DSVS at PAN 2024: Ensemble Approach of Transformer-based Language Models for Analyzing Conspiracy Theories Against Critical Thinking Narratives

Notebook for PAN at CLEF 2024

Sergio Damián[1,*], Brian Herrera[1], David Vázquez[1], Hiram Calvo[1], Edgardo Felipe-Riverón[1] and Cornelio Yáñez-Márquez[1]

[1]*Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Mexico City, Mexico*

Abstract

This paper presents a comprehensive analysis of ensemble models for the shared task "Conspiracy Theories Against Critical Thinking Narratives" for PAN at CLEF 2024. Through a data collection involving Telegram conversations on COVID-19, two distinct corpora in English and Spanish were assembled and manually labeled to differentiate between "critical" and "conspiracy" texts. The study employed ensemble models, comprising seven trained transformer-based models per language-task pair, to address two key tasks: distinguishing between critical and conspiracy texts (binary classification) and detecting spans for six different categories that can be found on the texts (multi-label span classification). The results unveiled the competitive performance of ensemble models, particularly in securing notable rankings surpassing the mean of all participants' results in both tasks.

Keywords

Conspiracy Theories, Critical Thinking Narratives, Multi-label Token Classification, Ensemble Model, Small Language Models

1. Introduction

Conspiracy theories (CT) are narratives that seek to explain the causes of significant situations or events for society, suggesting the existence of secret plans secretly carried out by actors who abuse their power to achieve their own objectives without caring about depriving people of their rights, freedoms, prosperity, health or knowledge [1, 2, 3]. These narratives can cause great harm, as they can modify the behavior of people who believe in them, fostering attitudes that put both believers and other members of society at risk. The potential risk increases when it comes to health-related conspiracy theories, as they can lead some people to make decisions that are detrimental to their well-being and that of those around them.

In addition to the behavioral change in believers of these theories, another significant harm is the mistrust they generate towards various medical treatments and the decrease in trust in public health institutions and health professionals. This hinders the implementation of public health measures and the response to health emergencies. For these reasons, it is urgent to identify and address conspiracy theories to mitigate their harmful effects.

CEUR-WS.org/Vol-3740/paper-233.pdf

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

sdamians2019@cic.ipn.mx (S. Damián); bherrerag2019@cic.ipn.mx (B. Herrera); dvazquezs2019@cic.ipn.mx (D. Vázquez); hcalvo@cic.ipn.mx (H. Calvo); edgardo@cic.ipn.mx (E. Felipe-Riverón); cyanez@cic.ipn.mx (C. Yáñez-Márquez)
https://github.com/sdamians (S. Damián); https://github.com/Hiram02 (D. Vázquez); http://hiramcalvo.com/ (H. Calvo)
0000-0003-2836-2102 (H. Calvo)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

1.1. Medical conspiracy theories

Although conspiracy theories are not limited to the field of health, they have been a persistent issue over the years, causing significant harm to the population. A clear example is the case of the smallpox vaccine, discovered by Edward Jenner in 1796, which represented a monumental advance with the potential to improve public health significantly. However, it also led to the creation of a CT [4]. It is likely that people did not properly understand how it worked, which led to the spread of rumors warning of horn growth resulting from its use.

And this is not the only case of conspiracy theories related to vaccines. In fact, they have been a recurring theme. For instance, in 1981, Dr. John Wilson claimed that the DPT vaccine caused convulsions and brain damage [5]. In 1998, Andrew Wakefield published an article suggesting a link between the MMR vaccine and autism [6], although it should be noted that this article was retracted by the journal in which it was published. More recently, the COVID-19 pandemic has fueled the spread of numerous conspiracy theories regarding vaccination against this virus [7].

1.2. Negative impacts of conspiracy theories

In general, the propagation of conspiracy theories could have several negative effects, among which we can highlight some of them:

- Social Division and Polarization: They exacerbate social divisions by promoting extreme and exclusionary beliefs, hindering rational dialogue and societal cohesion.
- Dissemination of Misinformation: They contribute to spreading false and unverified information, leading to confusion and potentially harmful decisions.
- Loss of trust in authorities and experts: They foster distrust towards governmental, scientific, and public health institutions, as well as towards experts in different areas.
- Psychological Impact: They induce anxiety, fear, and paranoia among believers, negatively affecting their emotional and mental well-being.
- Impaired Decision-Making: Believers may base decisions on misinformation or biased information, impeding informed and rational decision-making processes.

