Conspiracy Theory Detection using Transformers with Multi-task and Multilingual Approaches

Notebook for PAN at CLEF 2024

Leon Zrnić

University of Zagreb, Faculty of Electrical Engineering and Computing

Abstract

The COVID-19 pandemic sparked a new age of conspiracy theories in society. This has become an issue, especially since these theories are mixed in with reasonable arguments that criticize the measures taken by governments and their effects. A way to help differentiate these two narratives is using natural language processing (NLP) models such as Transformers. These working notes detail a few approaches to classifying conspiracy and critical narratives and the identification of the key narrative elements present in these texts. We employ these models on two datasets which encompass English and Spanish texts on Telegram talking about the COVID-19 pandemic. Our approaches include using pre-trained BERT and RoBERTa models on monolingual datasets, a multilingual approach in which we translate the Spanish texts into English and the use of a multilingual model on non-translated texts, and using a multi-task model architecture in the identification of narrative elements. Our results show that BERT pre-trained on COVID-19 tweets had similar results to RoBERTa in the binary classification task, while in the token classification task RoBERTa worked better. The monolingual English approach yielded better results than the multilingual one which was, however, better than the Spanish models. We conclude that transformer models can have good results in these classification tasks, making them an easy-to-deploy way to differentiate critical narratives from conspiracy theories.

Keywords

Machine learning, NLP, conspiracy theories, transformers, multilingual, multi-task model

1. Introduction

The COVID-19 pandemic has flooded digital platforms with both essential updates and conspiracy theories. This surge of information creates the challenge of distinguishing between legitimate critical narratives and harmful conspiracy theories. Critical narratives question established systems using evidence and reason, while conspiracy theories claim secret plots without substantial proof. Differentiating these is vital for effective public health communication and social stability. It ensures informed decision-making, as critical narratives drive constructive scrutiny based on evidence, while conspiracy theories spread misinformation and cause societal divisions.

One way to differentiate these narratives is through the use of natural language processing (NLP) models. These automatic classifiers can hasten the process of identifying conspiracy narratives, removing the need for human annotation in the process.

In these work notes, we describe our approach to creating and training these automatic classifiers on two separate classification tasks. The first task is a binary classification task in which an AI model differentiates between conspiracy and critical narratives. The second task sets the goal of identifying the key narrative elements present in conspiracy and critical narrative texts regarding the COVID-19 pandemic.

2. Task descriptions

As part of the PAN at CLEF 2024 *Oppositional thinking analysis: Conspiracy theories vs critical thinking narratives* shared task [1, 2], we partook in two different tasks. The first task was a binary classification

CLEF 2024: Conference and Labs of the Evaluation Forum, September 9-12, 2024, Grenoble, France

leon.zrnic@fer.hr (L. Zrnić)

task in which participants needed to make AI models that differentiated between conspiracy and critical narratives in Telegram messages about the COVID-19 pandemic. The second task was a token classification task in which models needed to find six different key narrative elements present in the Telegram messages.

3. Dataset

As mentioned, there are two `train` datasets: one with English Telegram messages and another with Spanish Telegram messages, each containing 4000 annotated messages about the COVID-19 pandemic.

Messages are labeled as either *Critical* or *Conspiracy. Critical* messages discuss the pandemic with reasoned arguments, questioning government measures. *Conspiracy* messages claim hidden plots aim to undermine freedom and establish a new world order.

Each message also has token-level annotations for six narrative elements, as defined by the dataset authors:

- **Agents**,
- **Facilitators**,
- **Victims**,
- **Campaigners**,
- **Objectives**, and
- **Negative effects**.

To further analyze the dataset, we examined the differences between texts labeled as conspiracy and critical in two ways. First, we looked at the lengths of the messages, as shown in Table 1. On average, conspiratorial messages are almost twice as long as critical messages and Spanish texts are longer than English texts.

Table 1
Statistics of text lengths (expressed in characters) by category.

