Small Language Models and Large Language Models in Oppositional Thinking Analysis: Capabilities, Biases and Challenges

Notebook for PAN at CLEF 2024

Álvaro Huertas-García[1,2], Carlos Martí-González[2], Javier Muñoz[2] and
Enrique De Miguel Ambite[2]

[1]Department of Computer System Engineering, Polytechnic University of Madrid, Calle de Alan Turing, 28031, Madrid, Spain
[2]Fundación Tecnológica Advantx – Funditec, Paseo de la Castellana, 28046 , Madrid, Spain

Abstract

The proliferation of misinformation and conspiracy theories needs robust methods to differentiate legitimate critical discourse from harmful conspiratorial narratives. This study investigates discerning critical messages from conspiracy theories within COVID-19 discussions on Telegram. Preserving information integrity on social media impacts vital public discourse on health, politics, and science.

The research employs two distinct approaches: linguistic style classification and contextual knowledge classification. The former leverages a diverse ensemble of Small Language Models (SLMs), Large Language Models (LLMs), and State-Space Models (SSMs), while the latter harnesses the capabilities of the Claude 2.0 Opus model for contextual analysis.

Empirical evaluations demonstrate that the SLM models using Matryoshka embedding and Mamba (SSM) models exhibit superior performance for the English language dataset, achieving a Matthews Correlation Coefficient (MCC) of 0.793. For the Spanish dataset, the Spanish BERT baseline (SLM) attains an MCC of 0.699. Notably, a multilingual model trained on a balanced combination of English and Spanish data outperforms its monolingual counterparts, with the multilingual-e5-large model (LLM) achieving an MCC of 0.768 for English and 0.725 for Spanish. This finding underscores the potential of multilingual models to mitigate the ”curse of multilinguality,” where performance often degrades on low-resource languages. However, the suboptimal performance of the Claude 2.0 Opus model, exhibiting a tendency to classify texts as conspiracy-related, highlights inherent biases that require further investigation.

Overall, this study contributes to the development of advanced models that can effectively differentiate critical thinking from conspiratorial narratives in various linguistic contexts. Future research should prioritize identifying and addressing biases in large language models to ensure fair treatment of diverse perspectives, as well as to preserve freedom of expression and ensure fair representation of narratives.

Keywords

PAN 2024, Oppositional Thinking Analysis, Transformers, Mamba, LLM, Claude, Bias

1. Introduction

The proliferation of misinformation and conspiracy theories has become a significant challenge in today’s digital age, impacting vital aspects of public discourse such as health, politics, and scientific discourse [1]. Conspiracy beliefs can shape human behaviour and decision-making processes, making understanding the cognitive styles and personality traits associated with such beliefs is crucial. Extensive psychological research has identified numerous predictors of conspiracy beliefs, including personality factors like low agreeableness and high openness to experience [1]. Moreover, studies on cognitive styles have revealed a correlation between belief in conspiracy theories and lower analytic thinking coupled with higher intuitive thinking [2].

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

ahuertas@funditec.es (Á. Huertas-García)
https://github.com/Huertas97 (Á. Huertas-García)
0000-0003-2165-0144 (Á. Huertas-García); 0009-0003-1387-1630 (C. Martí-González); 0000-0001-5068-3303 (J. Muñoz); 0009-0008-1026-2537 (E. D. M. Ambite)

Understanding these human aspects of conspiracy beliefs is not merely an academic exercise; it has far-reaching implications. This knowledge can inform behavioural interventions to mitigate the spread of misinformation [3]. Furthermore, integrating cognitive and personality factors into natural language processing (NLP) models can enhance their accuracy in distinguishing between critical and conspiratorial narratives, ultimately improving their performance and reliability.

In the realm of automatic content moderation, the challenge of distinguishing between conspiracy theories and critical thinking in NLP models has emerged as a vital area of study. The prevalence of conspiratorial content has escalated the need for robust methodologies that can accurately differentiate between legitimate critical discourse and harmful conspiracy narratives. Maintaining the integrity of information shared across social media platforms and other digital forums is crucial for preserving the credibility of public discourse.

