Detection of conspiracy-related messages in Telegram with anonymized named entities

Notebook for the PAN Lab at CLEF 2024

Juan Gómez-Romero[1,*], Santiago González-Silot[2], Andrés Montoro-Montarroso[3], Miguel Molina-Solana[1] and Eugenio Martínez Cámara[2]

[1]*Universidad de Granada, Daniel Saucedo Aranda s/n, 18014, Granada, Spain*
[2]*Centro de Estudios Avanzados en TIC, Universidad de Jaén, Campus Las Lagunillas s/n, 23007, Jaén, Spain*
[3]*Universidad de Castilla-La Mancha, Paseo de la Universidad, 4, 13071, Ciudad Real, Spain*

Abstract

This paper investigates the detection of conspiracy-related messages on Telegram within the PAN 2024 task on *oppositional thinking analysis*. The proposed approach aims to improve model generalization and reduce bias by anonymizing named entities during preprocessing. Two binary text classification models for Spanish and English were developed using sentence embeddings and feed-forward neural networks trained on an 8,000-message dataset (4,000 messages per language). Then, two modified models were trained with the same neural network architecture but with named entities replaced by type placeholders. Performance metrics showed that the modified models were competitive with other submissions, achieving MCC scores of 0.797 for English and 0.672 for Spanish.

Keywords

Natural Language Processing, Text Classification, Conspiracy Theories, Named Entity Recognition, Pseudo-Anonymization

1. Introduction

The task on *oppositional thinking analysis* [1] in PAN 2024 [2] focuses on the differentiation between conspiracy theories and critical thinking. Conspiracy theories often attribute significant events to covert, malevolent groups, whereas critical thinking involves scrutinizing established narratives without implying mal-intent. Specifically, this report describes our work in subtask 1, which is formulated as a binary text classification problem. The dataset for this task includes 8,000 messages extracted from Telegram labelled with CONSPIRACY or CRITICAL, 4,000 in English and 4,000 in Spanish. More details about the annotation procedures and the structure of the dataset can be found at [3].

Our previous work on disinformation detection has revealed that text classification methods are often not extensive enough to cover different contexts [4]. Furthermore, through the lens

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
jgomez@ugr.es (J. Gómez-Romero); sgs00034@red.ujaen.es (S. González-Silot); andres.montoro@uclm.es (A. Montoro-Montarroso); miguelmolina@ugr.es (M. Molina-Solana); emcamara@ujaen.es (E. M. Cámara)
0000-0003-0439-3692 (J. Gómez-Romero); 0000-0001-8378-5840 (S. González-Silot); 0000-0003-1893-3346 (A. Montoro-Montarroso); 0000-0001-5688-2039 (M. Molina-Solana); 0000-0002-5279-8355 (E. M. Cámara)

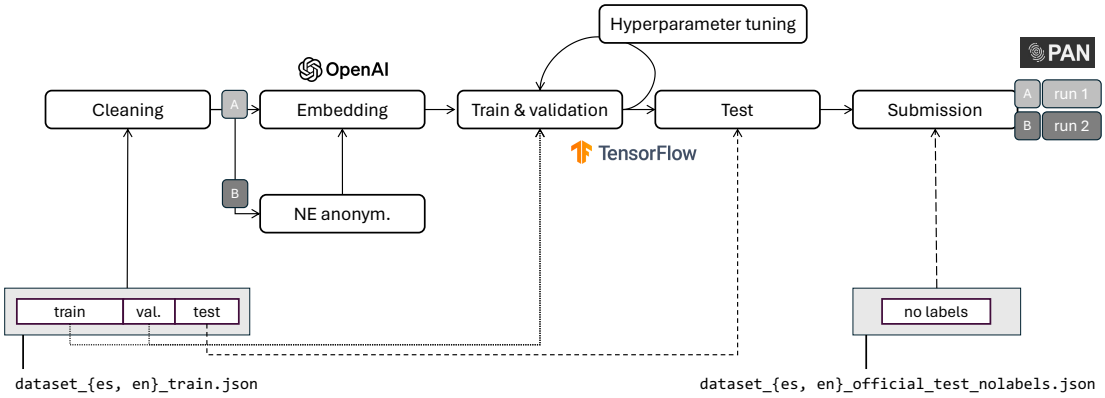


Figure 1: Methodology of the proposal to address subtask 1.

of explainable artificial intelligence [5], we have found that named entities (NE) are usually overrepresented in the disinformation categories and serve as a significant bias factor [6].