In the health field, conspiracy theories have had significant adverse effects. In Pakistan, for example, there is a belief that the polio vaccine was developed by the CIA to sterilize Muslim men [8], which has led many people to reject it. Another example is the theory that the U.S. government created HIV/AIDS to reduce the African-American population, a widespread belief among this community that has resulted in less frequent condom use [9].

Furthermore, certain sectors of society maintain mistrust towards specific drugs, alleging they inflict greater harm than the diseases they aim to cure. For instance, there exists a theory attributing the majority of deaths among AIDS patients to retroviral drugs. This conspiracy theory holds particular influence in sub-Saharan Africa, where it receives support from influential figures [10].

There are several reasons why conspiracy theories can be widely spread. Among the most prominent ones is their propagation by celebrities through digital media [11], which causes many of their followers to start believing in them. In addition, it is difficult to absolutely determine their falsity, together with the degree of plausibility attributed to them by each person [12], significantly contributes to their dissemination. Critical thinking can help people to better evaluate the information they receive in daily life and thus avoid fraud and harmful habits. For example, critical thinking can be useful in differentiating reliable medical information from unfounded claims, helping in decision-making about appropriate treatments and lifestyle. When a person with high levels of intelligence, but low levels of critical thinking, believes in a conspiracy theory, they can generate very well-supported arguments to support the false information [13]. These arguments can be quickly propagated through digital media and are difficult to detect.

This year's goal at PAN 2024 is to analyze texts reflecting oppositional thinking, specifically distinguishing between conspiracy theories and critical thinking narratives. This task addresses two

significant challenges for the NLP community: (subtask 1, a binary classification task) differentiating between conspiracy and critical narratives, and (subtask 2, a multi-label span classification task) identifying key elements of narratives that fuel intergroup conflict. Making this distinction is crucial because mislabeling a text as conspiratorial when it is merely oppositional to mainstream views could push individuals who are simply questioning mainstream perspectives closer to conspiracy communities [14, 15].

2. Related Work

Conspiracy theories represent a significant danger, as they can negatively influence people's behavior, affecting trust in institutions and fostering disinformation. Intelligence is often thought of as synonymous with critical thinking, however, these terms are not the same. In reality, intelligence alone does not always translate into critical thinking. Over the course of history, there have been people with high levels of intelligence who nevertheless have demonstrated a lack of critical thinking in some areas, for instance, Sir. Arthur Conan Doyle a brilliant writer who believed in spiritualism and fairies, despite clear evidence to the contrary [16].

A recent study [13] explores the connections between critical thinking, intelligence and the predisposition to believe in conspiracy theories. The authors note that while intelligence can help people formulate more sophisticated arguments, it does not always protect them from false beliefs. On the other hand, critical thinkers use logical rules, standards of evidence and other criteria that must be met for the product of a thought to be considered good, making them less likely to believe in unsubstantiated claims.

Intelligence is generally associated with good cognitive processing or intellectual abilities and the potential to learn and reason well. Intelligent people tend to perform well in basic real-world domains, such as academic performance and job success but sometimes find it difficult to adapt in other real-world situations [17]. Intelligence without critical thinking can sometimes result in more convincing arguments that support false beliefs. These persuasive arguments can mislead many people into accepting these false ideas.

In a companion study [18], the impact of cognitive styles, such as analytical thinking, critical thinking, and scientific reasoning, on the propensity to believe in conspiracy theories was examined. The findings suggest that individuals who exhibit a stronger inclination towards analytical thinking and scientific reasoning are less susceptible to conspiracy theories due to their more rigorous and evidence-based approach to evaluating information.

As a matter of fact, in recent years, there has been a notable surge in the recognition and analysis of conspiracy theories. This trend mirrors the growing acknowledgment of the significant impact that misinformation and disinformation can have on societies, particularly in the age of digital interconnectedness. Research endeavors[19], have increasingly focused on understanding the dynamics behind the propagation of conspiracy theories.

However, it's crucial to recognize that the identification and mitigation of conspiracy theories are part of a broader spectrum of tasks aimed at combating misinformation and preserving the integrity of information ecosystems. Alongside the detection of conspiracy theories, researchers and practitioners are also confronted with related challenges, including the identification and containment of rumors[20], the mitigation of the spread of fake news[21], the recognition of clickbait content [22] designed to manipulate user engagement, and the indispensable task of fact-checking[23].