2. Background

While the focus on this topic remains relatively limited, related studies provide valuable insights into methodologies and applications in adjacent areas. For instance, a significant contribution by [4] presents a framework for detecting conspiracy theories on Twitter using a novel recurrent model called BORJIS, highlighting the efficacy and challenges of NLP techniques in identifying that conspiratorial content is often found within vast amounts of social media data. Similarly, [5] explored fake news detection related to COVID-19 and 5G conspiracy theories using BERT embeddings and Graph Neural Networks, showcasing advanced NLP techniques for distinguishing misinformation from legitimate critical analysis.

Other studies follow a different approach focusing on tracking the spread across social networks, such as the FacTeR-Check semi-automated fact-checking tool that uses semantic similarity and natural language inference (NLI) to monitor the evolution of misinformation or disinformation on online social networks [3]. Additionally, the use of camouflage for content evasion has also been reported, and works have developed multilingual NER NLP models to counter these strategies, like the ”pyleetspeak” tool for simulating word camouflage and a NER Transformer model for its detection [6].

Furthermore, the research conducted by [7] explores the potential of NLP techniques in fostering critical thinking skills within educational settings. It offers valuable insights into the systematic instruction and assessment of critical thinking, specifically in comparison to conspiratorial thinking. Finally, [8] emphasize the importance of explicit theorization in developing models that can accurately differentiate between critical and conspiratorial thinking in their paper on gender bias in NLP research.

While significant progress has been made in the field, there is still much to explore in analyzing oppositional thinking using NLP. This research article addresses this issue in both English and Spanish, contributing to the development of more sophisticated NLP systems for real-world scenarios.

2.1. Competition Description

The competition, titled “Oppositional Thinking Analysis: Conspiracy vs Critical Narratives” is part of the PAN at CLEF 2024 event [9, 10]. Our focus is on the first subtask, which involves analyzing texts from the Telegram platform related to the COVID-19 pandemic. The objective is to perform a binary classification to differentiate between two types of narratives:

- Critical comment: Messages that question major decisions in the public health domain without promoting a conspiracist mentality. These are critical opinions based on information that may not be commonly accepted but do not imply secret plots or malevolent intentions.
- Conspiracy comment: Messages that portray the pandemic or public health decisions as results of malevolent conspiracies by secret, influential groups. These messages often encourage distrust based on unverified or poorly explained evidence.

The official evaluation metric for this subtask is the Matthews Correlation Coefficient (MCC). MCC is a measure of the quality of binary classifications, providing a balanced evaluation even when the

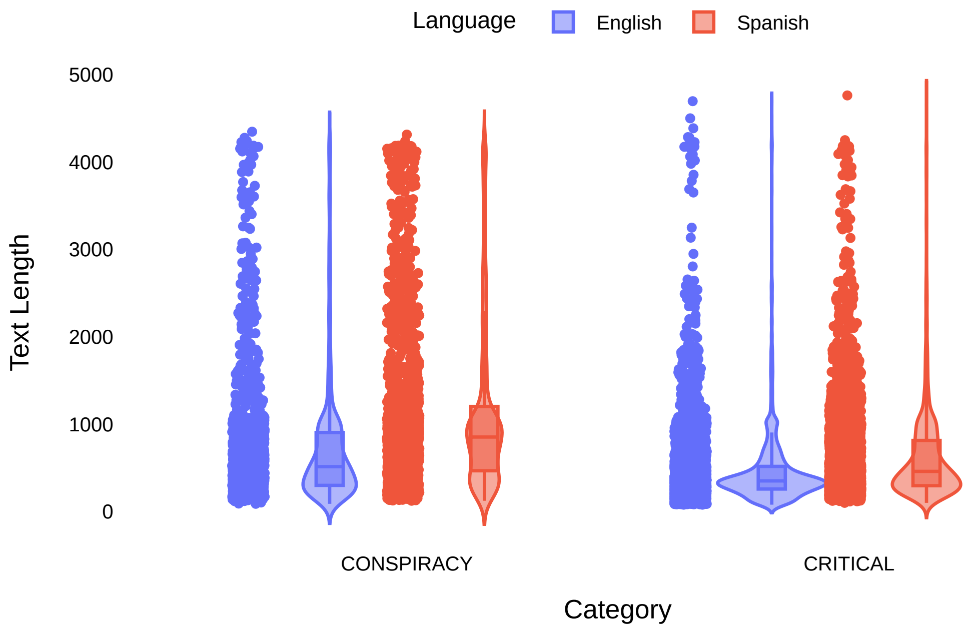

Figure 1: Distribution of Text Length in Training Set by Category and Language

classes are of very different sizes. It is normalized, making it applicable to other datasets and ensuring robust performance assessment across diverse scenarios.

