Conspiracy vs Critical Thinking Using an Ensemble of Transformers with Data Augmentation Techniques

Notebook for PAN at CLEF 2024

Angelo Maximilian Tulbure[1,2], Mariona Coll Ardanuy[3]

[1]Universitat Politècnica de València, València, Spain
[2]Politecnico di Milano, Milan, Italy
[3]Pattern Recognition and Human Language Technology Research Center, Universitat Politècnica de València, València, Spain

Abstract

This paper provides an overview of our contributions to the PAN at CLEF2024 *Oppositional thinking analysis* shared task, which focuses on distinguishing between conspiratorial and critical thinking narratives. The competition featured two main tasks. The first task is a binary classification task that aims at determining whether a text is conspiratorial or critical. The second task is a span-level detection task, in which the goal is to detect elements of oppositional narratives in the texts. Two annotated datasets, one in English and one in Spanish, were provided, each of 5K telegram comments. Our best-performing approaches combined custom fine-tuned Transformer models with data augmentation techniques. We achieved an F1-Score of 0.8917 for English and of 0.8293 for Spanish for task 1, and a span-F1 score of 0.6279 for English and 0.6129 for Spanish for task 2. Our task 2 approach achieved the best results in the shared task for both English and Spanish.

Keywords

PAN 2024, oppositional thinking, critical thinking, conspiracy theories, binary classification, token classification, ensembling models, data augmentation

1. Introduction

Conspiracy theories offer elaborate explanations for significant events, attributing them to hidden schemes by secretive and powerful groups. Recently, there has been increasing interest in automatically detecting these theories in text, often framed as a binary classification problem, with some approaches exploring multi-label or multi-class classification. However, key issue with existing methods is their difficulty in distinguishing between critical thinking and conspiracy theories [1]. This distinction is crucial, because misclassifying critical perspectives as conspiracies can inadvertently lead individuals to engage more deeply with conspiracy communities. As argued in Korenčić et al. [1], conspiracy theories often proliferate rapidly on social media, leading to widespread misinformation and potential harm. In contrast, critical thinking is essential for informed decision-making and healthy public discourse.

The shared task "Oppositional thinking analysis: Conspiracy theories vs critical thinking narratives"[1] [1], which is part of PAN at CLEF2024 [2, 3] deals with the problem of distinguishing between conspiracy theories and critical thinking. The task focuses on two primary objectives: the binary classification of texts as either conspiratorial or critical, and the span-level detection of specific elements within these oppositional narratives. The datasets provided for these tasks include annotated texts from English and Spanish sources, each consisting of 5,000 telegram comments. These datasets serve as a comprehensive resource for developing and evaluating models capable of handling multilingual data and diverse narrative structures.

Our approach to these tasks involved the use of custom fine-tuned Transformer models, which were enhanced through data augmentation techniques. For the binary classification task, we employed a Soft Voting Ensembling method combining multiple Transformer models to improve robustness and accuracy.

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

angelomaximilian.tulbure@mail.polimi.it (A. M. Tulbure); mcoll@prhlt.upv.es (M. Coll Ardanuy)
0000-0001-8455-7196 (M. Coll Ardanuy)

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

For the span-level detection task, we treated the problem as a token classification task, segmenting text into sentences to mitigate issues related to text length limitations in Transformer models.

2. Systems Overview

In this section, we describe our submitted systems. Our approaches for distinguishing between critical and conspiracy texts (Task 1) are described in Section 2.1, and our approaches for detecting elements of oppositional narratives (Task 2) in Section 2.2.

2.1. Task 1: Distinguishing between Critical and Conspiracy Texts

For Task 1, the general approach involved fine-tuning Transformer-based models and applying data augmentation techniques. Both English and Spanish datasets were processed similarly. The main difference between Run 1 and Run 2 lies in the method used to make predictions. In Run 1, only the best model checkpoint was used to make predictions. In Run 2, an ensembling method was employed, which combined the predictions from multiple models.