This study proposes replacing specific entities with generic-type placeholders during preprocessing. Although we refer to this procedure as anonymization, its primary purpose is not to keep entities unrecognizable but to improve model generalization, reduce bias, and decrease vulnerability to adversarial attacks. We evaluate the effectiveness of this approach by comparing the performance of models trained with and without named entity anonymization using the provided dataset and after submission. The results obtained in the task show that classifiers trained with anonymized named entities can compete with similar approaches, ranking 15th (English, $MCC = 0.797$) and 12th (Spanish, $MCC = 0.672$).

The remainder of the paper describes the details of the methodology and the models used in our submission to subtask 1, tagged as *sail*. The final models and test scripts are publicly available[1].

2. Method

To address subtask 1, we firstly developed two binary classifiers for the English and Spanish datasets (Figure 1, **A**), leading to *run 1*. Both employ sentence embeddings obtained with the OpenAI API[2]. The classification models are feed-forward neural networks (FFN). Data preprocessing involved the removal of URLs and emojis. Afterwards, we developed the approach with named entity anonymization (Figure 1, **B**), also for the Spanish and English datasets. Specifically, named entities of the types location, organization, geopolitical and person were replaced by placeholders of the form `<TYPE LABEL>`, e.g. `AstraZeneca` was replaced by `<ORG>`. The rest of the stages of the pipeline remained unchanged. This was our *run 2*, the one that is finally included in the ranking.

[1] https://github.com/ugr-sail/pan2024-oppositional-subtask1
[2] https://platform.openai.com/docs/guides/embeddings

2.1. Data preprocessing

The preparation of the data consisted of three steps: cleaning, named entity anonymization and embedding calculation. We subsequently detail each of these three steps.

Cleaning: Data cleaning involved the removal of URLs and emojis using `spacy`[3] pipelines, namely `en_core_web_lg` for English and `es_core_news_lg` for Spanish. Our previous work revealed that retaining these elements typically results in better classification metrics, but the resulting model is less capable of generalizing [6].

Named entity anonymization: We used the named entity recognition (NER) method of `spacy` with the model `en_core_web_lg` for English and the model `es_core_news_lg` for Spanish. The total number of entities in each dataset is larger than 25,000, distributed into entity types as shown in Table 1. The codification of the types is different in Spanish and English, e.g., PER (SP) vs PERSON (EN) or GPE (geopolitical entity) + LOC (EN) vs LOC (SP).[4]

Entity Type	#English dataset	#Spanish dataset
ORG ✓	8155	7537
PERSON ✓	5183	0
CARDINAL	4074	0
DATE	3906	0
GPE ✓	3544	0
NORP	1455	0
WORK_OF_ART	541	0
PERCENT	539	0
ORDINAL	426	0
TIME	311	0
MONEY	289	0
EVENT	233	0
LOC ✓	215	8223
FAC	198	0
PRODUCT	186	0
LAW	108	0
QUANTITY	53	0
LANGUAGE	27	0
MISC	0	17779
PER ✓	0	4714
Total	**28442**	**38253**

Table 1
Entity counts, ✓means that entities of this type are replaced.