3. Methodology

Our methodology is based on two main approaches: one that classifies texts according to their linguistic style and content, which we refer to as the **Linguistic Style Classification Approach**; and another that uses the input text, combined with contextual knowledge and reasoning from large language models (LLMs), referred to as the **Contextual Knowledge Classification Approach**.

3.1. Linguistic Style Classification Approach

3.1.1. Dataset Preprocessing

The dataset comprises texts in both English and Spanish, categorized into CRITICAL and CONSPIRACY narratives. The English dataset consists of 2,621 CRITICAL texts and 1,379 CONSPIRACY texts. The Spanish dataset includes 2,538 CRITICAL texts and 1,462 CONSPIRACY texts. Both datasets were divided into training (80%) and validation (20%) sets using a random seed of 42.

The preprocessing involved analyzing the prevalence of URLs, emojis, and text length distributions. URLs were removed to standardize the text data. The text length distributions for the English dataset were found to be 743±740 characters for CONSPIRACY and 476±479 characters for CRITICAL. For the Spanish dataset, the distributions were 1112±946 characters for CONSPIRACY and 641±577 characters for CRITICAL. These distributions are illustrated in Figure 1.

3.1.2. Models

We employed a diverse range of models, both monolingual and multilingual. Except for Mamba, all models are based on the Transformer architecture. The significance of Transformer models lies in their attention mechanism, which allows them to efficiently handle dependencies in long sequences and capture intricate patterns within the data. According to Vaswani et al.[11], the self-attention mechanism of Transformers enables them to dynamically weigh the importance of different tokens in a sequence, making them highly effective for various NLP tasks. Mamba, in contrast, is an advanced state-space model (SSM) designed for efficient handling of complex sequences with large datasets, as detailed by Gu and Dao[12].

Below, we list the models used in our research along with brief descriptions:

Monolingual

- **BERT-base-uncased**[13]: A foundational model that effectively applies Transformers at scale, expanding our understanding of linguistic context. We selected the largest variant to ensure a comprehensive analysis and to compare historical model design evolution.
- **DistilBERT**[1]: A compact version of BERT by Hugging Face, offering a smaller and faster alternative while maintaining similar performance. Suitable for various NLP tasks.
- **Nomic**: Nomic Embed [14] innovates in embedding techniques to provide dynamic, context-aware representations, surpassing leading models as of February 2024. With a compact size, low memory usage, and advanced training methods, Nomic Embed efficiently processes up to 8192 tokens, making it ideal for analyzing extensive online materials.
- **DistilRoBERTa**[15]: A faster and smaller version of RoBERTa, trained on the same corpus in a self-supervised manner using BERT as a teacher.
- **twitter-roberta-base-sentiment-latest**[16]: A RoBERTa-base model fine-tuned for sentiment analysis using tweets from January 2018 to December 2021, benchmarked with TweetEval.
- **all-MiniLM-L6-v2**[2]: A Transformer model trained with contrastive loss on 1B sentence pairs to encode sentences and short paragraphs into a dense vector space of 384 dimensions, suitable for tasks like clustering or semantic search.
- **mxbai-embed-large-v1**[17]: A powerful English embedding model known for its efficient size and high performance. Using Matryoshka Embedding [18], it trains hidden layers to generate high-quality embeddings independently of higher layers, reducing both the number of layers and embedding dimensions. Ranked in the top 25 on the MTEB leaderboard[3] for sentence embedding tasks, it outperforms commercial models like OpenAI's text-embedding-3-large, making it a top choice for our research.
- **Mamba**[4] [12]: An advanced state-space model designed for efficient handling of complex sequences with large datasets. It uses a selection mechanism to decide whether to propagate or discard information based on token relevance, providing a viable method for assessing intricate controversial and critical comments on social media.
- **dccuchile/bert-base-spanish-wwm-uncased** [19]: Also known as BETO, this is a BERT model trained on a large Spanish corpus using a vocabulary of about 31k BPE subwords constructed with SentencePiece.

