An Oppositional Thinking Analysis Method Using BERT-based Model with BiGRU

Notebook for PAN at CLEF 2024

Qingbiao Hu, Zhongyuan Han*, Jiangao Peng, Mingcan Guo and Chang Liu

Foshan University, Foshan, China

Abstract

The Oppositional thinking analysis: Conspiracy theories vs critical thinking narratives task of PAN at CLEF 2024 involves two challenges: first, distinguishing between conspiracy and critical narratives as Subtask 1, and second, identifying key elements of oppositional narratives as Subtask 2. We consider these two challenges as binary classification and sequence labeling problems, respectively. We will perform both tasks in English and Spanish. In this paper, we introduce our method to address these challenges by fine-tuning a BERT-based model with an added BiGRU layer for Subtask 1 and employing a multi-task learning method for Subtask 2. Finally, our model for English achieves MCC scores of 0.821 in Subtask 1 and Span-F1 scores of 0.569 in Subtask 2 on the official test set.

Keywords

PAN 2024, Oppositional Thinking Analysis, BERT-based Model, Multi-task Learning

1. Introduction

As it is acknowledged that conspiracy theories pose significant harm to society and are challenging to identify [1], the difficulty lies in distinguishing them from critical thinking narratives, as both share similarities in oppositional thinking. However, it is crucial to differentiate between them, as failure to do so could push people toward conspiracy communities, as shown in [2]. The PAN at CLEF 2024 task [3] on oppositional thinking analysis [4] aims to address this problem. It includes two subtasks framed as a binary classification task and a token-level classification task, respectively.

The automatic detection of conspiracy theories in text using pre-trained language models has proven effective [5] in recent years. Combining the transformer-based model with downstream neural networks has achieved state-of-the-art performance in similar tasks [6]. Inspired by related works, we employ CT-BERT [7] and BiGRU (Bidirectional Gated Recurrent Units) [8] to address this task. By integrating the BERT-based layer with the BiGRU layer, we leverage the benefits of deep contextual embeddings and sequence-sensitive features.

2. Oppositional thinking analysis Task

At PAN 2024 there are two subtasks proposed for oppositional thinking analysis:

- ***Subtask 1: Distinguishing between critical and conspiracy texts***. It is a binary classification task that aims to distinguish between two types of messages: the first contains critical messages that scrutinize significant decisions within the public health sector without endorsing a conspiratorial mindset; the second includes messages that interpret the pandemic or public health decisions as the result of a malignant conspiracy orchestrated by secretive, powerful entities. Our task is to categorize these texts into distinct categories: CONSPIRACY or CRITICAL.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

ezio411152084@gmail.com (Q. Hu); hanzhongyuan@gmail.com (Z. Han); wyd1n910@gmail.com (J. Peng); gmc9812@163.com (M. Guo); lc965024004@gmail.com (C. Liu)

0009-0004-8237-0044 (Q. Hu); 0000-0001-8960-9872 (Z. Han); 0009-0006-3780-5023 (J. Peng); 0000-0002-4977-2138 (M. Guo); 0009-0000-0887-9273 (C. Liu)

- ***Subtask 2: Detecting elements of the oppositional narratives.*** It is a token-level classification task aimed at recognizing text spans corresponding to the key elements of oppositional narratives. A span-level annotation scheme that identifies the Agents (A), Facilitators (F), Campaigners (C), Victims (V), Effects (E), Objectives (O) in the oppositional narratives was developed. Our task is to identify specific spans in texts that should be annotated with the corresponding labels.

3. Method

Generally speaking, our method consists of two main parts: the BERT-based encoder and the BiGRU downstream neural network layer for both Subtask 1 and Subtask 2. Our method involves three primary steps: 1) fine-tune the pre-trained BERT-based model with the given training dataset, 2) feed the sequence of embeddings from the BERT-based model into a BiGRU layer and 3) Use the outputs from the BiGRU layer, typically the final hidden states that encapsulate the information from the entire sequence, to classify the text into categories (e.g., critical or conspiracy) in Subtask 1 or to combine with different task heads for span annotation in Subtask 2.

3.1. BERT-based Model with BiGRU Layer Architecture for Subtask 1

In this section, we introduce the architecture for Subtask 1. Figure 1 shows the whole architecture.

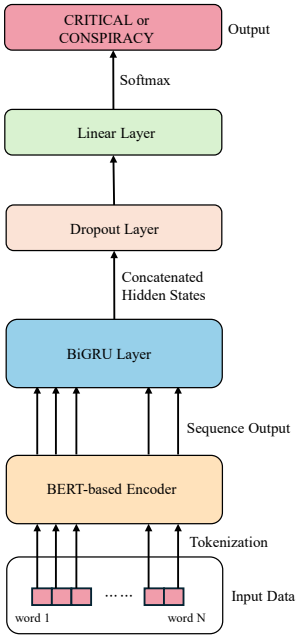

Figure 1: Model Architecture for Subtask 1. This architecture enhances BERT's contextual embeddings with a BiGRU layer for bidirectional sequential processing, which, after dropout regularization, feeds into a linear layer for final classification.