The approach for both languages involved several key steps. We experimented with various Transformer-based models for fine-tuning and different hyperparameters. In addition to model selection and training, data augmentation played an important role in our approach. To increase the diversity and quantity of training data, we applied translation-based augmentation. For this, for English, the Spanish dataset was translated into English using the `Helsinki-NLP/opus-mt-es-en`[2] model [4]; and, for Spanish, the English dataset was translated into Spanish using the `Helsinki-NLP/opus-mt-en-es`[3] model [4]. This method helped in creating a more varied training set, enabling the models to generalize better and perform more effectively on unseen data.

In Run 1, predictions were made using the best model checkpoints identified during the training phase. However, in Run 2, we enhanced the prediction process by employing an ensembling method. This is described in more detail in the following two subsections.

2.1.1. Run 1 task 1

In Run 1, we focused on using the best model checkpoint to make predictions for both English and Spanish datasets. For English, the `facebook/roberta-base`[4] [5] model was used. This model is known for its robust performance on various NLP tasks. For Spanish, the `dccuchile/bert-base-spanish-wwm-uncased`[5] [6] model was used.

2.1.2. Run 2 task 1

In Run 2, an ensembling approach was used. The general approach involved using a Soft Voting Ensembling method composed of three custom fine-tuned Transformer-based models and data augmentation. We used the following three models for English:

1. `facebook/roberta-base`[6] [5]
2. `google/bert-base-uncased`[7] [7]
3. `allenai/scibert_scivocab_uncased`[8] [8]

For Spanish, we used the following three models:

[2] https://huggingface.co/Helsinki-NLP/opus-mt-es-en
[3] https://huggingface.co/Helsinki-NLP/opus-mt-en-es
[4] https://huggingface.co/FacebookAI/roberta-base
[5] https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased
[6] https://huggingface.co/FacebookAI/roberta-base
[7] https://huggingface.co/google-bert/bert-base-uncased
[8] https://huggingface.co/allenai/scibert_scivocab_uncased

1. `dccuchile/bert-base-spanish-wwm-uncased`[9] [6]
2. `PlanTL-GOB-ES/bsc-bio-ehr-es-pharmaconer`[10] [9]
3. `bertin-project/bertin-roberta-base-spanish`[11] [10]

For ensembling, the best checkpoint from each fine-tuned Spanish model was used in a Soft Voting ensemble for predictions. This approach involved combining predictions from multiple models using a Soft Voting ensemble [11], which significantly improved the overall accuracy and robustness of the system. The Soft Voting process involves averaging the predicted probabilities of each category from the different models and then making the final prediction based on the highest average probability. By integrating the strengths of different models, the ensembling method provided a more reliable and precise set of predictions, ensuring higher performance in distinguishing between critical and conspiracy texts.

2.2. Task 2: Detecting Elements of Oppositional Narratives

We approached Task 2 by fine-tuning a transformer model with a token classification head, therefore treating it as a token classification problem. Having only one head (instead of a classification head per label, as is implemented in the provided baseline [1]) precluded the possibility of overlapping spans, but offered increased simplicity and reduced computational expense instead. While the provided data was annotated at the document-level, we transformed it so that we could train the token classifier at the sentence-level instead. Segmenting the text into sentences overcame the problem of transformers truncating texts that are longer than the maximum length size, ensuring no data was lost during training or testing. The main difference between Run 1 and Run 2 was that, in Run 1, the best model checkpoint was used without additional training, while in Run 2, the best model checkpoint was retrained for one more epoch using the entire dataset as training.

We also performed data augmentation. We could not easily augment the data through translation, because working at the span-level means that the annotated spans are provided in terms of indices that match the original text. Therefore, we used an alternative data augmentation technique which consisted in replacing words in the texts by synonyms or semantically-related words, using static word embeddings (word2vec) [12] in combination with SpaCy [13]. This process ensured that the total number of words remained the same to maintain consistency with the start and end spacy tokens. By introducing synonym replacements, we created a more varied dataset, which helped the models generalize better and perform more effectively on unseen data. In English, we used the `GoogleNews-vectors-negative300`[12] static word embeddings while, in Spanish, we used the `FastText` embeddings[13] from the *Spanish Unannotated Corpora*.