Multilingual

- **XLM-RoBERTa**: A scaled cross-lingual multilingual sentence encoder version of the RoBERTa model, trained on 2.5TB of data across 100 languages filtered from Common Crawl.
- **LLAMA 2**: A family of pre-trained and fine-tuned large language models (LLMs) by Meta AI, useful for various research and commercial purposes.
- **multilingual-e5**: Developed at Microsoft, this sophisticated embedding model excels in tasks requiring robust text representation, such as information retrieval, semantic textual similarity, and text reranking. Initialized from xlm-roberta-large, it is continually trained on a mixture of multilingual datasets, supporting 100 languages from xlm-roberta with potential performance degradation for low-resource languages.

In this style-based strategy, all these models are employed as the encoder body of the texts to which a layer of 1024 classifier neurons is added.

[1] https://huggingface.co/distilbert/distilbert-base-uncased

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[3] https://huggingface.co/spaces/mteb/leaderboard

[4] https://huggingface.co/state-spaces/mamba-370m-hf

Table 1
Hyperparameters Explored in Bayesian Optimization

Parameter	Range/Values	Distribution/Method
Epochs	1 to 4	Uniform
Accumulation Steps	[1, 2, 3, 4]	Discrete
Learning Rate (lr)	1e-6 to 1e-4	Log-Uniform
Weight Decay	0.0 to 1.0	Uniform
Batch Size	[4]	Discrete
Loss Function	[BCEWithLogitsLoss, sigmoid_focal_loss]	Discrete
Sigmoid Focal Loss Alpha	0.25 to 1.0	Uniform
Sigmoid Focal Loss Gamma	1.0 to 4.0	Uniform
Scheduler Name	[PolynomialLR, CosineAnnealingWarmRestarts, LinearLR ConstantLR]	Discrete
Optimizer	[AdamW, Adam]	Discrete

3.1.3. Hyperparameter Tuning, Importance, and Correlation

For hyperparameter tuning, we used Bayesian optimization, which leverages prior evaluations to guide its search process, enhancing model performance. Table 1 lists the explored hyperparameters, including their ranges and sampling distributions.

We analyzed the importance and correlation of hyperparameters with the Matthews Correlation Coefficient (MCC). Correlation measures the linear relationship between hyperparameters and MCC, indicating how changes in hyperparameters affect performance.

Additionally, we calculated an importance metric exploiting the feature importance of a random forest model, based on the idea that the more important features appear more often in the trees of the forest. Hyperparameters served as input features, with MCC as the target output. This provided feature importance values, showing each hyperparameter's contribution to predicting performance. These analyses offer insights into how hyperparameters influence model performance.

The experimental tracking can be consulted in Weight and Biases[5]

3.2. Contextual Knowledge Classification Approach

For this approach, we utilized the Claude 2.0 Opus model for zero-shot classification. This approach relies on prompt engineering using tempeareture equals to 1 and without fine-tuning the model, leveraging the extensive knowledge of language and context up to early 2023. Additionally, its multilingual capability is well-suited to this task, as approximately 10% of the data used was non-English, according to Anthropic. This model was accessed via Anthropic's public API before May 6.

Below is the final prompt used for the Zero-shot classification:

Claude 2.0 Opus Prompt

Your role is to analyze text inputs to identify whether they represent critical commentary or conspiracy theories, each with distinct characteristics:
Critical Commentary:
Definition: Critical messages that question major decisions in the public health domain, but do not promote a conspiracist mentality. It is an opinion, it may not be correct but do not consider that the revendication belongs to a secret or a plot against the population in terms of influential groups. It can be a critic based on information that may not be the common opinion, and it could be wrong, but it is expressing a point of view that another can criticize.
Characteristics: Applicability: Applies even when the topic might be susceptible to conspiratorial interpretations.
Conspiracy Commentary:
Definition: Messages that view the pandemic or public health decisions as a result of a malev-

[5] https://wandb.ai/huertas_97/PAN_2024_Opposing/workspace

olent conspiracy by secret, influential groups. It can be an opinion but the main problem is that it tries to convince you to distrust based on evidences that are not well trusted or explained and leave open the door to be distrustful instead of being critical based on information that may not be the common opinion.