The CT-BERT model is selected as our encoder, which was trained on a large dataset of COVID-19 Twitter messages. The corpus for this PAN 2024 task consists of COVID-19 Telegram texts, making our model particularly well-suited due to its training on similar content. Consequently, this model is expected to outperform other BERT-based models due to its superior understanding of this specific domain. Additionally, we have chosen RoBERTa [9] as a contrasting model to verify whether these expectations hold.

The BERT-based model provides rich contextual embeddings by considering the left and right contexts within the transformer architecture. The addition of a BiGRU layer introduces an extra level of sequential processing. It processes information in both forward and backward directions across the text, offering

a comprehensive view of the temporal dependencies. Once the BERT-based layer has generated the sequence outputs, they are fed into the BiGRU layer. The BiGRU layer synthesizes the information captured by the BERT layer, adding a layer of understanding. This enhancement aids in detecting subtle cues and patterns that differentiate various narrative types.

The BiGRU outputs are then passed through additional dropout layers for regularization, followed by a linear classification layer that maps the BiGRU outputs to the target category.

3.2. Multi-task Learning Architecture for Subtask 2

The core architecture for Subtask 2 remains the same, however, we employ a multi-task learning method to more effectively address the specific challenges posed by Subtask 2, as shown in Figure 2.

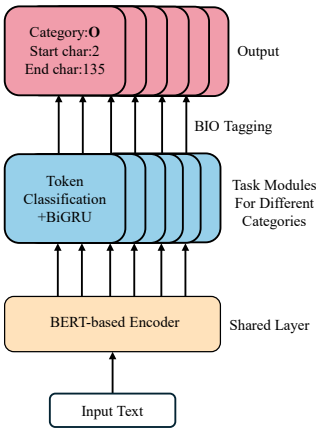

Figure 2: Model Architecture for Subtask 2.This architecture uses a BERT-based encoder shared layer and BiGRU-enhanced token classification layers with BIO tagging for different categories, creating a multi-task classifier that identifies text elements in six categories.

Given that the key elements to be identified in a text fall under one of six categories— Agents (A), Facilitators (F), Campaigners (C), Victims (V), Effects (E), and Objectives (O)—each can be considered a separate token classification task. All these tasks share the same need for embeddings. Therefore, we utilize a BERT-based encoder (primarily CT-BERT) as the backbone of our architecture, with token classification layers serving as task-specific heads. This forms our multi-task classifier architecture. Additionally, the token classification layer is integrated with a BiGRU layer, and through BIO tagging, we achieve the span output for each category.

Recent research [10] has proven the effectiveness of a multi-task classifier based on the domain-specific CT-BERT model. Utilizing a shared encoder, our model efficiently learns universal representations beneficial across all tasks, while the dedicated task modules concentrate on task-specific features.

4. Experiments and Results

4.1. Datasets

Given these two subtasks, the oppositional thinking analysis task has provided datasets [11] consisting of Telegram texts related to COVID-19 from a list of oppositional Telegram channels, available in both English and Spanish. The data has been pre-processed and tokenized for convenience, with emojis and other non-text content removed. The training datasets include lists of texts fully annotated with categories and spans of key elements, whereas the test datasets contain only the input texts. A total of 5000 texts for each language have been provided.

4.2. Evaluation

For evaluation, we used the official metrics provided to evaluate Subtask 1: **Matthews Correlation Coefficient (MCC)** [12], per-class F1 scores: **F1-Consp** and **F1-Crit** and **macro-averaged F1**.

And we used the following metrics in Subtask 2: **span-F1** [13], **span-recall**, **span-precision** and **micro-span-F1**.

4.3. Baseline

The organisers of each subtask provided baselines in both languages for each subtask. BERT classifier is used for Subtask 1, and BERT-based multi-task token classifier is used for Subtask 2.

4.4. Settings

While training, we preprocessed the training set and divided it using stratified 3-fold cross-validation.

Our model is trained using a cross-entropy loss function and utilizes the AdamW optimizer with a learning rate of 2e-5, incorporating a scheduler for learning rate adjustments. Other hyperparameters include a batch size of 16 and a training duration of three epochs.

In Subtask 1, we selected CT-BERT and RoBERTa for experiments on the English corpus, and bert-spanish [14] for the Spanish corpus. Each model was tested both with and without an added BiGRU layer. In Subtask 2, we selected CT-BERT as backbone on the English corpus, and bert-spanish for the Spanish corpus. Each model was tested both with an added BiGRU layer.

4.5. Results

During the training process for Subtask 1, we evaluated our models and compared them with the official baselines. We anticipate that the CT-BERT + BiGRU model will outperform other models on the English corpus. For the Spanish corpus, due to the limited availability of multilingual models for experimentation, we used BERT-Spanish with a BiGRU layer.

As shown in Table 1, our model performed better than both the baseline and RoBERTa + BiGRU, demonstrating the effectiveness of the CT-BERT + BiGRU model in this binary classification task. When compared with CT-BERT without the BiGRU, the version with BiGRU showed slight improvement. However, the BERT-Spanish + BiGRU model slightly fell short of the Spanish baseline.