2.2.1. Run 1 task 2

In Run 1, we used the best model checkpoint without additional training. For Task 2 in English, the `facebook/roberta-base`[14] [5] model was employed. For the Spanish dataset, the `PlanTL-GOB-ES/roberta-base-bne`[15] [14] model was used.

2.2.2. Run 2 task 2

In Run 2, the best model checkpoint from Run 1 was trained for one more epoch using the augmented dataset. This final model checkpoint was then used to detect the elements of oppositional narratives in the test dataset.

[9] https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased
[10] https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es-pharmaconer
[11] https://huggingface.co/bertin-project/bertin-roberta-base-spanish
[12] https://github.com/mmihaltz/word2vec-GoogleNews-vectors
[13] https://github.com/dccuchile/spanish-word-embeddings
[14] https://huggingface.co/FacebookAI/roberta-base
[15] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne

2.3. Experiments Setup

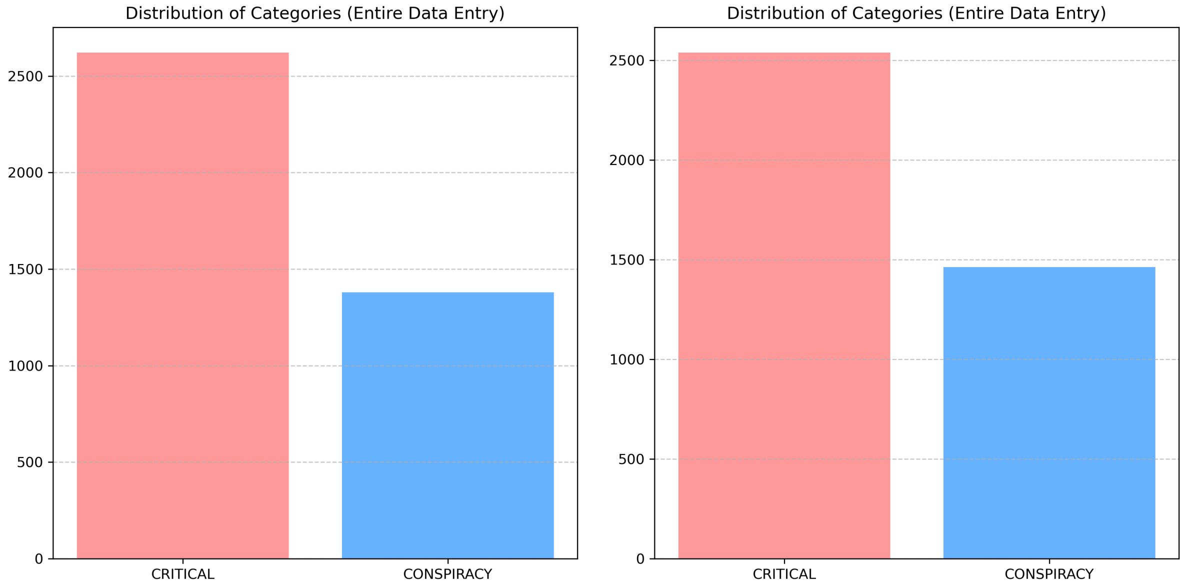
Experiments were conducted on an NVIDIA GeForce GTX 1080 (8 GB memory). To ensure uniformity and comparability of results, the same experimental setting was consistently applied across all tasks, runs, and languages under study. The experimental setup was meticulously designed to ensure optimal performance and efficient resource utilization.

StratifiedKFold cross-validation [15] with 3 folds was used to ensure robust performance across different subsets of the data. This method involves splitting the dataset into k folds while preserving the percentage of samples for each class. For each run, models were trained on $k - 1$ folds and validated on the remaining fold, rotating this process k times to ensure every data point was used for both training and validation.