In this contemporary landscape, where information dissemination is facilitated by sophisticated technologies and platforms, the importance of discerning false information from genuine content cannot be overstated. The rise of AI-driven text generation capabilities, for instance, presents both opportunities and challenges. On one hand, these advancements offer innovative approaches to understanding and combating misinformation. On the other hand, they underscore the urgency of developing robust mechanisms to differentiate between authentic and fabricated texts.

3. Dataset Preprocessing

The data collection involved gathering textual data from Telegram conversations concerning COVID-19. These texts were then manually labeled to distinguish between "critical" and "conspiracy" categories. Two corpora were employed for this study, one in English and the other in Spanish. Each set contained 4000 entries for training purposes and an additional 1000 entries for testing [24]. In this work, the use of k-fold cross validation was not implemented due to limited computational resources. Instead, the 10% of the training set was split for validation experiments, preserving the initial class balance provided, as shown in Table 1. The main hypothesis was that a single train-validation split could lead to a scenario where the model stability were more consistent than averaging results over multiple folds specially for subtask 2.

Table 1
Statistics of both corpora.

Language	Label	Numerical Label	Train	Val	Total
English	Conspiracy	0	1105	274	1379
English	Critical	1	2466	155	2621
Total	-	-	3571	429	4000
Spanish	Conspiracy	0	1178	284	1462
Spanish	Critical	1	2373	165	2538
Total	-	-	3551	449	4000

The dataset entries had two different representations: the original sentence and the sentence split by tokens designed for subtask 2. The approach implemented was to use the original sentence representation for subtask 1, leveraging the tokenization step to each model's tokenizer and to use the list of tokens for subtask 2, trying to preserve the majority of the tokens labeled after the preprocess and cleaning step. The following procedures for text cleaning were implemented for both tasks:

- Small combinations of numbers and letters (with lengths ranging from 2 to 4) were removed.
- Combinations of alternating letters and numbers were removed (e.g. tokens such as *1df324D* identified in URLs).
- Special words for URLs were removed.
- English contractions such as *'re* or *n't* were normalized by using the complete word (*are* and *not*).
- Numbers in date format and hour were tagged using the labels *date*, *hour* for English and *fecha*, *hora* for Spanish.
- The rest of the numbers were tagged using the label *number*.
- Repeated strings of three or more characters were normalized (e.g. *aaa* to *a*).

Significantly, both corpora manifest an inherent class distribution imbalance, characterized by a larger proportion of inputs labeled as "critical" in contrast to those categorized as "conspiracy", which is illustrated in Figure 1.

4. Methodology

The baseline model provided was a transformer-based model designed for multitask learning to address both tasks. While this generally leads to better results, it can also make the model more complex and difficult to train, particularly in balancing the loss of both tasks to prevent one task from negatively impacting the performance of the other.

The proposed solution in this work involved using an ensemble of transformer-based models in the form of several Small Language Models (SLMs) to address each task-language pair independently, thus training them as single-task learning models using low computational resources. This methodology

