A Conspiracy Theory Text Detection Method based on RoBERTa and XLM-RoBERTa Models

Notebook for PAN at CLEF 2024

Zhengqiao Zeng[1], Zhongyuan Han[1,*], Jingyan Ye[1], Yaozu Tan[1], Haojie Cao[1], Zengyao Li[1] and Runjin Huang[2]

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

Abstract

Conspiracy theories are complex narratives that attempt to explain the ultimate causes of significant events as cover plots orchestrated by secret, powerful, and malicious groups. To analyze texts affecting adversarial thinking containing conspiratorial or critical narratives, PAN 2024 introduces Adversarial Thinking Analysis: Conspiracy vs. Critical Thinking Narratives. This evaluation subtask comprises two sub-tasks: subtask 1 requires distinguishing texts questioning public health decisions without endorsing conspiracy theories from those attributing these decisions to malicious conspiracies; subtask 2 involves extracting, identifying, and classifying key elements of adversarial narratives. The hyperparameters of the RoBERTa and XLM-RoBERTa models are tuned to accomplish these tasks. After the training, these models are employed to make predictions and evaluate the test set. Ultimately, the following metrics are achieved in performance: in subtask 1, an MCC of 0.7758 for English texts and an MCC of 0.6871 for Spanish texts are obtained. In subtask 2, the span-F1 score reached 0.5666 for English and 0.4903 for Spanish texts.

Keywords

PAN 2024, RoBERTa, XLM-RoBERTa, Conspiracy, Critical Thinking

1. Introduction

In the current epoch of information, an open communication environment brings convenience and freedom; it also allows malicious groups to confuse conspiracy with critical texts through intricate narratives, leading to a misguided public understanding of significant events and steering them in the wrong direction. PAN 2024 proposes Adversarial Thinking Analysis: Conspiracy vs. Critical Thinking Narratives to address this issue[1]. This evaluation subtask comprises two subtasks: subtask 1 is a binary classification subtask differentiating between 1) critical messages that question significant decisions in the public health domain but do not promote a conspiracist mentality and 2) messages that view the pandemic or public health decisions as a result of an evil conspiracy by secret, influential groups. Subtask 2 is a token-level classification subtask aimed at recognizing text spans corresponding to the key elements of oppositional narratives[2]. The primary work involves classifying critical and conspiracy texts based on RoBERTa[3] and XLM-RoBERTa[4] models and identifying narrative elements corresponding to each text. The RoBERTa model is utilized for English and the XLM-RoBERTa model for Spanish. Text preprocessing is performed before the model training phase.

In subtask 1, the cross-validation method is applied to determine the hyperparameters of the RoBERTa and XLM-RoBERTa models, followed by model training. In subtask 2, the cross-validation method is applied to determine the hyperparameters of the RoBERTa and XLM-RoBERTa models, and then a

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
zhengqiaozeng@163.com (Z. Zeng); hanzhongyuan@gmail.com (Z. Han); yyyxy0604@163.com (J. Ye); tanyaozu2023@163.com (Y. Tan); caohaojie0322@163.com (H. Cao); Izy1512192979@gmail.com (Z. Li); ruijin_huang@163.com (R. Huang)
0009-0000-5415-349X (Z. Zeng); 0000-0001-8960-9872 (Z. Han); 0009-0002-2749-9422 (J. Ye); 0009-0003-0184-3050 (Y. Tan); 0000-0002-8365-168X (H. Cao); 0000-0001-8472-4150 (Z. Li)

multi-tasking approach is conducted. Subsequently, these models are used to predict and evaluate the test set.

2. Method

Subtask 1 is approached as a binary classification task, and Subtask 2 is approached as a token-level classification subtask. The research methodology consists of three steps: 1) data preprocessing, 2) model training, and 3) prediction.

For subtask 1, in the first step, text preprocessing is performed. In the second step, due to the limited sample size, the cross-validation method is applied to determine the hyperparameters of the RoBERTa and XLM-RoBERTa models for subtask 1. In the third step, the trained models are used to classify the text and tested against the official test set.

For subtask 2, the first step is to transform the text into a format acceptable to the models. Then, the RoBERTa and XLM-RoBERTa models are used for training. In the third step, the trained models are used for text recognition, followed by testing on the official test set.

3. Experiment

3.1. Dataset

The model is trained on the training data provided by the evaluation party. The training data can be accessed through a JSON file, part of the PAN@CLEF2024 shared task **Oppositional Thinking Analysis**. This JSON file contains 4000 texts in English and Spanish, encompassing all texts and their respective annotations in the training dataset. Each text entry is in dictionary format, recording the text ID, tokenized text content, binary category label, and span annotations. The span annotations consist of a series of dictionaries detailing a specific annotation span, including its category, start and end character indices, and corresponding text snippet. This link allows researchers to request access to dataset [1] designed specifically for subtasks 1 and 2.

3.2. Data Processing

3.2.1. Subtask1 Data Processing

For subtask 1, a binary classification approach based on RoBERTa and XLM-RoBERTa is adopted to distinguish between two types of texts: Those questioning public health decisions without propagating conspiracy theories and those attributing these decisions to malicious conspiracies. Initially, for datasets in English and Spanish, the 'text,' 'id,' and 'category' are extracted from each data entry and subsequently regarded as forming a new dataset.

3.2.2. Subtask2 Data Processing

For subtask 2, the BIO tagging method is employed to identify and classify narrative elements within the text. This approach can identify the starting and ending positions and hierarchical structure of each narrative element. The text is tokenized, and each token is labeled (where 'B' denoted the beginning of a narrative element, 'I' represented the inside of a narrative element, and 'O' stood for outside a narrative element) based on a predefined set of tags. Subsequently, the annotated data is fed into RoBERTa and XLM-RoBERTa models for model training.