The training process spanned 15 epochs. A weight decay of 0.01 was applied as a regularization technique to penalize large weights. A custom linear learning rate scheduler was employed, adjusting the learning rate from an initial value of 2e-5 to a final value of 2e-6 over the total number of training epochs. Gradient accumulation steps were set to 4, effectively increasing the batch size without inflating the memory footprint by accumulating gradients over multiple steps before updating the model's weights. The training batch size per device was dynamically set based on available GPU memory, managed by a custom callback designed to dynamically adjust the batch size used during training and evaluation based on the available GPU memory. This ensures efficient resource utilization and prevents memory-related issues during training, especially when dealing with varying data sizes and model complexities. Upon completion of training, the best model, as determined by the F1 score, was loaded.

3. Dataset

Participants were provided with a JSON file containing all texts in the training dataset along with their annotations. Each text is represented by a dictionary that includes the ID, tokenized text, binary category, and span annotations. Span annotations are provided as a list of dictionaries, with each dictionary representing an annotated span and detailing the span's category and text, specified by the start and end characters. The training subset, comprising 4000 records, was released with all annotations, while the test subset, consisting of 1000 records, was released only with "id" and the "text" field [1, 2].



(a) binary categories for the English dataset

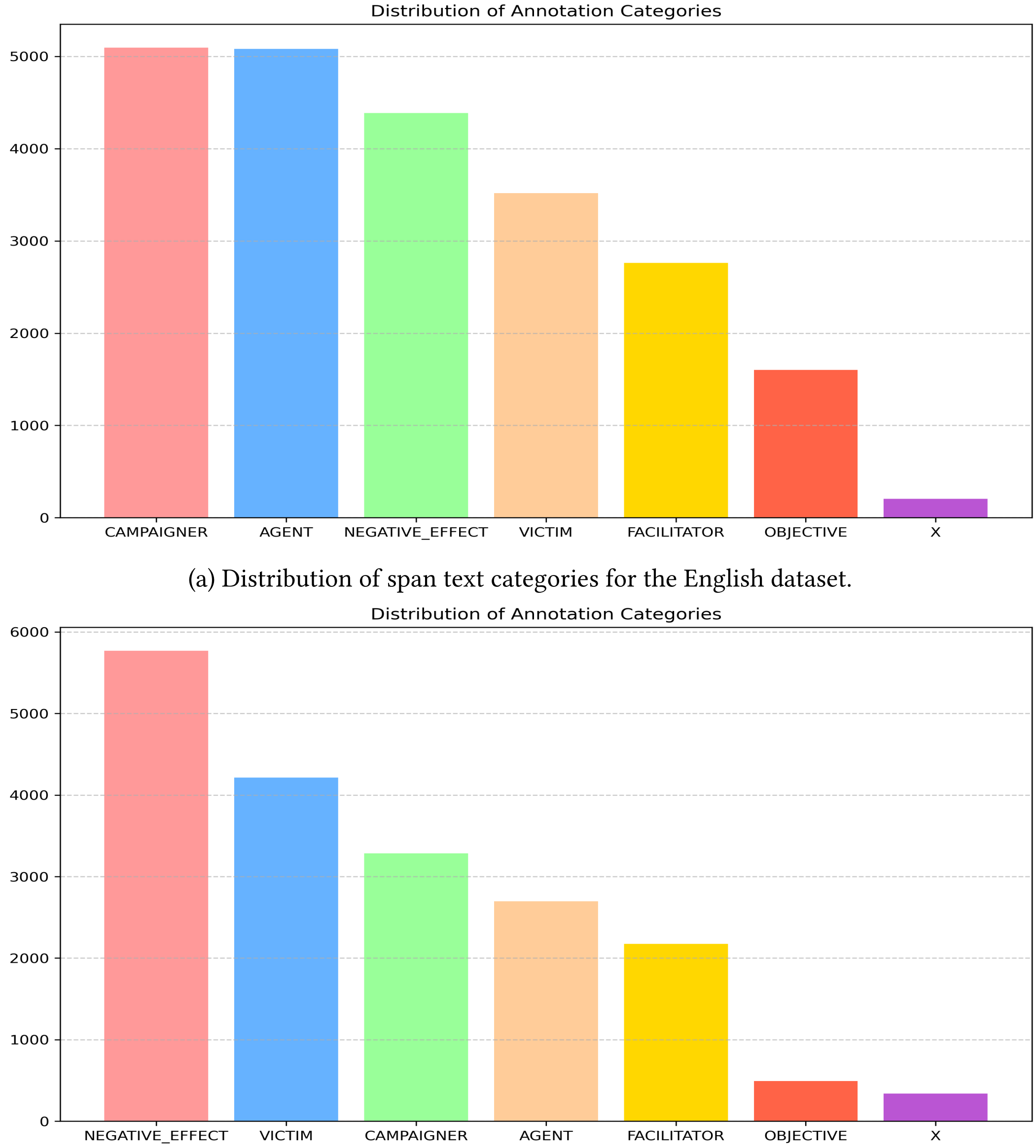
(b) binary categories for the Spanish dataset