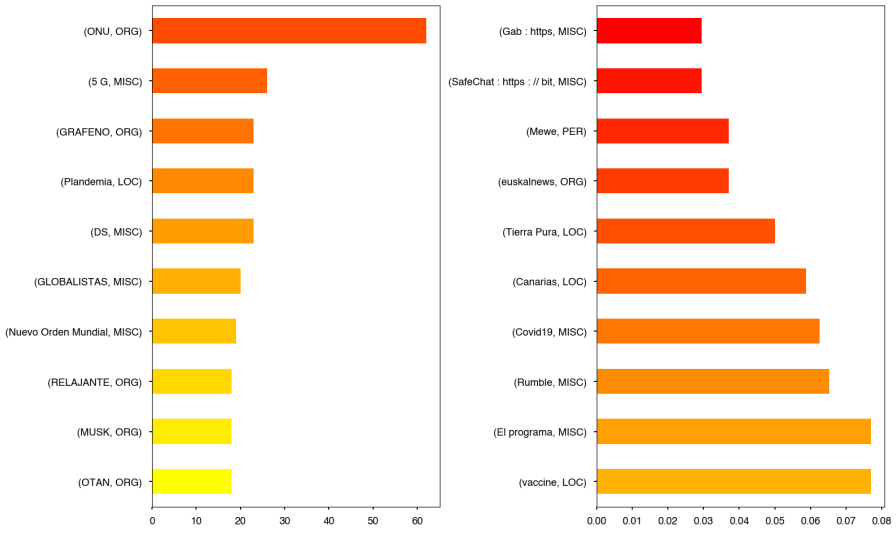
The overrepresentation of certain entities in a specific target category is illustrated in Figure 2. We define the *disparity* ratio of a named entity E as its relative frequency of occurrence in the

[3]https://spacy.io

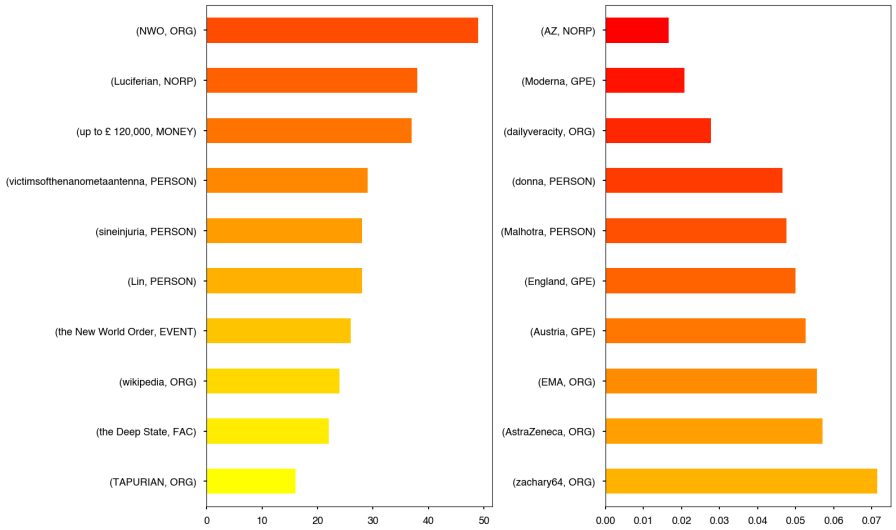
[4]Although the Spanish model recognizes less entity types than the English one, we prefer to keep the homogeneity of the approach and select the appropriate types to anonymize in each case.

CONSPIRACY category compared to the CRITICAL category. Accordingly, the entities with the highest and lowest disparity values mean that they are overrepresented. The figure also shows how many of these entities are not correctly identified or assigned to their type. However, we did not improve the named entity recognition component further.

$$\text{disparity}(E) = \frac{N_{\text{CONSPIRACY}}(\text{entity})}{N_{\text{CRITICAL}}(\text{entity}) + 1} \tag{1}$$



(a) Spanish dataset.



(b) English dataset.

Figure 2: Disparity ratio showing the most overrepresented entities in each classification category. Left: more frequent in CONSPIRACY, Right: more frequent in CRITICAL. Red shades mean more overrepresented.

We observe that certain entities, such as ONU, 5G, and NWO, are more frequent in the CONSPIRACY category. This suggests that these words could significantly aid the model in correctly classifying texts. However, relying on these words would make the model less general and more biased; for instance, any new sentence containing these words would likely be classified as CONSPIRACY. Moreover, this reliance would increase the model's sensitivity to adversarial attacks [7]. We leave a more comprehensive analysis of the impact of specific tokens on the model's results for future work.