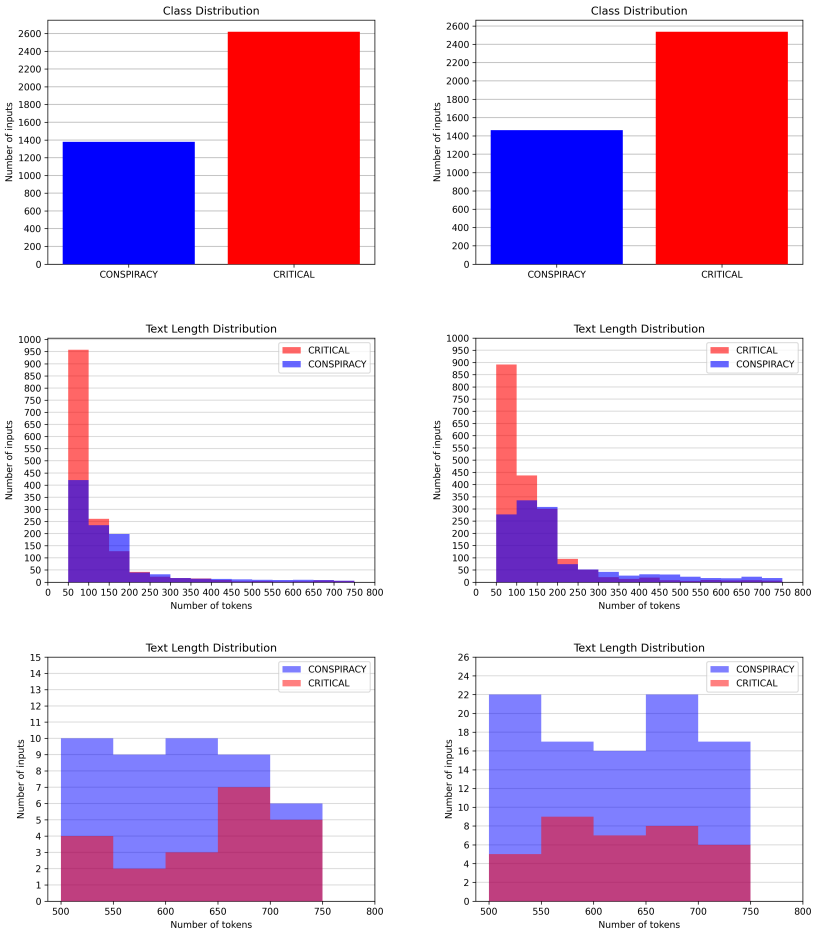

Figure 1: Dataset statistics for both the English (left) and Spanish (right) corpora. The first pair of charts illustrates an imbalance of classes for both datasets. The second pair denotes the distribution of inputs by length (number of tokens) and class, and the third pair provides a zoomed-in version of the second pair, allowing for a clearer view of the distribution of inputs that are longer than the common length typically encountered in SLMs.

facilitated the aggregation of multiple logits and aimed to improve overall performance. The training process consisted of developing seven distinct SLMs for each language and task. Subsequently, the top five models for each language-task pair were selected based on specific evaluation metrics. For subtask 1, the official evaluation metrics were the Matthews correlation coefficient (MCC) [25] and macro F1-score, while subtask 2 was evaluated using the span-F1 metric [26].

Figure 2 illustrates the ensemble strategy utilized in this work for subtask 1. All logits obtained by each SLM were multiplied by a weight based on the scores of the evaluation metrics. Subsequently, the logits were aggregated and rounded, to get the final outcome of the ensemble model. The same strategy was applied for subtask 2, where instead of getting a single outcome per SLM, a matrix $Y_n \in \mathbb{R}^{j \times k}$ was obtained and aggregated aftwewards, as depicted in Figure 3. In summary, two ensemble models were evaluated per task-language pair, one using all seven trained SLMs and another using the top 5 best trained SLMs. Each ensemble model employed a mean voting classifier.

4.1. Small Language Models employed for English Corpus

This work's rigorous selection process led to the identification of several transformer-based models for both subtasks within the English corpus. The transformer library by huggingface provides wrappers for sequence classification and token classification tasks. The following enumeration provides a concise description of the models assessed.

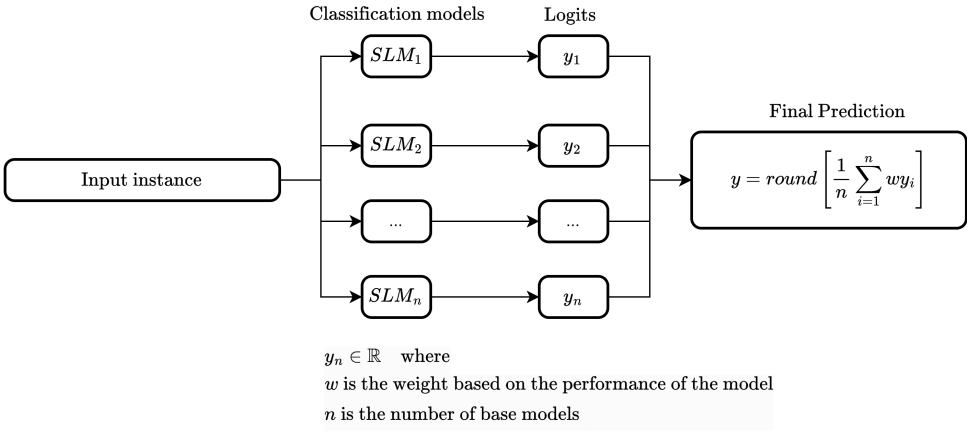

Figure 2: Diagram of the ensemble model approach for subtask 1. All logits obtained by each model are averaged and rounded to get the final prediction for each input.