Dataset	Category	Count	Mean
English	Conspiracy	1379	742.93
	Critical	2621	476.00
Spanish	Conspiracy	1462	1112.04
	Critical	2538	641.24

4. Methodology

This section describes the methodology used in our work. Subsection 4.1 briefly explains the transformer architecture and the models we used. Subsection 4.2 details the two multilingual approaches we used, namely text translation and multilingual models. In Subsection 4.3 we explain the multi-task model architecture that was used for the token classification task, and in Subsection 4.4 we detail Stratified K-fold cross-validation with which we evaluated the performance of our models during training.

4.1. Transformer models

Transformers [3] are deep learning neural network architectures primarily used in NLP. They leverage the attention mechanism proposed by [4], allowing the model to focus on crucial parts of a text for a given task. Transformers can be applied to various tasks, including text summarization, question answering, and binary and token classification, as explored in our work.

Since their introduction, many transformer architectures have emerged, with BERT (Bidirectional Encoder Representations from Transformers) [5] being one of the most popular. BERT improves on the original transformer by using a bidirectional approach, analyzing the entire sentence to determine the importance of each word, unlike the original architecture, which only considered preceding words.

The availability of pre-trained models has significantly contributed to the widespread use of transformers in the NLP community. All pre-trained models were sourced from the HuggingFace transformers library [6]. The following models were used for the binary classification task:

- English
  - `bert-base-cased` [5],
  - `bert-large-cased` [5],
  - `roberta-base` [7],
  - `roberta-large` [7], and
  - `digitalepidemiologylab/covid-twitter-bert-v2` [8] (referred to as ct-bert)
- Spanish
  - `dccuchile/bert-base-spanish-wwm-cased` [9] (referred to as bert-spanish) and
  - `PlanTL-GOB-ES/roberta-large-bne` [10] (referred to as roberta-spanish)

4.2. Multilingual approach

4.2.1. Text translation

One way to utilize all the available data is by translating one language data into another. This way we will have available twice the amount of data instead of the monolingual approach in which we use half of all the data. To this end, we used the `translate` Python package[1]. From this Python package, we implemented the `MyMemory` [11] translation provider that has several different machine translation models with a linguistic database.

Since English is a high-resource language, we decided that for this approach we would use monolingual English models that worked on a dataset that combined the original English texts and the Spanish texts translated into English.

4.2.2. Multilingual models

Using multilingual models can address the issue of utilizing only half the available data. An example is the `xlm-roberta-base` [12] model, which was "pre-trained on 2.5TB of CommonCrawl data in 100 languages".

Multilingual models leverage shared learning across languages, allowing for the use of double the data compared to monolingual models. English, a high-resource language, can enhance the classification of Spanish texts, improving performance through shared representations and transfer learning, which also aids generalization.

4.3. Multi-task model

Multi-task Learning (MTL) models [13] are trained on related tasks to create representations that can handle multiple objectives. They use two main architectures: hard parameter sharing and soft parameter sharing.

In hard parameter sharing, a single pipeline of shared layers is used, with separate task-specific layers. Figure 1 showcases this MTL architecture. In hard parameter sharing, the model has one main pipeline of shared layers while keeping task-specific layers separate for each task. This approach reduces overfitting and enables knowledge transfer between tasks. For example, representations learned from a binary classification task can aid in token classification.

[1] https://pypi.org/project/translate/#description

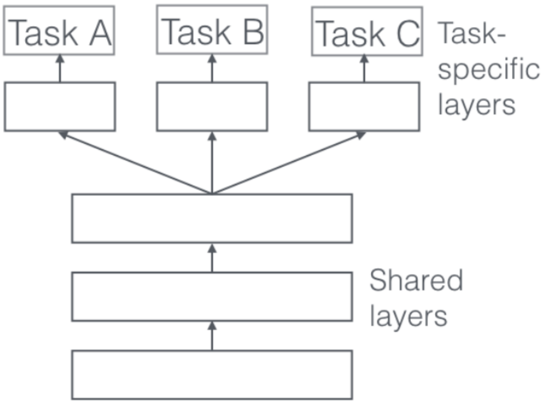

Figure 1: Hard parameter sharing MTL model [14].