Characteristics:

- Suspicion and Paranoia: Thrives on distrust of official narratives and institutions.
- Simplistic Explanations: Oversimplifies complexities by attributing them to the actions of a few.
- Resistance to Evidence: Dismisses contrary evidence as part of the cover-up.

You are required to utilize web browser research extensively to verify claims and gather context before making your classification. Be sure to adhere strictly to the output format, especially in reporting URLs used in your research to ensure transparency and accountability.

Additional Instruction:

- Always use a web browser to search for information related to the text. Your classification should be informed by credible online sources. Include URLs of these sources in your explanation to validate your findings and reasoning.
- Maintain neutrality in your classification process. Do not classify a text as ”CONSPIRACY” solely because the topic is related to commonly misunderstood or hot-button issues. Instead, use clear evidence from the text and supporting information from web searches to distinguish between critical perspectives and actual conspiracy theories. Include URLs of these sources in your explanation to validate your findings and reasoning.

Task Requirements:

- Classify the narrative of the text based on the categories above.
- Determine the main topics of the text in 2-3 words.
- Assign a Confidence Score from 0 to 1, indicating the certainty of your classification.
- Your explanation must reflect how the information sourced online influenced your classification and must include URLs for verification.

Output Format: (It is crucial that the output strictly follows this format)

```
{
  "Prediction": "CATEGORY_NAME",
  "Confidence": [Confidence Score],
  "Topic": ["topic1", "topic2"],
  "Reason": "A concise explanation based on the
            characteristics with URLs of the sources used."
}
```

It is crucial that the output strictly follows this format.

4. Experiments and Results

4.1. Training and Developing Results

This section presents the results obtained from developing and evaluating various models. As shown in Table 2, the best monolingual model for English are mxbai-embed-large-v1 and Mamba 370m model, achieving an MCC of 0.793, while the best model for Spanish is bert-base-spanish-wwm-uncased, with an MCC of 0.699.

Table 2
Model performance metrics on validation set for English

Language	Model	Accuracy	F1-macro	MCC
EN	**mamba-370m**	**0.908**	**0.861**	**0.793**
	mxbai-embed-large-v1	**0.906**	**0.865**	**0.793**
	Baseline (BERT)	0.905	0.854	0.787
	twitter-roberta-base-sentiment-latest	0.900	0.851	0.776
	DistilRoBERTa	0.896	0.856	0.776
	all-MiniLM-L6-v2_EN	0.896	0.851	0.771
	DistilBERT	0.889	0.829	0.750
	XLM-Roberta	0.869	0.826	0.729
	nomic-embed-text-v1.5	0.870	0.782	0.710
	LLAMA 2	0.803	0.730	0.578
	Claude 2.0 Opus	0.457	0.366	0.236

Table 3
Model performance metrics on validation set for Spanish

Language	Model	Accuracy	F1-macro	MCC
ES	**bert-base-spanish-wwm-uncased**	**0.863**	**0.793**	**0.699**
	multilingual-e5-large	0.861	0.785	0.698
	mamba-370m	0.843	0.770	0.654
	nomic-embed-text-v1.5	0.838	0.763	0.643
	XLM-Roberta	0.829	0.732	0.624
	LLAMA 2	0.705	0.658	0.424

As shown in Table 2, the superior performance of the mxbai-embed-large-v1 model underscores its effectiveness in encoding texts into embeddings. This model employs the Matryoshka Embedding technique [18], where each layer is trained to produce high-quality embeddings independently, thus enhancing the model's overall performance. This approach contrasts sharply with the performance of larger models such as Llama 2 [20], which, despite having over 7 billion parameters, underperforms when a classifier head is added. This observation corroborates the notion that model size does not necessarily correlate with task-specific performance. Optimizing the model architecture to improve linguistic encoding, as demonstrated by mxbai-embed-large-v1, proves more beneficial than merely increasing the number of parameters.

Additionally, the performance of the Mamba 370m model, which matches mxbai-embed-large-v1 with an MCC of 0.793, highlights the potential of alternative architectures beyond Transformers. The Mamba model, with its state-space approach and selective propagation mechanism, presents a compelling case for further exploration of non-Transformer architectures in NLP tasks.