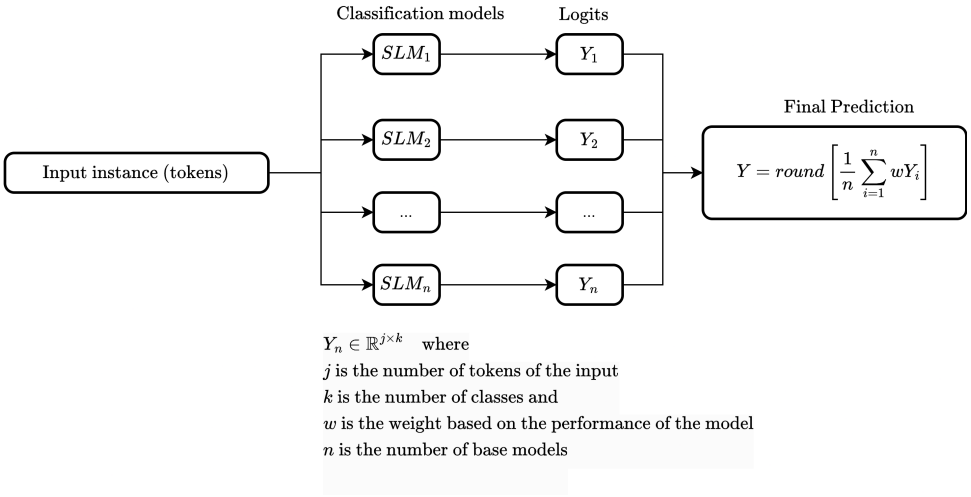

Figure 3: Diagram of the ensemble model approach for subtask 2. All logits obtained by each model are averaged and rounded to get the final prediction for each input.

- BERT [27]: Demonstrates a significant performance in understanding context and semantics, making it a natural choice. The baseline provided was constructed utilizing it.
- RoBERTa [28]: Employs an optimized pretraining and can achieve better results than BERT.
- BigBird [29]: Handles long sequences through sparse attention mechanisms. The English corpus comprises several long sequences of tokens that exceed the typical maximum length (512) accepted by SLMs.
- Electra [30]: Utilizes a generator-discriminator architecture for enhanced efficiency, offering robustness against adversarial attacks and enhancing generalization capabilities.
- T5 [31]: Adopts a text-to-text framework that can handle diverse tasks. Although it is a text-generating model, it can be used as a binary classification by adding a classification module (a linear layer on top of the pooled output). For classification tasks, the output of the first token is processed and classified. Huggingface has an implementation of this model's variant.
- XLM-RoBERTa [32]: Extends RoBERTa to multiple languages, producing distinct representations of the inputs, potentially offering a complementary perspective on the tasks.
- MDeBERTa [33]: Designs efficient multilingual representations like XLM-RoBERTa, thereby providing another perspective of the tasks.

Table 2 displays the metric outcomes for each SLM to subtask 1 on the English corpus. Notably,

MDeBERTa and T5 models achieved the most favorable results, outperforming the rest. Conversely, Table 3 showcases the results for the macro span-f1 metric associated with subtask 2 on the English corpus. Here, the multilingual model MDeBERTa and Electra emerged as the best models, while T5 exhibited comparatively insignificant results.

Table 2
Evaluation results for each SLM trained for Task 1 on the English corpus

Model	MCC	F1-Macro	F1-Conspiracy	F1-Critical
XLM-RoBERTa	0.8005	0.8960	0.9256	0.8664
T5	**0.8358**	0.9123	0.9395	0.8851
RoBERTa	0.8204	0.9053	0.9336	0.8771
MDeBERTa	0.8347	**0.9131**	**0.9388**	**0.8874**
Electra	0.8292	0.9110	0.9367	0.8852
BigBird	0.8315	0.9096	0.9378	0.8814
BERT	0.8247	0.9086	0.9348	0.8824

Table 3
Evaluation results for each SLM trained for Task 2 on the English corpus

Model	Campaigner	Neg Effect	Objective	Victim	Agent	Facilitator	F1-Macro
BERT	0.5712	0.4983	0.3305	0.5544	0.6631	0.3434	0.4935
BigBird	0.6107	0.4967	0.3991	0.5903	0.5814	0.3852	0.5106
Electra	0.5962	0.5140	0.3651	**0.6121**	**0.6873**	0.3964	0.5285
MDeBERTa	0.6257	0.5160	**0.4048**	0.6118	0.6857	**0.4263**	**0.5450**
RoBERTa	**0.6365**	**0.5205**	0.3650	0.5897	0.6724	0.3561	0.5233
T5	0.4632	0.4042	0.2621	0.5344	0.5959	0.2231	0.4138
XLM-RoBERTa	0.6274	0.5093	0.3500	0.5947	0.6457	0.3433	0.5117