Consequently, we decided to replace all the entities of the following types with a placeholder <ENTITY TYPE>:

- Spanish: ORG, PER, LOC.
- English: ORG, PERSON, GPE, LOC.

Calculation of embeddings: We explored two multilingual sentence embedding models from OpenAI, namely `text-embedding-3-large` and `text-embedding-ada-002`. The resulting vectors' dimensions were neither reduced from their original size, respectively 3072 and 1536, nor scaled or normalized. As described in Section 3, `text-embedding-3-large` performed in general better for the classification tasks. From the projection of the embeddings depicted in Figure 3, it appears that the classification tasks can be effectively solved in English. However, in Spanish, the task seems more challenging, and there is no significant difference between the embeddings of the original and anonymized datasets.

2.2. Model

The classification model was a feed-forward neural network with five hidden layers of sizes $\{512, 256, 128, 64, 32\}$. Dropout of 0.4 was enabled after each layer. The input is adapted to the size of the embedding vector, and the output is two values corresponding to each class label with *softmax* activation.

For comparison purposes, we also trained a random forest classifier with `sklearn`[5] and a ridge regressor with `pyCaret`[6]. These methods offer worse performance than the neural network in most cases, although it should be taken into account that we used default hyperparameter values.

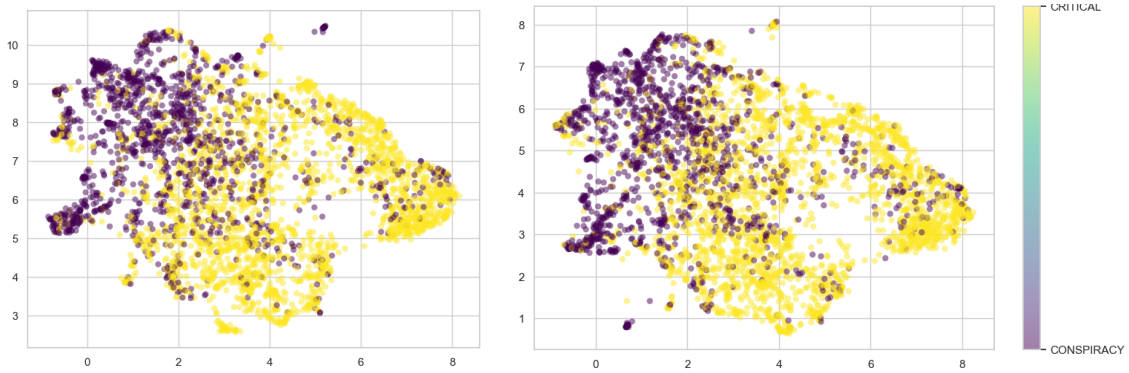
2.3. Train and validation

The labelled dataset was partitioned into training (60%), validation (15%), and test (25%) splits. The neural network model was trained for 10 epochs. The metric used to select the best configuration was the Matthews Correlation Coefficient (MCC).[7] The convergence of the two models (Spanish and English) with NE anonymization is illustrated in Figure 4. However, selecting different batch sizes, optimizers, and loss functions did not significantly impact the results. The