The second way of making an MTL model is by soft parameter sharing. Figure 2 shows the structure of one such model. Soft parameter sharing involves separate models for each task, with regularized layers to keep parameters similar. [14] state that there are different ways of regularizing these models such as L2 distance [15] or the trace norm [16].

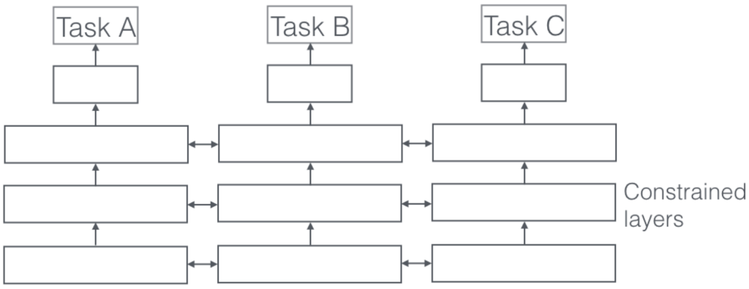

Figure 2: Soft parameter sharing MTL model [14].

We employ an MTL model for token classification using a hard parameter-sharing transformer model. It shares a common hidden layer backbone with six separate classification heads for different narrative elements. Different pre-trained transformers serve as backbones for the two datasets. Figure 3 visualizes the model architecture used.

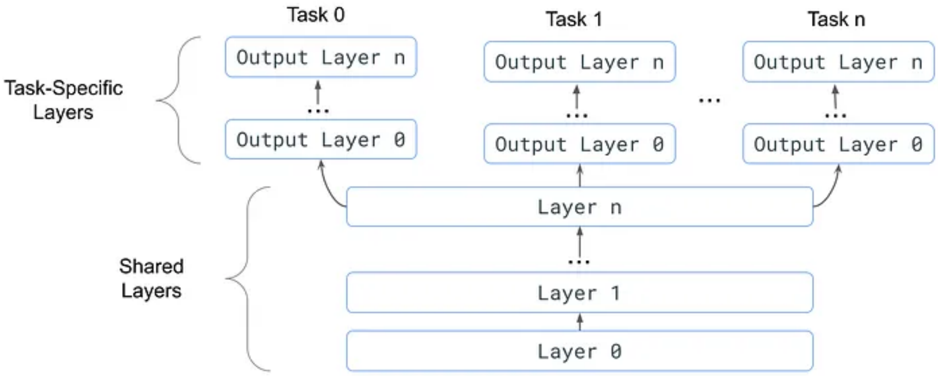

Figure 3: MTL model architecture that we used.[2]

4.4. Stratified K-fold cross-validation

Since only the `train` dataset was available for most of our work, we used an artificial `test` dataset for evaluation during training. We created this dataset using Stratified k-fold cross-validation, which splits the training set into k equal-sized subsets while preserving the class label ratio in each fold. In each epoch, the model is trained k times, using $k-1$ folds for training and the remaining fold for validation. The model's performance is then averaged across all folds and epochs. This method, implemented with Scikit-learn [17], allowed us to obtain performance scores without the official `test` dataset. The best models were ultimately evaluated on the official `test` dataset at the competition's end.

5. Experimental setup

In this section, we present the technical details regarding our setup. For both tasks, we used Stratified 5-fold cross-validation. The hyperparameters for the transformers were 10 epochs, a learning rate of $2e^{-5}$, a batch size of 32, weight decay was set to 0.01, and we set the value of warmup steps to 0.1. We also increased the maximum sequence length from the base length of 256 to 512. The models were trained on an Nvidia A100 graphics card with 40GB of RAM. All of the models and their tokenizers were from the HuggingFace [6] library.