The Table 2 shows that our model still holds up, indicating that our model is robust and neither overfits nor underfits the training set. However, the BERT-Spanish + BiGRU model performed worse than the baseline.

Table 1
Results for SubTask 1 on training sets

Model	Language	MCC	F1-Consp	F1-Crit	F1-avg
Baseline	English	0.729	0.819	0.908	0.863
CT-BERT + BiGRU	English	**0.815**	0.878	**0.936**	**0.907**
CT-BERT	English	0.808	0.872	0.935	0.903
RoBERTa + BiGRU	English	0.789	0.859	0.928	0.894
RoBERTa	English	0.783	**0.928**	0.853	0.890
Baseline	Spanish	**0.677**	**0.790**	**0.886**	**0.838**
BERT-spanish + BiGRU	Spanish	0.662	0.776	0.882	0.829

In relation to Subtask 2, and similar to the approach in Subtask 1, we compared the CT-BERT + BiGRU model and the BERT-Spanish + BiGRU model with the baseline model during training to evaluate if this multi-task architecture still performs better. Subsequently, we submitted our best model for testing on the official test sets. Table 3 and Table 4 demonstrate the results obtained in Subtask 2.

Table 2
Results for Subtask 1 on official testing sets

Model	Language	MCC	F1-Consp	F1-Crit	F1-avg
Baseline	English	0.796	**0.863**	0.931	0.897
CT-BERT + BiGRU	English	**0.821**	0.821	**0.940**	**0.909**
Baseline	Spanish	**0.668**	**0.787**	0.880	**0.833**
BERT-spanish + BiGRU	Spanish	0.653	0.768	**0.880**	0.824

Table 3
Results for Subtask 2 on training sets

Model	Language	span-F1	span-P	span-R	micro-span-F1
Baseline	English	0.522	0.453	0.640	0.510
CT-BERT + BiGRU	English	**0.576**	**0.516**	**0.667**	**0.542**
Baseline	Spanish	0.475	0.429	**0.544**	0.475
BERT-spanish + BiGRU	Spanish	**0.475**	**0.440**	0.527	**0.483**

Table 4
Results for Subtask 2 on official testing sets

Model	Language	span-F1	span-P	span-R	micro-span-F1
Baseline	English	0.532	0.468	0.633	0.499
CT-BERT + BiGRU	English	**0.569**	**0.522**	**0.633**	**0.538**
Baseline	Spanish	**0.493**	0.453	**0.562**	**0.495**
BERT-spanish + BiGRU	Spanish	0.486	**0.462**	0.522	0.494

5. Conclusion

This paper mainly introduces our work on oppositional thinking analysis at PAN 2024. Our work utilizes a BERT-based model with a BiGRU layer to enhance performance in both binary classification and sequence labeling tasks within this domain. The results from the official testing datasets indicate that our method achieved an improvement of approximately 0.04 MCC scores in Subtask 1 and reached 4th place in the Official Ranking for the English corpus.

While the English model demonstrated strong performance, the Spanish model was less successful, with only marginal improvements attributed to the BiGRU layer. Therefore, future work should focus on investigating how this method impacts multilingual tasks.

Acknowledgments

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

References

[1] K. M. Douglas, J. E. Uscinski, R. M. Sutton, A. Cichocka, T. Nefes, C. S. Ang, F. Deravi, Understanding conspiracy theories, Political psychology 40 (2019) 3–35.

[2] S. Phadke, M. Samory, T. Mitra, What makes people join conspiracy communities? Role of social factors in conspiracy engagement, Proceedings of the ACM on Human-Computer Interaction 4 (2021) 1–30.

[3] 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.

[4] 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.

[5] K. Pogorelov, D. T. Schroeder, S. Brenner, J. Langguth, FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task at MediaEval 2021., in: MediaEval, 2021.

[6] J. Alghamdi, Y. Lin, S. Luo, Towards covid-19 fake news detection using transformer-based models, Knowledge-Based Systems 274 (2023) 110642.

[7] M. Müller, M. Salathé, P. E. Kummervold, Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, Frontiers in artificial intelligence 6 (2023) 1023281.

[8] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, arXiv preprint arXiv:1412.3555 (2014).

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[10] Y. Peskine, G. Alfarano, I. Harrando, P. Papotti, R. Troncy, Detecting COVID-19-Related Conspiracy Theories in Tweets., in: MediaEval, 2021.

[11] D. Korenčić, B. Chulvi, X. Bonet Casals, M. Taulé, P. Rosso, PAN24 Oppositional Thinking Analysis [Data set], https://doi.org/10.5281/zenodo.11199642, 2024. Available from Zenodo.

[12] 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) 1–22.

[13] G. Da San Martino, Y. Seunghak, A. Barrón-Cedeno, R. Petrov, P. Nakov, et al., Fine-grained analysis of propaganda in news article, in: 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, 2019, pp. 5636–5646.

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