[1] https://zenodo.org/records/11199642

3.3. Model training

3.3.1. Subtask1 Model training

For subtask 1, to address the issue of limited data volume, a k -fold cross-validation method is employed, with the value of k set to 5, evenly dividing the dataset into five parts. During each round of cross-validation, one subset is selected as the validation set, while the remaining four subsets are combined to form the training set. Conducting five such tests allows each sample in the dataset can be used for training and validation. Ultimately, based on the results of these five sets of models, the hyperparameter combination that yields a better average performance is selected as the hyperparameter setting for the model. Ultimately, 25 epochs are selected for the training cycle (epochs attempted: 10, 15, 20, 25, and 30). The learning rate of $1e-5$ is chosen (learning rates attempted: $1e-5$, $2e-5$, and $3e-5$). The batch size of 64 is selected (batch sizes attempted: 32, 64, and 128). Model training is executed on NVIDIA A800 TENSOR CORE GPU hardware, where the cross-entropy loss function is opted for to measure the discrepancy between predictions and accurate labels, and the Adam [5] optimizer is employed to adjust the parameters of the model.

3.3.2. Subtask2 Model training

For subtask 2, the k-fold cross-validation method and multi-tasking approach are utilized, with the k value set to 5, thereby allowing each sample in the dataset to serve as training and validation data. Ultimately, 20 epochs are chosen for the training cycle of the model (epochs attempted: 10, 15, 20, 25, and 30). The learning rate of $1e-5$ is selected for the model (learning rates attempted: $1e-5$, $2e-5$, and $3e-5$). The batch size of 64 is selected (batch sizes attempted: 32, 64, and 128). During the training process, the multi-tasking approach is employed, where each named entity is assigned to a task, and tasks share parameters through RoBERTa or XLM-RoBERTa as they are trained. The training is conducted on NVIDIA GeForce RTX 3090 hardware. Regarding the selection of the loss function, the cross-entropy loss function is employed to calculate the discrepancy between predicted and actual labels, and the parameters are adjusted using the Adam optimizer.

4. Results

The official evaluation metric for subtask 1 (critical vs. conspiracy classification) is MCC, while the official metric for subtask 2 (span-level detection of narrative elements) is macro-averaged span-F1.[2] Based on the experiments, rankings of 6 out of 78 in task 1-SPANISH and 10 out of 28 in task 2-ENGLISH are achieved, surpassing baseline. Rankings of 31 out of 83 in task 1-ENGLISH and 12 out of 25 in task 2-SPANISH are obtained, neither of which exceeds baseline scores.

Table 1
TASK1 - ENGLISH

TEAM	MCC	F1-MACRO	F1-CONSPIRACY	F1-CRITICAL	POSITION
zhengqiaozeng	0.7758	0.8866	0.8476	0.9256	31
baseline-BERT	**0.7964**	**0.8975**	**0.8632**	**0.9318**	**18**

Table 2
TASK 1 - SPANISH

TEAM	MCC	F1-MACRO	F1-CONSPIRACY	F1-CRITICAL	POSITION
zhengqiaozeng	**0.6871**	**0.8417**	**0.7925**	**0.8909**	**6**
baseline-BERT	0.6681	0.8339	0.7872	0.8806	14

Table 3
TASK2 - ENGLISH

TEAM	SPAN-F1	SPAN-P	SPAN-R	MICRO-SPAN-F1	POSITION
zhengqiaozeng	**0.5666**	**0.5122**	**0.6485**	**0.5421**	**10**
baseline-BETO	0.5323	0.4684	0.6334	0.4998	16

Table 4
TASK 2 - SPANISH

TEAM	SPAN-F1	SPAN-P	SPAN-R	MICRO-SPAN-F1	POSITION
zhengqiaozeng	0.4903	0.4507	0.5494	0.4874	12
baseline-BETO	**0.4934**	**0.4533**	**0.5621**	**0.4952**	**11**

5. Conclusion

For subtask 1 in English using RoBERTa, the score does not exceed the baseline when compared with it. The reason may be that the parameters determined by the k-fold cross-validation method are not effective, thus affecting the performance of the model. The score for subtask 2 in English using RoBERTa exceeds the baseline model, possibly because the k-fold cross-validation method determines a hyperparameter combination with better average performance.

For the Spanish texts of subtask 1 and subtask 2, a multi-tasking approach is employed to train the XLM-RoBERTa model, with the anticipation that the scores will surpass the baseline. Upon comparing the outcomes with the baseline, the score for subtask 1 in Spanish exceeds the baseline by 0.019 points. For subtask 2 with Spanish texts, the score does not surpass the baseline. This outcome is not anticipated, which is the research focus of our further research.

Acknowledgments

This work is supported by the Natural Science Platforms and Projects of Guangdong Province Ordinary Universities (Key Field Special Projects) (No. 2023ZDZX1023)

References

[1] 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.

[2] D. Korenčić, B. Chulvi, X. Bonet Casals, M. Taulé, P. Rosso, F. Rangel, Overview of the oppositional thinking analysis pan task at clef 2024, in: G. Faggioli, N. Ferro, P. Galuscakova, A. García Seco de Herrera (Eds.), Working Notes of CLEF 2024–Conference and Labs of the Evaluation Forum, 2024. URL: https://doi.org/10.5281/zenodo.10680586.

[3] 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, 2019. arXiv:1907.11692.

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott,

L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. arXiv:1911.02116.
[5] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. arXiv:1711.05101.
[6] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.