Conspiracy Theory Text Classification Based on CT-BERT and BETO Models

Notebook for PAN at CLEF 2024

Jiahong Huang[1], Zhongyuan Han[1,*], Ruihao Zhu[1], Mingcan Guo[1] and Kaiyin Sun[2]

[1]Foshan University, Foshan, China
[2]Foshan Huaying School, Foshan, China

Abstract

Conspiracy theories serve as significant conduits for disseminating false information, leading to the misguidance of the public and impacting the ability to make informed decisions. Identifying such theories poses numerous challenges. For this, PAN 2024 has introduced two tasks: the binary classification task that distinguishes between conspiratorial and critical texts, and the task of detecting elements of oppositional narratives. In this paper, Covid-Twitter-BERT and BETO are fine-tuned to address these challenges. The assessment results indicate that the model achieved MCC scores of 0.8198 and 0.6192 on the English and Spanish test sets for subtask 1, respectively. For subtask 2, span-F1 scores of 0.5886 and 0.4994 were attained on the English and Spanish test sets, respectively.

Keywords

PAN 2024, Conspiracy Theories, Binary Classification, Covid-Twitter-BERT, BETO

1. Introduction

Conspiracy theories [1] disseminate incorrect, unverified, or deliberately misleading information, which misguides the public and affects people's ability to make fact-based judgments and decisions. The automatic identification of conspiracy theories on social networks is crucial for combating false information and maintaining social stability. A significant challenge in detecting conspiracy theories using natural language processing (NLP) models is the complexity of distinguishing between critical analysis and conspiratorial ideation during automated content moderation. The PAN 2024 [2] Oppositional Thinking Analysis [3] comprises two subtasks: distinguishing between critical and conspiracy texts (subtask 1) and detecting elements of the oppositional narratives (subtask 2).

In this study, the tasks of distinguishing between critical and conspiracy texts and detecting elements within oppositional narratives are accomplished by fine-tuning models such as Covid-Twitter-BERT [4] (referred to as CT-BERT) and BETO [5].

2. Method

The contest encompasses two distinct subtasks. Subtask 1 is a binary text classification task, which requires determining whether the text falls into the category of "critical" or "conspiracy." Subtask 2 is a token-level classification task, which involves identifying the goals, effects, agents, and groups-in-conflict within the text. Both tasks include texts in both English and Spanish.

Numerous studies have demonstrated the outstanding performance of Transformer-based models [6] in classification tasks. In this competition, the tasks were primarily accomplished by fine-tuning BERT [7] models and constructing classifiers. To better adapt to the application scenarios of each task, different BERT models were fine-tuned, and classifiers were constructed accordingly. Two models were primarily

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
hinlole@qq.com (J. Huang); hanzhongyuan@gmail.com (Z. Han); brotherblaozhu281@gmail.com (R. Zhu); gmc9812@163.com (M. Guo); sunkaiyin123@163.com (K. Sun)
0009-0004-6192-5274 (J. Huang); 0000-0001-8960-9872 (Z. Han); 0009-0000-7521-946X (R. Zhu); 0000-0002-4977-2138 (M. Guo)

utilized: CT-BERT and BETO. CT-BERT, a Transformer-based natural language processing (NLP) model, has been pre-trained on a large corpus of Twitter messages related to COVID-19, showcasing strong comprehension and analytical capabilities for COVID-19 information on social media. BETO, on the other hand, is a BERT model trained on a large Spanish corpus, enabling it to more accurately understand word meanings, syntactic structures, and semantic relationships when processing Spanish texts.

Data processing was initially conducted for model fine-tuning. The datasets for this task are presented in two languages, English and Spanish. Each dataset consists of 4000 texts with their annotations. Each text is associated with a dictionary containing the ID, tokenized text, the binary category, and the span annotations. [1]Span annotations comprise a list of dictionaries, with each dictionary corresponding to an annotated span, which includes the span's category and text, identified by the start and end characters. The data for each language was partitioned into a training set and a validation set in a 9:1 ratio.

2.1. The method for subtask 1

For subtask 1, the BERT model is fine-tuned using the AutoModelForSequenceClassification as the model architecture to construct a binary classifier. This architecture is designed for sequence classification tasks, capturing the sequential characteristics of text data, thereby enhancing the model's generalization capability. The fully connected layer of the model outputs the probability distribution of the target categories, which is then transformed into the probabilities of each class using the softmax activation function. During fine-tuning, the Adam optimizer and the cross-entropy loss function are employed, with the number of epochs set to 100. The model that performs best on the validation set, according to the evaluation results of each epoch, is selected and saved.