[5] https://scikit-learn.org

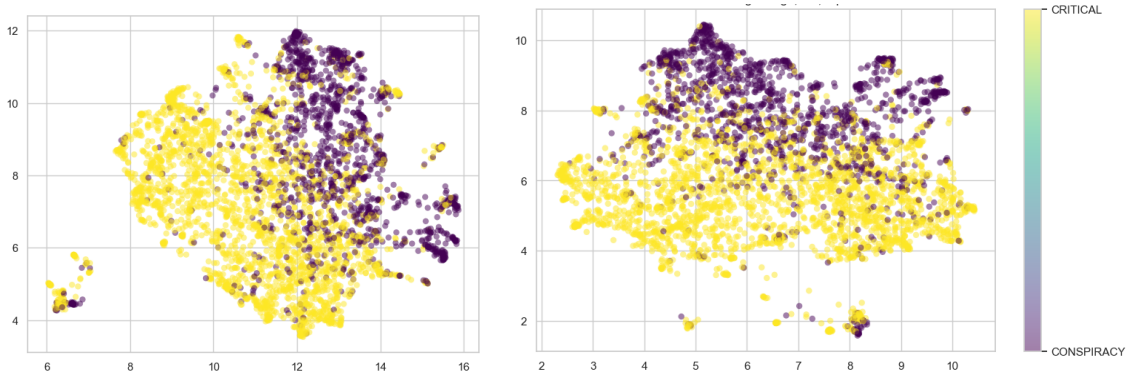
[6] https://pycaret.readthedocs.io

[7] The Matthews Correlation Coefficient (MCC) [8] measures the quality of binary classifications, considering true and false positives and negatives. It returns a value between -1 (total disagreement) and +1 (perfect prediction), making it useful for imbalanced classes.



(a) Spanish dataset, no NE anonymization.

(b) Spanish dataset, NE anonymization.



(c) English dataset, no NE anonymization.

(d) English dataset, NE anonymization.

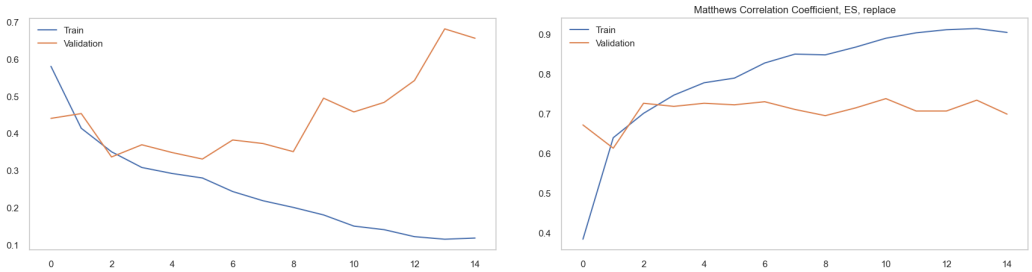
Figure 3: 2-D Projection using UMAP of the messages' embeddings with `text-embedding-3-large`.

final model used a batch size of 32, the `adam` optimizer, and the `categorical_crossentropy` loss function.

3. Results

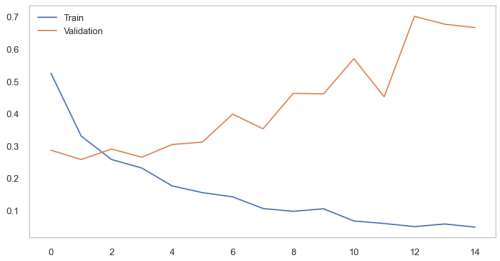
Test: The trained models were evaluated using the 25% test split before preparing the submission. Figure 5 reveals that the combination of `text-embedding-3-large` and feedforward networks typically performs the best. Also, the results with the Spanish dataset before and after embedding does not change very much —as one can expect from the embeddings projections of Figure 3. Therefore, for simplicity, we used this combination of `text-embedding-3-large` embeddings and FFN model for the final submissions. Note that the baseline depicted in the image provided by the competition organizers (red line) is calculated using cross-validation on a model trained with the complete dataset.

Submission: The final model used for the submission was trained with the complete dataset using the best configuration found after tuning with the validation split. We used early stopping

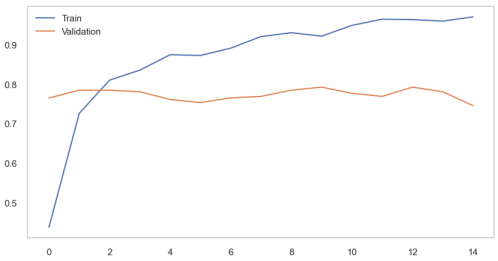


(a) Spanish dataset, loss.

(b) Spanish dataset, MCC.



(c) English dataset, loss.



(d) English dataset, MCC.

Figure 4: Train and validation of the FNN model with `text-embedding-3-large` and named entity anonymization.

applied after 3 consecutive epochs of increasing validation loss. The submission results are shown in Tables 2 and 3.