6. Results

Here we present the results from our different approaches to the two classification tasks. In Subsection 6.1 we detail the results we got while self-validating on the `train` dataset. Subsection 6.2 shows the results of the models we submitted to evaluation on the official `test` dataset.

6.1. Experimental results

6.1.1. Task 1: Binary classification

Table 2 contains the results for the monolingual models on the English dataset, Table 3 has the results for the Spanish models, and Table 4 the results for the multilingual approaches. The `ct-bert` model had the best results on the English dataset while the `roberta-spanish` was the best in the Spanish variant. Despite `xlm-roberta-base` having slightly better results than `roberta-spanish`, we decided that this advantage was not sufficient enough to employ the model in the official evaluation for the Spanish variant of the binary classification task.

Table 2
Binary classification model performance metrics for English dataset.

Model	F1_macro	F1	F1_neg	ACC	Prec	Recall	MCC
bert-base-cased	0.87	0.91	0.83	0.88	0.91	0.91	0.74
bert-large-cased	0.90	0.93	0.86	0.91	0.92	**0.94**	0.79
ct-bert	**0.91**	**0.94**	**0.88**	**0.92**	**0.94**	**0.94**	**0.82**
roberta-base	0.89	0.93	0.86	0.90	0.92	0.93	0.78
roberta-large	**0.91**	**0.94**	**0.88**	**0.92**	0.93	**0.94**	0.81

[2] https://towardsdatascience.com/how-to-create-and-train-a-multi-task-transformer-model-18c54a146240

Table 3
Binary classification model performance metrics for Spanish dataset.

Model	F1_macro	F1	F1_neg	ACC	Prec	Recall	MCC
bert-spanish	0.85	0.89	0.80	0.86	0.88	**0.91**	0.69
roberta-spanish	0.85	0.89	**0.81**	0.86	**0.89**	0.90	**0.70**

Table 4
Binary classification model performance metrics for combined language datasets. As already stated in Subsubsection 4.2.1, the two monolingual models were trained on the translated Spanish messages combined with the English dataset. The xlm-roberta-base model worked on the combined English and Spanish dataset.

Model	F1_macro	F1	F1_neg	ACC	Prec	Recall	MCC
xlm-roberta-base	**0.87**	**0.91**	**0.83**	**0.88**	**0.90**	0.92	**0.74**
roberta-base	0.75	0.85	0.65	0.79	0.79	**0.93**	0.53
ct-bert	0.69	0.83	0.54	0.76	0.77	0.92	0.42

6.1.2. Task 2: Token classification

Here we detail the results the token classification models had during our own evaluation. Table 5 and Table 6 have the results for the English and Spanish models respectively. The `roberta-large` achieved the best results on the English dataset while `roberta-spanish` was the best model on the Spanish dataset.

Table 5
Token classification model performance metrics for English dataset.

Model	F1	A-F1	C-F1	F-F1	N_E-F1	O-F1	V-F1
bert-base-cased	0.49	0.62	0.56	0.41	0.58	0.45	0.61
bert-large-cased	0.53	0.64	0.59	0.44	0.63	0.47	0.63
ct-bert	0.52	0.66	0.61	0.48	**0.66**	0.50	0.63
roberta-base	0.55	0.64	0.65	0.45	0.61	0.48	0.66
roberta-large	**0.60**	**0.69**	**0.71**	**0.51**	0.65	**0.51**	**0.69**

Table 6
Token classification model performance metrics for Spanish dataset.

Model	F1	A-F1	C-F1	F-F1	N_E-F1	O-F1	V-F1
bert-spanish	0.50	0.50	0.52	0.40	0.66	0.35	0.58
roberta-spanish	**0.60**	**0.60**	**0.64**	**0.50**	**0.73**	**0.43**	**0.71**

6.2. Official results

This subsection contains the tables with the results we achieved on the official test dataset. Table 7 contains the binary classification results and Table 8 the results for the token classification task.