The horrible performance of the large language model Claude 2.0 Opus[6] in English, which is typically a benchmark model for complex reasoning tasks, warrants further investigation. Despite its state-of-the-art status in reasoning datasets, Claude 2.0 Opus showed a tendency to classify texts as conspiracy-related. This bias was evident even when the model provided reasoning for its classifications, suggesting a predisposition influenced by the sensitive nature of the subject matter. This finding highlights the need for ongoing research into model biases and their impact on classification tasks, particularly for topics with significant socio-cultural implications such as conspiracy theories.

In the Spanish dataset (see Table 3, the bert-base-spanish-wwm-uncased model achieved lower performance compared to its English counterparts, indicating potential limitations in the Spanish training data or model architecture. However, when using the multilingual model multilingual-e5-large, which was trained on Spanish data alone, still not surpassing the monolingual model. This suggests that

[6] https://www.anthropic.com/news/claude-2

Table 4
Model performance metrics on validation set for English and Spanish

Language	Model	EN Score			ES Score		
		Accuracy	F1-macro	MCC	Accuracy	F1-macro	MCC
EN-ES	**multilingual-e5-large**	**0.896**	**0.845**	**0.768**	**0.874**	**0.821**	**0.725**
	multilingual-e5-base	0.895	0.850	0.769	0.868	0.819	0.714

in this context, monolingual models might be more effective than multilingual ones just using one language data for training.

Interestingly, when the multilingual model was trained on both English and Spanish datasets, it achieved an MCC of 0.725, as shown in Table 4. This indicates that multilingual models can leverage larger and more diverse datasets to enhance their understanding of the task. The ability of multilingual models to generalize across languages is particularly evident when they are exposed to substantial amounts of well-represented data, demonstrating their potential to exploit linguistic diversity for improved performance.

Overall, we select the following models for the two trial of the competition:

- RUN 1 - Monolingual approach consists of Mamba 370m model for English and bert-base-spanish-wwm-uncased model for Spanish.
- RUN 2 - Multilingual approach consists of multilingual model multilingual-e5-large, trained in both languages together.

4.2. Parameters Tuning, Importance and Correlation

The results of our hyperparameter tuning highlight the significant influence of learning rate (lr) on model performance, with an importance score of 0.531 and a negative correlation of -0.535 with MCC. This suggests that optimizing the learning rate is crucial for achieving higher performance, as inappropriate values can lead to suboptimal results. Runtime also showed considerable importance (0.196) and a positive correlation (0.410), indicating that longer training times generally improve model performance.

Weight decay and sigmoid focal loss parameters, while less influential than the learning rate, still play vital roles. The weight decay parameter had an importance of 0.163 and a negative correlation of -0.323, suggesting that higher weight decay might adversely affect the model. Sigmoid focal loss [21] parameters (alpha and gamma) demonstrated moderate importance, with alpha showing a positive correlation (0.156) and gamma a negative one (-0.325). The focal loss function, designed to address class imbalance, is given by:

$$\mathrm{FL}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t)$$

where p_t is the model's estimated probability for the true class label, α_t is a weighting factor for class imbalance, and γ is a focusing parameter that adjusts the rate at which easy examples are down-weighted. This indicates a complex relationship where these parameters can be fine-tuned to balance the model's sensitivity to class imbalances effectively.

Other parameters, such as epochs, batch size, and accumulation steps, showed lower importance scores. Interestingly, batch size had a positive correlation with MCC (0.287), indicating that larger batch sizes might contribute to better performance. However, the relatively low importance scores for these parameters suggest that while they do influence performance, their impact is less critical compared to the learning rate and regularization parameters.