4.2. Models employed with Spanish corpus

In alignment with the specific demands of the Spanish corpus, a tailored selection of seven models was employed, all implemented using the Hugging Face Transformers library. The following list provides a description of these models:

- BETO [34]: Encompasses proficient linguistic understanding and contextual comprehension of the Spanish language, and it served as the baseline model for the subtasks.
- Bertin [35]: Contributes to the linguistic analysis of Spanish language, providing an alternative model for addressing linguistic nuances.
- MarIA [36]: Demonstrates proficiency and efficacy in addressing the complexities of the Spanish language, being trained by large amounts of Spanish texts.
- TwHIN-BERT [37]: Enhances capabilities in processing linguistic structures, being tailored for hate speech detection in Spanish, particularly on social media.
- mT5 [38]: Offers a multilingual variant of the T5 model, and enriches the analytical repertoire available for the Spanish language. It is also a generative text model.
- XLM-RoBERTa [32]: Proposes another variant of the inputs, offering an additional multilingual perspective on the tasks.
- MDeBERTa [33]: As a third multilingual representation, it offers valuable insights, augmenting the analytical approach of the solution approach.

Table 4 presents the metric outcomes for each trained model concerning subtask 1 for Spanish language. Remarkably, MarIA and MDeBERTa demonstrated the most promising results, surpassing its

counterparts. On the other hand, Table 5 delineates the results for subtask 2 on the Spanish corpus. For this language, the multilingual model MDeBERTa emerged as the leading performer, while mT5 displayed relatively negligible results, mirroring the outcomes obtained by its counterpart in the English experiments.

Table 4
Evaluation results for each SLM trained for subtask 1 on the Spanish corpus

Model	MCC	F1-Macro	F1-Conspiracy	F1-Critical
Bertin	0.6204	0.8033	0.8515	0.7552
BETO	0.6694	0.8284	0.8723	0.7844
MarIA	**0.7029**	**0.8437**	**0.8862**	**0.8012**
MDeBERTa	0.6882	0.8371	0.8803	0.7939
mT5	0.2750	0.4782	0.7362	0.2203
TwHIN-BERT	0.6539	0.8154	0.8671	0.7669
XLM-RoBERTA	0.6250	0.8113	0.8508	0.7718

Table 5
Evaluation results for each SLM trained for subtask 2 on the Spanish corpus

Model	Campaigner	Neg Effect	Objective	Victim	Agent	Facilitator	F1-Macro
Bertin	0.6559	0.6354	0.2850	0.6239	0.5556	0.4279	0.5306
BETO	0.6785	0.6506	0.3104	0.5965	0.5314	0.4350	0.5561
MarIA	0.6740	0.6055	0.3214	0.5965	0.5314	0.4350	0.5273
MDeBERTa	**0.7117**	0.6554	0.3291	0.6392	**0.6064**	0.4988	**0.5742**
mT5	0.6312	0.6034	0.3474	0.5801	0.5246	0.3268	0.5022
TwHIN-BERT	0.6597	0.6369	**0.3475**	**0.6396**	0.5700	**0.5015**	0.5592
XLM-RoBERTA	0.6908	**0.6756**	0.3309	0.6367	0.5939	0.4861	0.5690

5. Results

The shared task allowed a maximum of two submissions per subtask. For our submissions, we opted to present two ensemble models per subtask: an ensemble version comprising all seven models trained per language-task pair, alongside another submission featuring the top five models. Table 6 provides a comprehensive overview of the official results attained per submission for subtask 1, incorporating the attained placement, while Table 7 delineates the results for subtask 2. The best models were determined on their competitiveness across the Matthews Correlation Coefficient (MCC) metric and span-F1 metric, for both subtasks respectively. Due to complications encountered during the experimentation phase, the evaluation of the ensemble model comprising the top 5 models for Spanish was precluded. For subtask 1, the optimal ensemble model surpassed the baseline performance for the English language. However, the submitted ensemble model for the Spanish language did not exhibit a similar performance. Conversely, for subtask 2, the optimal ensemble model successfully outperformed the baselines for both languages. In this subtask, the ensemble model with five learners was the best approach for the English language, while the ensemble model with seven learners was the best for the Spanish language. The results obtained for the Spanish language were significantly higher than its baseline, which implies the learners successfully contributed different information to the final solutions.