When comparing these results to the other teams competing at PAN, we placed sixth in the English variant of the first task and ninth in the Spanish variant. In the second task, we placed second in both the English and Spanish variants.

Table 7
Official test results on the PAN binary classification task. The ct-bert model was evaluated on the English dataset while the roberta-spanish model was evaluated on the Spanish dataset. F1-consp. and F1-crit. are the F1 scores when looking at only the *conspiracy* or *critical* texts.

Model	MCC	F1	F1-consp.	F1-crit.
ct-bert	0.82	0.91	0.88	0.94
roberta-spanish	0.68	0.84	0.80	0.89

Table 8
Official test results on the PAN token classification task. The roberta-large model was evaluated on the English dataset while the roberta-spanish model was evaluated on the Spanish dataset.

Model	Span-P	Span-R	Span-F1
roberta-large	0.55	0.71	0.62
roberta-spanish	0.55	0.66	0.59

7. Conclusion

In our work, we explored the application of transformer models on the tasks of binary classification of conspiracy and critical narratives and the token classification of key narrative elements in the dataset. Our results show that transformer models such as BERT and RoBERTa are highly effective in both binary and token classification tasks in the domain of COVID-19 messages. In the binary classification task, the English transformers performed better than the Spanish ones. There are many possible reasons for this, such as the data quality and size of pre-training data, the pre-training approaches, and the differences between English and Spanish. The translation approach did not succeed in achieving good results. We attribute this to the poor translation capabilities of `MyMemory`. Further work could use different translation methods, such as transformer-based machine translation [18, 19]. On the other hand, the multilingual transformer model had good results when compared to the monolingual approaches. In the token classification task, the best-performing English and Spanish models had the same F1 score. However, there were differences in the performance of the models when looking at the F1 scores for each annotation. For example, the Spanish models could detect better the *Negative effect* and *Victim* annotations. Further work should explore the differences between English and Spanish conspiracy theories on a semantical level. Multilingual models could perhaps leverage these differences to achieve better results than the monolingual models.

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, 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 - condensed lab overview, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association CLEF-2024, 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. Galuvakova, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[4] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2016. arXiv:1409.0473.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2018. URL: https://arxiv.org/abs/1810.04805. doi:10.48550/ARXIV.1810.04805.

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[7] 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.

[8] 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). URL: https://www.frontiersin.org/articles/10.3389/frai.2023.1023281. doi:10.3389/frai.2023.1023281.

[9] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[10] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[11] M. Trombetti, MyMemory: creating the world's largest translation memory, in: Proceedings of Translating and the Computer 31, Aslib, London, UK, 2009. URL: https://aclanthology.org/2009.tc-1.12.

[12] 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, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[13] R. Caruana, Multitask learning: A knowledge-based source of inductive bias, in: International Conference on Machine Learning, 1993. URL: https://api.semanticscholar.org/CorpusID:18522085.

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

[15] L. Duong, T. Cohn, S. Bird, P. Cook, Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser, in: C. Zong, M. Strube (Eds.), Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Beijing, China, 2015, pp. 845–850. URL: https://aclanthology.org/P15-2139. doi:10.3115/v1/P15-2139.

[16] Y. Yang, T. M. Hospedales, Trace norm regularised deep multi-task learning, 2017. arXiv:1606.04038.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[18] T. Tian, C. Song, J. Ting, H. Huang, A french-to-english machine translation model using transformer network, Procedia Computer Science 199 (2022) 1438–1443. URL: https://www.sciencedirect.com/science/article/pii/S1877050922001831. doi:https://doi.org/10.1016/j.procs.2022.01.182, the 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 2021): Developing Global Digital Economy after COVID-19.

[19] T. J. Sefara, S. G. Zwane, N. Gama, H. Sibisi, P. N. Senoamadi, V. Marivate, Transformer-based

machine translation for low-resourced languages embedded with language identification, in: 2021 Conference on Information Communications Technology and Society (ICTAS), 2021, pp. 127–132. doi:10.1109/ICTAS50802.2021.9394996.