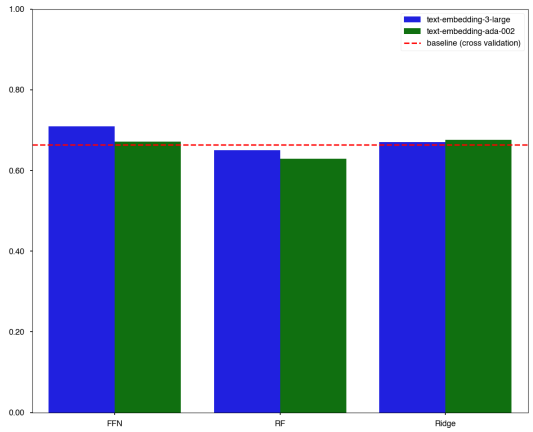
Model	MCC	F1-MACRO	F1-CONSPIRACY	F1-CRITICAL
Regular (A)	**0,676**	**0,837**	**0,798**	0,877
With NE anonymization (B)	0,672	0,830	0,771	**0,888**
BERT baseline	0,668	0,834	0,787	0,881
Best (SINAI)	0,743	0,871	0,832	0,909

Table 2
Submission results of the FNN final model with the Spanish dataset compared to the baseline and the best one.

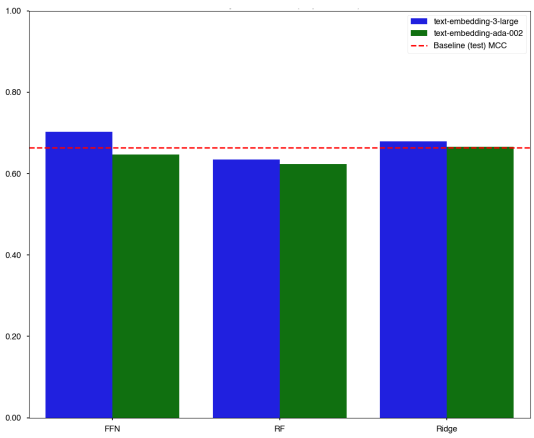
While there are no significant differences in the results with NE anonymization, the generalization capabilities of such models are improved. Let us consider the following message:

> *The concept of the New World Order (NWO) has been a subject of much debate and speculation. However, it is important to approach this topic with a rational perspective. One key criticism of the NWO is the potential for centralized power to undermine democratic principles and individual freedoms.*

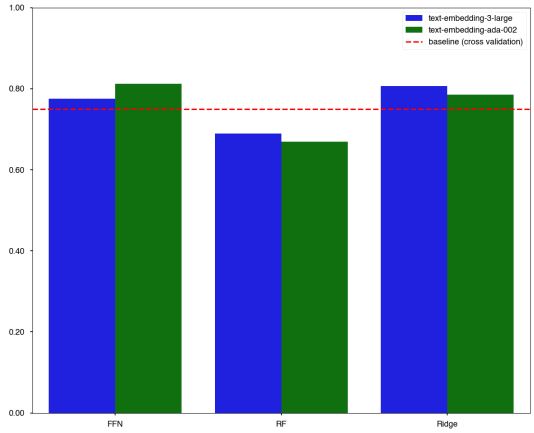
This text is not included in the datasets and can be clearly identified as CRITICAL. However, the regular model without NE anonymization classifies the text as CONSPIRACY with value



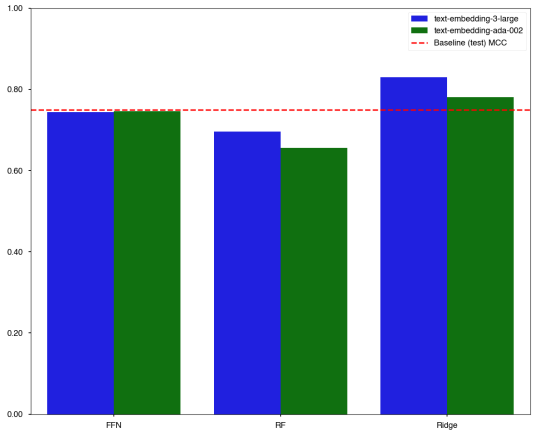
(a) Spanish dataset, no NE anonymization.