Table 6
Test results for each submission for subtask 1. The baseline is included for comparison purposes.

Language	Model	MCC	F1-Macro	F1-Conspiracy	F1-Critical	Rank
English	Ensemble (7 models)	**0.7970**	0.8985	0.8674	0.9296	**14/83**
English	Ensemble (5 models)	0.7943	**0.9071**	**0.9080**	0.9061	
English	Baseline-BERT	0.7964	0.8975	0.8632	**0.9318**	
Spanish	Ensemble (7 models)	0.6462	0.8231	0.7753	0.8708	**29/78**
Spanish	Baseline-BETO	**0.6681**	**0.8339**	**0.7872**	**0.8806**	

Table 7
Test results for each submission for subtask 2. The baseline is included for comparison purposes.

Model	Campaigner	span-F1	span-P	span-R	micro-span-F1	Rank
English	Ensemble (7 models)	0.5460	0.5287	0.5774	0.5133	
English	Ensemble (5 models)	**0.5598**	**0.5332**	0.6012	**0.5287**	**12/28**
English	Baseline-BERT	0.5323	0.4684	**0.6334**	0.4998	
Spanish	Ensemble (7 models)	**0.5529**	**0.5384**	0.5785	0.5323	**07/25**
Spanish	Ensemble (5 models)	0.5483	0.5210	**0.5873**	**0.5383**	
English	Baseline-BETO	0.4934	0.4533	0.5621	0.4952	

6. Conclusions

The ensemble model's combination of diverse SLM architectures contributed to robustness and generalization, thereby enhancing performance across both tasks. However, certain limitations and areas for improvement were identified. A small fixed validation set was used, but a cross-validation strategy might lead to better performance, specially for obtaining more accurate weights for the base models. The ensemble used a weighted mean voting classifier that can be replaced for a more sophisticated meta model like a logistic regression classifier. The single-task learning approach did not outperform all the baseline results obtained using a multitask learning approach. The shared knowledge from both subtasks might enhance the results and the generalization of the final predictions.

The disparities in performance between tasks could be attributed to the inherent complexity and ambiguity associated with detecting different classes among texts, necessitating further exploration and refinement of the approach's methodologies and feature representations. By leveraging insights gleaned from the model performance analysis, future iterations of the ensemble model can be refined to enhance robustness and efficacy within the domain of conspiracy theories and critical thinking narratives.

Acknowledgments

This work was done with partial support from the Mexican Government through Consejo Nacional de Humanidades Ciencias y Tecnologías (CONAHCYT) and Instituto Politécnico Nacional (IPN).

References

[1] M. R. X. Dentith, M. Orr, Secrecy and conspiracy, Episteme 15 (2018) 433–450. doi:10.1017/epi.2017.9.

[2] C. R. Sunstein, A. Vermeule, Conspiracy theories: Causes and cures*, Journal of Political Philosophy 17 (2009) 202–227. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9760.2008.00325.x. doi:https://doi.org/10.1111/j.1467-9760.2008.00325.x.

[3] J. E. Uscinski, J. M. Parent, American Conspiracy Theories, Oxford University Press, 2014. URL: https://doi.org/10.1093/acprof:oso/9780199351800.001.0001. doi:10.1093/acprof:oso/9780199351800.001.0001.