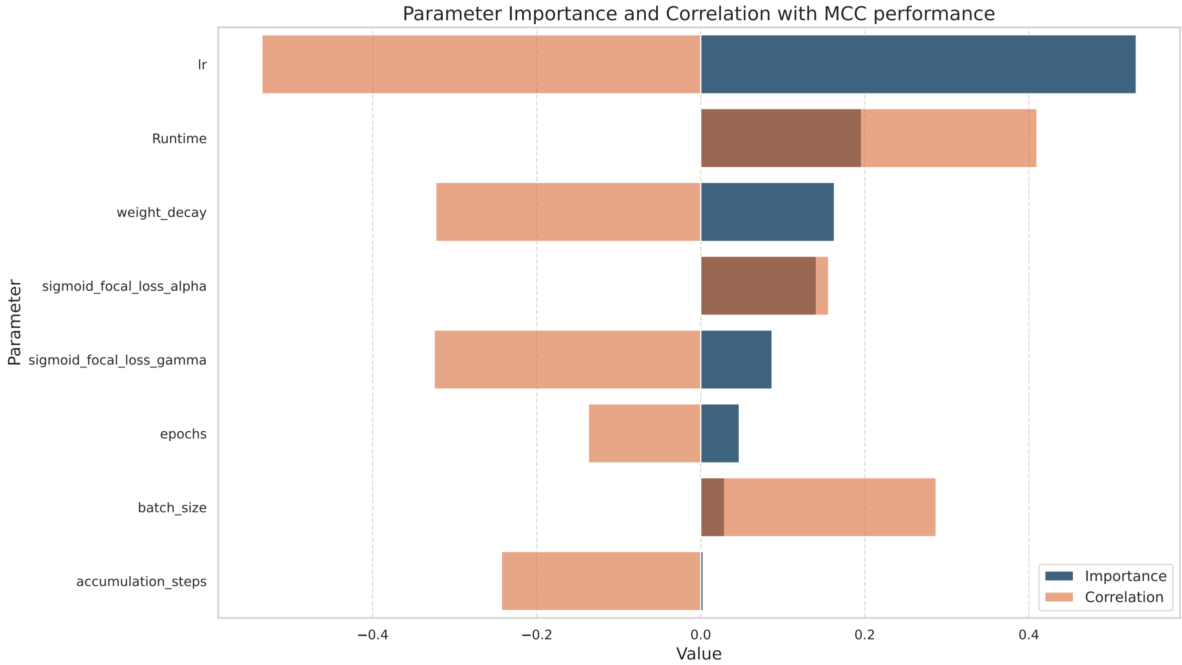

Figure 2: Parameter Importance and Correlation with MCC performance

Table 5
Comparison of Monolingual and Multilingual Approaches

	Run 1 - Monolingual Approach		**Run 2 - Multilingual Approach**	
	EN	ES	EN	ES
MCC	0.7894	0.6445	0.7965	0.7028
F1-macro	0.8947	0.8160	0.8977	0.8497
F1-conspiracy	0.8617	0.7523	0.8637	0.8035
F1-critical	0.9276	0.8796	0.9317	0.8960

5. Official Competition Results and Conclusion

Table 5 presents the test results, demonstrating that Run 2, which employs a single multilingual model, outperforms the monolingual models. Notably, in both languages, the baseline performance of BERT (MCC 0.7964) is exceeded. Out of 82 teams, only 17 surpass this threshold. Particularly striking is the performance in Spanish, where only 13 teams exceed the MCC threshold of 0.6681, placing us in the top three.

These results highlight the potential advantages of multilingual models in achieving robust performance across languages. The improved performance in Run 2 suggests that training a multilingual model on data from both languages mitigates the so-called "curse of multilinguality", where multilingual models often struggle to distribute their knowledge equally across all languages. This phenomenon has been documented in the literature, where multilingual models tend to underperform on low-resource languages due to an imbalance in data distribution and representation [22, 23]. Our findings supports that providing a balanced dataset across languages can significantly enhance the fairness and effectiveness of multilingual models, as other works have applied these to counter content evasion on social media platforms [6, 24].

Furthermore, this study underscores the importance of addressing biases in large language models (LLMs). The bias observed in the Claude 2.0 Opus model, which showed a tendency to classify texts as conspiracy-related, raises critical questions about the ethical deployment of AI technologies. Such biases can have profound implications for freedom of expression and the equitable treatment of diverse perspectives. Future research should focus on developing techniques to identify and mitigate these biases, ensuring that LLMs operate fairly across different socio-cultural contexts.

In conclusion, our findings support the utility of multilingual models in handling diverse linguistic data, provided that training data is well-distributed across languages. The competition has allow us to conduct a research that demonstrate the potential of such models in achieving high performance and also emphasizes the necessity of continuous efforts to address and mitigate inherent biases in AI systems. Moving forward, it is essential to explore advanced methodologies for bias detection and mitigation, which will be crucial for the ethical and effective application of AI technologies in real-world scenarios.