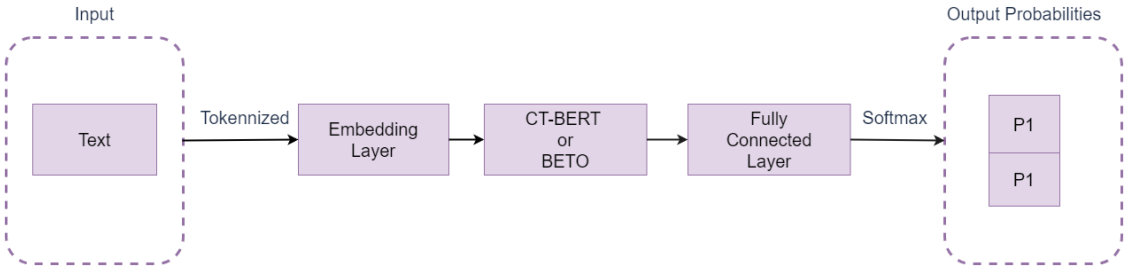

Figure 1: Training Framework of our models for distinguishing between critical and conspiracy texts

The CT-BERT and BETO models were fine-tuned, with CT-BERT being applied to English texts and BETO to Spanish texts, followed by an evaluation of their performance on the validation set. During the experimentation, it was observed that CT-BERT, once trained on a Spanish dataset, also demonstrated effective processing capabilities on the English dataset. This led to the hypothesis that CT-BERT possesses multilingual learning capabilities. Consequently, an attempt was made to merge the English and Spanish training datasets into a single multilingual dataset for training CT-BERT. It was found that, for Task 1-Spanish, the CT-BERT model trained with the multilingual dataset showed specific improvements over the model trained with a single language dataset, with an approximate increase of 0.012 in MCC on the validation set.

Furthermore, the model's performance in Spanish was relatively poor. A multi-model voting approach was also attempted to enhance the model's accuracy in Spanish texts. The models used for voting include:

- BETO
- CT-BERT trained with a Spanish dataset
- CT-BERT trained with a merged bilingual dataset

[1]https://pan.webis.de/clef24/pan24-web/oppositional-thinking-analysis.html#data

It was discovered that post-voting, the model's predictive accuracy on Spanish texts significantly increased.

Table 1 shows the results of various models on validation set.

2.2. The method for subtask 2

For subtask 2, based on the Baseline [8] approach, the BERT models are fine-tuned using the Adam optimizer and the cross-entropy loss function, with the number of epochs set to 15. The Baseline for subtask 2 achieves token-level classification through a BERT-based multi-task token classifier (separate classification heads, common transformer backbone). The CT-BERT and BETO models are fine-tuned individually, with the CT-BERT model being applied to English texts and the BETO model to Spanish texts.

Table 3 shows the results of various models on validation set.

3. Results

3.1. The result of subtask 1

For subtask 1, on task1-English, results were submitted utilizing the CT-BERT model trained with the English dataset. On task1-Spanish, two submissions were made: one from the CT-BERT model trained with the Spanish dataset and another employing a multi-model voting approach. The official evaluation metric for subtask 1 is the MCC [9], Table 2 shows the results of various models for subtask 1.

On task1-English, the model's performance surpassed the baseline, ranking fifth among all participants and outperforming the majority of competing methods. This suggests that the CT-BERT model is able to obtain higher-quality text encoding in this task, demonstrating its superior performance. On task1-Spanish, the multi-model voting method showed an improvement over the single CT-BERT model prediction method, with an approximate increase of 0.03 in the MCC score. This suggests that the multi-model voting approach can enhance model performance to a certain degree. Additionally, a significant discrepancy was observed between the model's performance on the test set and the validation set. This suggests that our model is suffering from overfitting, which may be due to the validation set being too small or the selected samples not being representative enough. To resolve this issue, we can fine-tune the hyperparameters and retrain the model. Moreover, employing k-fold cross-validation can provide a more thorough assessment of the model's performance, thereby reducing the risk of overfitting.