Figure 1: Distribution of binary categories for the English and Spanish datasets.

Table 1
Distribution of binary labels for the English and Spanish datasets.

Label	English	(%)	Spanish	(%)
CRITICAL	2621	65.53%	2538	63.45%
CONSPIRACY	1379	34.48%	1462	36.55%

As summarized in Table 1 and shown in Figures 1a and 1b, the binary classification task for distinguishing critical from conspiracy texts reveals an inherent class imbalance in both the English and Spanish datasets. The English dataset has 65.53% critical texts compared to 34.48% conspiracy texts, while the Spanish dataset shows a similar trend with 63.45% critical and 36.55% conspiracy texts. This imbalance poses a challenge for model training. The similarity in the proportions of critical and conspiracy texts across both languages suggested that the models trained on one dataset might be adaptable to the other with minimal adjustments. That was indeed the approach adopted for the shared task.



(a) Distribution of span text categories for the English dataset.

(b) Distribution of span text categories for the Spanish dataset.

Figure 2: Distribution of span text categories for the English and Spanish datasets. The label "x" represents the texts where no label appears for the task.

Table 2
Distribution of span text labels for the English and Spanish datasets.

Label	English	(%)	Spanish	(%)
CAMPAIGNER	5096	22.70%	3285	17.63%
AGENT	5082	22.63%	2698	14.47%
NEGATIVE_EFFECT	4387	19.54%	5770	30.96%
VICTIM	3517	15.67%	4213	22.61%
FACILITATOR	2763	12.31%	2174	11.67%
OBJECTIVE	1602	7.14%	493	2.65%

The token classification task aims to identify key elements within oppositional narratives. As shown in Figures 2a and 2b and in Table 2, the distribution of these elements varies significantly between the English and Spanish datasets.

4. Results

The official evaluation metrics were different for each subtask. For subtask 1, which involved distinguishing between critical and conspiracy narratives, the Matthews Correlation Coefficient (MCC) was used as the primary evaluation metric [16]. Additionally, the F1-macro, F1-conspiracy and F1-critical metrics were provided for subtask 1. For subtask 2, focusing on span-level detection of narrative elements, the macro-averaged span-F1 was used as the primary evaluation metric [17]. In addition, the span-P, span-R and micro-span-F1 metrics were reported for subtask 2.

The shared task organizers provided two baseline models for these tasks. For subtask 1, the baseline is a standard BERT classifier [7]. For subtask 2, the baseline is a BERT-based multi-task token classifier with separate classification heads and a common transformer backbone [18]. The baselines utilize either English or Spanish BERT models, depending on the language. The performances of our approaches and the baselines are reported in Table 3 (Subtask 1) and in Table 4 (Subtask 2).

Table 3
Performance metrics for Subtask 1.

	Subtask 1			
	MCC	F1-macro	F1-conspiracy	F1-critical
English_baseline	*0.7964*	*0.8975*	*0.8632*	*0.9318*
English_run1	0.7574	0.8769	0.8338	0.9200
English_run2	0.7872	0.8917	0.8536	0.9297
Spanish_baseline	*0.6681*	*0.8339*	*0.7872*	*0.8806*
Spanish_run1	0.6147	0.7950	0.7179	0.8720
Spanish_run2	0.6722	0.8293	0.7699	0.8887

Table 4
Performance metrics for Subtask 2.