References

[1] A. Goreis, M. Voracek, A Systematic Review and Meta-Analysis of Psychological Research on Conspiracy Beliefs: Field Characteristics, Measurement Instruments, and Associations With Personality Traits, Frontiers in Psychology 10 (2019) 205. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2019.00205/full. doi:10.3389/fpsyg.2019.00205.

[2] B. Gjoneska, Conspiratorial Beliefs and Cognitive Styles: An Integrated Look on Analytic Thinking, Critical Thinking, and Scientific Reasoning in Relation to (Dis)trust in Conspiracy Theories, Frontiers in Psychology 12 (2021) 736838. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.736838/full. doi:10.3389/fpsyg.2021.736838.

[3] A. Martín, J. Huertas-Tato, Álvaro Huertas-García, G. Villar-Rodríguez, D. Camacho, Factercheck: Semi-automated fact-checking through semantic similarity and natural language inference, Knowledge-Based Systems 251 (2022) 109265. doi:https://doi.org/10.1016/j.knosys.2022.109265.

[4] B. A. Galende, G. Hernández-Peñaloza, S. Uribe, F. A. García, Conspiracy or not? a deep learning approach to spot it on twitter, IEEE Access 10 (2022) 38370–38378. doi:10.1109/ACCESS.2022.3165226.

[5] A. Hamid, N. Shiekh, N. Said, K. Ahmad, A. Gul, L. Hassan, A. Al-Fuqaha, Fake news detection in social media using graph neural networks and nlp techniques: A covid-19 use-case, 2020. arXiv:2012.07517.

[6] Á. Huertas-García, A. Martín, J. Huertas-Tato, D. Camacho, Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage, Applied Soft Computing 145 (2023) 110552. doi:https://doi.org/10.1016/j.asoc.2023.110552.

[7] Editor Pje, Nourishing critical thinking skills using neuro-linguistic programming: farah hashmi, PJE 39 (2023). URL: https://ojs.aiou.edu.pk/index.php/pje/article/view/865. doi:10.30971/pje.v39i1.865.

[8] H. Devinney, J. Björklund, H. Björklund, Theories of "gender" in nlp bias research, 2022. arXiv:2205.02526.

[9] J. Bevendorff, X. B. 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.

[10] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, et al., Overview of pan 2024: multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative ai authorship verification, in: European Conference on Information Retrieval, Springer, 2024, pp. 3–10.

[11] 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).

[12] A. Gu, T. Dao, Mamba: Linear-time sequence modeling with selective state spaces, 2024. arXiv:2312.00752.

[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[14] Z. Nussbaum, J. X. Morris, B. Duderstadt, A. Mulyar, Nomic embed: Training a reproducible long context text embedder, arXiv preprint arXiv:2402.01613 (2024).

[15] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[16] J. Camacho-collados, K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. Espinosa Anke, F. Liu, E. Martínez Cámara, TweetNLP: Cutting-edge natural language processing for social media, in: W. Che, E. Shutova (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Abu Dhabi, UAE, 2022, pp. 38–49. doi:10.18653/v1/2022.emnlp-demos.5.

[17] S. Lee, A. Shakir, D. Koenig, J. Lipp, Open Source Strikes Bread - New Fluffy Embedding Model, 2024. URL: https://www.mixedbread.ai/blog/mxbai-embed-large-v1.

[18] A. Kusupati, G. Bhatt, A. Rege, M. Wallingford, A. Sinha, V. Ramanujan, W. Howard-Snyder, K. Chen, S. Kakade, P. Jain, A. Farhadi, Matryoshka representation learning, 2024. arXiv:2205.13147.

[19] 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.

[20] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. arXiv:2307.09288.

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2018. arXiv:1708.02002.

[22] 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.

[23] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, M. Johnson, Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization, 2020. arXiv:2003.11080.

[24] Á. Huertas-García, A. Martín, J. Huertas-Tato, D. Camacho, Countering Misinformation Through Semantic-Aware Multilingual Models, in: H. Yin, D. Camacho, P. Tino, R. Allmendinger, A. J. Tallón-Ballesteros, K. Tang, S.-B. Cho, P. Novais, S. Nascimento (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2021, volume 13113, Springer International Publishing, Cham, 2021, pp. 312–323. doi:10.1007/978-3-030-91608-4_31.