, G., Pradhan, N., Saini, Y.: Ensemble of deep neural networks based on Condorcet's jury theorem for screening Covid-19 and pneumonia from radiograph images. Computers in Biology and Medicine, vol. 149, p. 105979 (2022) doi: 10.1016/j.compbiomed. 2022.105979
9. Romero, M.: BETO (Spanish BERT) + Spanish SQuAD2.0. Hugging Face (2023) https://huggingface.co/mrm8488/bert-base-spanish-wwm-cased-finetuned-spa-squad2-es
10. De-la-Rosa, J., González, E., Villegas, P., González-de-Prado, P., Romero, M., Grandury, M.: Hugging Face. bertin-project/bertin-roberta-base-spanish. Hugging Face (2022) https://huggingface.co/ bertin- project/ bertin-roberta-base-spanish
11. DCCUChile: Department of computer sciences. University of Chile (2022) dccuchile/distilbert-base-spanish-uncased-finetuned-mldoc. Hugging Face. https://huggingface. co/ dccuchile/ distilbert-base-spanish-uncased-finetuned-mldoc
12. Fernandez, L.: IMDB Dataset of 50K movie reviews (Spanish). Kaggle (2021)
13. Mollanorozy, S. sepidmnorozy/Spanish_sentiment. Hugging Face (2022) https://hugging face. co/ datasets/ sepidmnorozy/Spanish_sentiment/tree/main
14. Cortés-Miranda, I.: Hugging Face. (2023) https://huggingface.co/datasets/alexcom/analisis-sentimientos-textos-turisitcos-mx-polaridad
15. Carneiro, T., Medeiros-Da-Nóbrega, R. V., Nepomuceno, T., Gui-Bin, B., De-Albuquerque, V. H. C., Rebouças-Filho, P. P.: Performance analysis of google colaboratory as a tool for accelerating deep learning applications. IEEE Access, vol. 6, pp. 61677–61685 (2018) doi: 10.1109/ACCESS.2018.2874767
16. Komatsuzaki, A.: One epoch is all you need. (2019) http://arxiv.org/abs/1906.06669
17. Berend, D., Paroush, J.: When is Condorcet's jury theorem valid? Social Choice and Welfare, vol. 15, no. 4, pp. 481–488 (1998) doi: 10.1007/s003550050118
18. Gillioz, A., Casas, J., Mugellini, E., Abou-Khaled, O.: Overview of the transformer-based models for NLP tasks. In: 2020 15th Conference on computer science and information systems (FedCSIS), pp. 179–183 (2020) doi: 10.15439/2020F20
19. Dogra, V., Verma, S., Kavita, Chatterjee, P., Shafi, J., Choi, J., Ijaz, M. F.: A complete process of text classification system using state-of-the-art NLP models. Computational Intelligence and Neuroscience, vol. 2022, no. 1, p. 1883698 (2022) doi: 10.1155/2022/ 1883698
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I.: Attention is all you need. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, pp. 1-11 (2017)

21. Lin, T., Wang, Y., Liu, X., Qiu, X.: A survey of transformers. AI Open, vol. 3, pp. 111–132 (2022) doi: 10.1016/j.aiopen.2022.10.001