[4] M. V. Eve Dubé, N. E. MacDonald, Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications, Expert Review of Vaccines 14 (2015) 99–117. URL: https://doi.org/10.1586/14760584.2015.964212. doi:10.1586/14760584.2015.964212. arXiv:https://doi.org/10.1586/14760584.2015.964212, pMID: 25373435.
[5] J. T. Wilson, Dpt vaccine and serious neurological illness: current status of the controversy, Pediatrics 68 (1981) 650–651.
[6] A. Wakefield, S. Murch, A. Anthony, J. Linnell, D. Casson, M. Malik, M. Berelowitz, A. Dhillon, M. Thomson, P. Harvey, A. Valentine, S. Davies, J. Walker-Smith, Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children, The Lancet 351 (1998) 637–641.
[7] N. Corbu, R. Buturoiu, V. Frunzaru, G. Guiu, Vaccine-related conspiracy and counter-conspiracy narratives. silencing effects, Communications 49 (2024) 339–360. URL: https://doi.org/10.1515/commun-2022-0022. doi:doi:10.1515/commun-2022-0022.
[8] G. E. Andrade, A. Hussain, Polio in pakistan: Political, sociological, and epidemiological factors, Cureus 10 (2018) e3502. doi:10.7759/cureus.3502.
[9] L. M. Bogart, S. T. Bird, Exploring the relationship of conspiracy beliefs about hiv/aids to sexual behaviors and attitudes among african-american adults, Journal of the National Medical Association 95 (2003) 1057.
[10] P. Fourie, M. Meyer, The Politics of AIDS Denialism, Routledge, New York, 2010.
[11] G. Andrade, Medical conspiracy theories: cognitive science and implications for ethics, Medicine, Health Care and Philosophy 23 (2020) 505–518. URL: https://doi.org/10.1007/s11019-020-09951-6. doi:10.1007/s11019-020-09951-6.
[12] M. Frenken, A. Reusch, R. Imhoff, "just because it's a conspiracy theory doesn't mean they're not out to get you": Differentiating the correlates of judgments of plausible versus implausible conspiracy theories, Social Psychological and Personality Science (2024) 19485506241240506. URL: https://doi.org/10.1177/19485506241240506. doi:10.1177/19485506241240506.
[13] D. A. Bensley, Critical thinking, intelligence, and unsubstantiated beliefs: An integrative review, Journal of Intelligence 11 (2023). URL: https://www.mdpi.com/2079-3200/11/11/207. doi:10.3390/jintelligence11110207.
[14] 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: Proceedings of the Fifteenth International Conference of the CLEF Association CLEF-2024, Springer, 2024, pp. 3–10.
[15] 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.
[16] T. Waters, Magic and the british middle classes, 1750–1900, Journal of British Studies 54 (2015) 632–653. URL: http://www.jstor.org/stable/24702123.
[17] D. F. Halpern, D. S. Dunn, Critical thinking: A model of intelligence for solving real-world problems, Journal of Intelligence 9 (2021). URL: https://www.mdpi.com/2079-3200/9/2/22. doi:10.3390/jintelligence9020022.
[18] 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). URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.736838. doi:10.3389/fpsyg.2021.736838.
[19] A. Giachanou, B. Ghanem, P. Rosso, Detection of conspiracy propagators using psycho-linguistic characteristics, Journal of Information Science 49 (2023) 3–17. URL: https://doi.org/10.1177/0165551520985486. doi:10.1177/0165551520985486. arXiv:https://doi.org/10.1177/0165551520985486.

[20] G. Gorrell, E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, L. Derczynski, Semeval-2019 task 7: Rumoureval 2019: Determining rumour veracity and support for rumours, in: Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019, Association for Computational Linguistics, 2019, pp. 845–854.
[21] N. Capuano, G. Fenza, V. Loia, F. D. Nota, Content-based fake news detection with machine and deep learning: A systematic review, Neurocomputing 530 (2023) 91–103.
[22] A. Anand, T. Chakraborty, N. Park, We used neural networks to detect clickbaits: You won't believe what happened next!, in: Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings 39, Springer, 2017, pp. 541–547.
[23] N. Walter, J. Cohen, R. L. Holbert, Y. Morag, Fact-checking: A meta-analysis of what works and for whom, Political communication 37 (2020) 350–375.
[24] D. Korenčić, B. Chulvi, X. B. Casals, M. Taulé, P. Rosso, Pan24 oppositional thinking analysis [data set] (2024). URL: https://doi.org/10.5281/zenodo.11199642. doi:10.5281/zenodo.11199642.
[25] 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.
[26] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, et al., Fine-grained analysis of propaganda in news articles, in: Proceedings of EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, 2019, pp. 5636–5646.
[27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[28] 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, arXiv preprint arXiv:1907.11692 (2019).
[29] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big bird: Transformers for longer sequences, Advances in neural information processing systems 33 (2020) 17283–17297.
[30] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, arXiv preprint arXiv:2003.10555 (2020).
[31] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (2020) 1–67.
[32] 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, arXiv preprint arXiv:1911.02116 (2019).
[33] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).
[34] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, arXiv preprint arXiv:2308.02976 (2023).
[35] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.
[36] 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.
[37] X. Zhang, Y. Malkov, O. Florez, S. Park, B. McWilliams, J. Han, A. El-Kishky, Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations, arXiv preprint arXiv:2209.07562 (2022).
[38] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A

massively multilingual pre-trained text-to-text transformer, arXiv preprint arXiv:2010.11934 (2020).