(b) Spanish dataset, NE anonymization.



(c) English dataset, no NE anonymization.



(d) English dataset, NE anonymization.

Figure 5: MCC of the models with the test split (20 %). In each pane, from left to right, FFN (feed-forward neural network), RF (random forest), Ridge (ridge regression). Average values after 5 runs.

0.999. In contrast, if we replace the entities of this text (`New World Order` and `NWO`, both with high disparity ratio), the model with NE replacement classifies the text as CRITICAL with value 0.609.

Similarly, the following text, including several entities overrepresented in the category CRITICAL, is classified by the Spanish model without anonymization as CRITICAL with 0.543 (wrong). Interestingly enough, the output of the Spanish model with anonymization is CONSPIRACY with value 0.538 (right).

> *En Canarias se está produciendo un golpe de estado encubierto, reporta Mewe para euskalnews.*

Model	MCC	F1-MACRO	F1-CONSPIRACY	F1-CRITICAL
Regular (A)	0,736	0,862	0,810	0,914
With NE anonymization (B)	**0,797**	**0,898**	**0,869**	0,927
BERT baseline	0,796	0,898	0,863	**0,932**
Best (IUCL)	0,839	0,919	0,895	0,944

Table 3
Submission results of the FNN final model with the English dataset compared to the baseline and the best one.

4. Conclusions and future work

This study shows that replacing named entities with generic placeholders to classify conspirative and critical messages can enhance model's generalization capabilities and reduce bias without significant performance decreases. The results indicate that the English dataset's classification was more positively affected by such named entity anonymization. Some examples are provided to illustrate the changes in classification results, but a more extensive evaluation is required. The extended preprocessing performed in this study could be applied to similar datasets, not only in the context of automatic disinformation detection, to improve model generalization and mitigate bias.

Future work will focus on integrating automatic hyperparameter optimization methods and improving the named entity recognition and replacement process, particularly in Spanish, to enhance model performance. Additionally, we will perform a more comprehensive study of the impact of anonymizing only a subset of entities and apply explainability methods to quantify the impact of these entities on the models' outcomes. In the longer term, we also plan to investigate the role of embedding models, considering that the OpenAI embeddings used here may already employ some form of anonymization, and explore the potential of fine-tuning embeddings post-anonymization. Another interesting direction is the development of multilingual models to avoid having separate ones for different languages.

Acknowledgments

This publication is part of the projects XAI-DISINFODEMICS (PLEC2021-007681) funded by MICIU/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR, FedDAP (PID2020-116118GA-I00) funded by MCIN/AEI/10.13039/501100011033, and SAFER (PID2019-104735RB-C42) funded by MICIU/AEI/10.13039/501100011033.

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] 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.

[3] D. Korenčić, B. Chulvi, X. Bonet Casals, M. Taulé, P. Rosso, PAN24 oppositional thinking analysis, 2024. doi:10.5281/ZENODO.10680586.

[4] A. Montoro-Montarroso, J. Cantón-Correa, P. Rosso, B. Chulvi, A. Panizo-Lledot, J. Huertas-Tato, B. Calvo-Figueras, M. J. Rementeria, J. Gómez-Romero, Fighting disinformation with artificial intelligence: fundamentals, advances and challenges, El Profesional de la información 32 (2023) e320322. doi:10.3145/epi.2023.may.22.

[5] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115. doi:10.1016/j.inffus.2019.12.012.

[6] S. González-Silot, Procesamiento de Lenguaje Natural Explicable para Análisis de Desinformación, Master's thesis, Universidad de Granada, 2023.

[7] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, C. Li, Adversarial attacks on deep-learning models in natural language processing: A survey, ACM Transactions on Intelligent Systems and Technology (TIST) 11 (2020) 1–41.

[8] B. W. Matthews, Comparison of the predicted and observed secondary structure of t4 phage lysozyme, Biochimica et Biophysica Acta (BBA)-Protein Structure 405 (1975) 442–451. doi:10.1016/0005-2795(75)90109-9.