Table 1
Performance of Various Models on Task 1 Validation Set

Model	Training Set	Validation Set	MCC	F1-macro	F1-conspiracy	F1-critical
CT-BERT	English	English	0.8307	0.9149	0.8855	0.9442
CT-BERT	Spanish	Spanish	0.7265	0.8546	0.8074	0.9019
CT-BERT	Spanish	English	0.6676	0.8298	0.7642	0.8953
CT-BERT	Multilingual	Spanish	0.7383	0.8669	0.8293	0.9045
BETO	Spanish	Spanish	0.7382	0.8681	0.8328	0.9034
Multi-model Voting	-	Spanish	0.7960	0.8923	0.8592	0.9254

3.2. The result of subtask 2

For subtask 2, on task2-English, predictions were made using the CT-BERT model, while on task2-Spanish, the BETO model was utilized. The official evaluation metric for subtask 2 is span-F1 [10], and Table 4 shows the results of various models for subtask 2.

On task2-English, the model demonstrated superior performance, achieving an approximate increase of 0.06 in span-F1 score compared to the baseline, securing the third place among all participants and

Table 2
Performance of Various Models on Task 1 Test Set

Task	Model	MCC	F1-macro	F1-conspiracy	F1-critical
Task1-English	**CT-BERT**	**0.8198**	**0.9098**	**0.8811**	**0.9396**
Task1-English	Baseline	0.7964	0.8975	0.8632	0.9318
Task1-Spanish	CT-BERT	0.6192	0.8048	0.7391	0.8706
Task1-Spanish	Multi-model Voting	0.6465	0.8186	0.7579	0.8794
Task1-Spanish	Baseline	0.6681	0.8339	0.7872	0.8806

surpassing the majority of the competing methods. On task2-Spanish, the model also exceeded the baseline, meeting the expected outcomes. These results indicate that the CT-BERT and BETO models are suitable for this token-level classification task and exhibit commendable performance.

Table 3
Performance of Various Models on Task 2 Validation Set

Model	Training Set	Validation Set	Span-F1	Span-P	Span-R	Micro-span-F1
CT-BERT	English	English	0.5413	0.4756	0.6666	0.5165
BETO	Spanish	Spanish	0.4973	0.4165	0.6225	0.4798

Table 4
Performance of Various Models on Task 2 Test Set

Task	Model	Span-F1	Span-P	Span-R	Micro-span-F1
Task2-English	**CT-BERT**	**0.5886**	**0.5243**	**0.6834**	**0.5571**
Task2-English	Baseline	0.5323	0.4684	0.6334	0.4998
Task2-Spanish	**BETO**	**0.4994**	**0.4530**	**0.5740**	**0.4890**
Task2-Spanish	Baseline	0.4934	0.4533	0.5621	0.4952

4. Conclusion

In the present study, the text classification was realized by fine-tuning the BERT model, accomplishing all subtasks of the Oppositional Thinking Analysis 2024 task. This included distinguishing between texts of critical and conspiratorial nature and detecting elements within oppositional narratives. The results surpassed the baseline performance in the majority of the evaluated metrics. The top-performing model achieved an MCC score of 0.8198 on the task1-English test dataset and a span-F1 score of 0.5886 on task2-English, outperforming the majority of competing models. This suggests that the CT-BERT model is able to obtain higher-quality text encoding in this task, demonstrating its superior performance.

For future work, other outstanding models tailored to the task scenario can be explored, with the aim of leveraging ensemble learning to develop a robust learner and thereby enhance model performance. In addition, Analyzing the correlation between narrative elements and text categories statistically can reveal potential patterns and trends in text classification, providing additional features and information for the model, which may further improve the accuracy of classification.

Acknowledgments

This work is supported by the Social Science Foundation of Guangdong Province, China (No.GD24CZY02)

References

[1] K. M. Douglas, R. M. Sutton, What are conspiracy theories? a definitional approach to their correlates, consequences, and communication, Annual Review of Psychology 74 (2023) 271–298. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-psych-032420-031329. doi:https://doi.org/10.1146/annurev-psych-032420-031329.

[2] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, E. Stakovskii, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[3] D. Korenčić, B. Chulvi, X. B. Casals, M. Taulé, P. Rosso, F. Rangel, Overview of the Oppositional Thinking Analysis PAN Task at CLEF 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[4] M. Müller, M. Salathé, P. E. Kummervold, Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, 2020. arXiv:2005.07503.

[5] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, 2023. arXiv:2308.02976.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. arXiv:1706.03762.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[8] S. Ruder, An overview of multi-task learning in deep neural networks, 2017. arXiv:1706.05098.

[9] D. Chicco, N. Tötsch, G. Jurman, The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation, BioData Mining 14 (2021). URL: https://api.semanticscholar.org/CorpusID:231816620.

[10] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, Fine-grained analysis of propaganda in news article, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5636–5646. URL: https://aclanthology.org/D19-1565. doi:10.18653/v1/D19-1565.