	Subtask 2			
	span-F1	span-P	span-R	micro-span-F1
English_baseline	*0.5323*	*0.4684*	*0.6334*	*0.4998*
English_run1	0.6293	0.5832	0.6856	0.6074
English_run2	0.6279	0.5859	0.6790	0.6120
Spanish_baseline	*0.4934*	*0.4533*	*0.5621*	*0.4952*
Spanish_run1	0.6089	0.5997	0.6193	0.6051
Spanish_run2	0.6129	0.6159	0.6129	0.6108

5. Discussion

Tables 3 and 4 offer a summary of how the models performed in distinguishing between conspiratorial and critical thinking narratives and in detecting narrative elements within texts. In this section, we discuss the reported results.

5.1. Subtask 1: Binary Classification of Conspiratorial vs. Critical Thinking

5.1.1. English Results task 1

In our analysis of the English dataset, the first run exhibited good performance metrics. The Matthews Correlation Coefficient (MCC) was 0.7574, indicating a robust ability to distinguish between different narrative types. The F1-macro score of 0.8769 further supported the model's high overall classification capability. Notably, the F1 score for critical thinking texts was 0.92, compared to 0.8338 for conspiracy texts. This disparity suggests that the model more effectively identified critical thinking. However, it is important to note that these results are below the baseline, which had an MCC of 0.7964 and an F1-macro score of 0.8975. This baseline is indeed a hard baseline and difficult to beat.

In the second run, the ensembled model demonstrated improved reliability, with the MCC rising to 0.7872. The F1-macro score also increased to 0.8917, indicating enhanced overall performance. The F1 scores for conspiracy and critical texts were 0.8536 and 0.9297, respectively, showing more balanced and accurate classifications. Despite these improvements, the model still did not surpass the baseline, highlighting the baseline's strong performance and the challenges in achieving higher accuracy.

5.1.2. Spanish Results task 1

For the Spanish dataset, the first run showed moderate performance with an MCC of 0.6147, indicating a need for further improvement. The F1-macro score was 0.795, reflecting decent overall performance but highlighting areas for enhancement. The model struggled more with conspiratorial texts, achieving an F1 score of 0.7179 for conspiracy versus 0.872 for critical texts, likely due to the specific linguistic challenges presented by the Spanish language. The baseline for Spanish had an MCC of 0.6681 and an F1-macro score of 0.8339, indicating that the baseline was also strong for this language.

In the second run, there was a noticeable improvement in performance. The MCC increased to 0.6722, and the F1-macro score rose to 0.8293, indicating better overall performance. The F1 scores for conspiracy and critical texts improved significantly to 0.7699 and 0.8887, respectively. These improvements suggest that the ensembled model became more adept at handling linguistic features specific to Spanish, leading to more balanced and accurate classifications. This second run beat the strong baseline performance, reflecting significant progress.

5.2. Subtask 2: Span-level Detection of Narrative Elements

5.2.1. English Results task 2

In the first run for the English dataset, the model achieved a span-P of 0.5832 and a span-R of 0.6856, indicating moderate precision but better recall. The span-F1 score was 0.6293, and the micro-span-F1 score was 0.6074, suggesting a balanced performance with a need for improvement in precision. The baseline for this task had a span-F1 score of 0.5323 and a micro-span-F1 of 0.4998, showing that our model performed significantly better than the strong baseline.

In the second run, there was a slight improvement in precision, with a span-P of 0.5859 and a span-R of 0.679. The span-F1 score remained relatively consistent at 0.6279, and the micro-span-F1 score increased marginally to 0.6120. These modest enhancements reflect steady progress in performance and show that our model maintained competitive performance with the baseline.

5.2.2. Spanish Results task 2

For the Spanish dataset, the first run had a span-P of 0.5997 and a span-R of 0.6193. The span-F1 score was 0.6089, and the micro-span-F1 score was 0.6051. The model performed significantly better at span detection in Spanish compared to English, possibly due to distinct narrative markers in the language. The baseline for this task had a span-F1 score of 0.4934 and a micro-span-F1 of 0.4952, indicating that our model significantly outperformed the baseline in both metrics. In the second run, the performance improved, with a span-P of 0.6159, a span-R of 0.6129, and a span-F1 of 0.6129, maintaining a significant advantage over the baseline.

Overall, retraining the model on the entire dataset for one epoch, using the augmented dataset without a validation and test set, resulted in improved performance during the second run.

5.3. Error analysis

In the future, we will thoroughly investigate why we achieved the best results in the competition in the second, more challenging task, but not in the first, theoretically easier one. By examining the confusion matrices for binary classification, we can pinpoint where the model's predictions deviate from actual categories. For instance, in the Spanish dataset, the model misclassified "CRITICAL" texts as "CONSPIRACY" with a 5.52% error rate and "CONSPIRACY" texts as "CRITICAL" with a 31.4% error rate. In the English dataset, these error rates were 4.12% and 19.7%, respectively, indicating greater difficulty in detecting "CONSPIRACY" texts, especially in the Spanish dataset. Misclassifications may arise from language ambiguity, training data limitations, and cultural nuances. For token classification, both datasets exhibited significant issues with spans that should not be annotated, often being misclassified into various categories. This could be due to model overconfidence, lack of enough training data, or to the subjective nature of the task.

6. Conclusion and Future work

The detailed analysis of the results highlights the significant progress made in distinguishing between conspiratorial and critical thinking narratives and detecting narrative elements within texts. The continuous improvements observed between runs emphasize the importance of data augmentation, model fine-tuning, and language-specific adaptations.

Future work should focus on refining these models further, exploring advanced augmentation techniques, and incorporating diverse datasets to improve generalizability. Additionally, developing specialized tools for intergroup conflict analysis and enhancing content moderation strategies will be crucial for addressing the challenges posed by misinformation and fostering a healthier information environment.

Acknowledgements

The work of Mariona Coll Ardanuy is funded by the research project FairTransNLP, grant PID2021-124361OB-C31, funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe.

References

[1] 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. García Seco de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[2] J. Bevendorff, X. Bonet-Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative AI authorship verification, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[3] A. A. Ayele, N. Babakov, J. Bevendorff, X. Bonet-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.

[4] J. Tiedemann, S. Thottingal, OPUS-MT — Building open translation services for the World, in: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 2020.

[5] 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, CoRR abs/1907.11692 (2019). arXiv:1907.11692.

[6] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained BERT model and evaluation data, in: Practical ML for Developing Countries Workshop at the International Conference on Learning Representations (ICLR 2020), 2020.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.

[8] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, 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. 3615–3620. doi:10.18653/v1/D19-1371.

[9] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical nlp in spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, 2022, pp. 193–199.

[10] J. de la Rosa, E. G. Ponferrada, M. Romero, P. Villegas, P. González de Prado Salas, M. Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23.

[11] A. Manconi, G. Armano, M. Gnocchi, L. Milanesi, A soft-voting ensemble classifier for detecting patients affected by covid-19, Applied Sciences 12 (2022) 7554. doi:10.3390/app12157554.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems 26 (2013).

[13] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength natural language processing in python, 2020. URL: 10.5281/zenodo.1212303.

[14] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, MarIA: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). doi:10.26342/2022-68-3.

[15] T. Fontanari, T. C. Fróes, M. Recamonde-Mendoza, Cross-validation strategies for balanced and

imbalanced datasets, in: J. C. Xavier-Junior, R. A. Rios (Eds.), Intelligent Systems. BRACIS 2022. Lecture Notes in Computer Science, volume 13653 of *Lecture Notes in Computer Science*, Springer, Cham, 2022. doi:10.1007/978-3-031-21686-2_43.

[16] 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.

[17] 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.

[18] K.-H. Thung, C.-Y. Wee, A brief review on multi-task learning, Multimedia Tools and Applications 77 (2018) 29705–29725.