Overview of the Oppositional Thinking Analysis PAN Task at CLEF 2024

Damir Korenčić[1], Berta Chulvi[2,5], Xavier Bonet-Casals[3], Mariona Taulé[3], Paolo Rosso[4,6,*] and Francisco Rangel[5]

[1]*Ruđer Bošković Institute, Croatia*
[2]*Universitat de València, Spain*
[3]*Universitat de Barcelona, Spain*
[4]*Universitat Politècnica de València, Spain*
[5]*Symanto Research, Spain*
[6]*ValgrAI Valencian Graduate School and Research Network Analysis of Artificial Analysis, Spain*

Abstract
This paper describes the Oppositional Thinking Analysis task at CLEF 2024. The task focuses on analyzing conspiracy theories and critical thinking narratives, and is comprised of two subtasks. Subtask 1 is a binary classification task aimed at distinguishing between critical and conspiracy texts. Subtask 2 is a token classification task aimed at detecting text spans corresponding to the key elements of oppositional (critical and conspiracy) narratives. The subtasks are based on a dataset of English and Spanish COVID19-related texts obtained from oppositional Telegram channels, and labeled using a topic-agnostic annotation scheme [1]. A total of 82 teams participated in the challenge, and 17 teams published working notes papers with system descriptions. The participants employed a range of NLP methods and pushed the state-of-art performance on both subtasks beyond the performance of the strong baseline systems [1] that were provided.

Keywords
Conspiracy Theories, Oppositional Thinking, Computational Social Science, Natural Language Processing, Text Classification, Sequence Labeling

1. Introduction

The first edition of the Oppositional Thinking Task, held at CLEF 2024, focused on distinguishing automatically between conspiratorial narratives and critical narratives that do not convey a conspiratorial mentality. Conspiracy Theories (CTs) are causal explanations of significant events that present them as a result of cover plots orchestrated by secret, powerful, and malicious groups [2]. Since conspiracy narratives tend to convey a critical vision of mainstream policies, a common mistake, especially in the middle of a global crisis such as a pandemic or a war, is to categorize every critical narrative against the official discourse as *conspiratorial*. Criticism and free discussion are key values in democratic societies; however, conspiracy narratives severely weaken democratic systems because they place the ultimate agent of the crisis outside the control of our systems of governance. As a result, it is important not to confuse critical and conspiracy narratives.

The interest in the automatization of the critical-conspiracy distinction was recently highlighted by Korenčić et al. [1], who argued that, if models monitoring the social media messages do not differentiate between critical and conspiratorial thinking, there is a high risk of pushing people toward conspiracy communities. The sociopsychological basis of this process is based on Social Identity Theory. Social Identity Theory (SIT) has been a cornerstone in understanding group processes and intergroup relations since its inception in the early 1970s [3]. This theory posits that individuals derive a part of their self-concept from their membership in social groups, which influences their behavior and attitudes

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
prosso@dsic.upv.es (P. Rosso)
0000-0003-4645-2937 (D. Korenčić); 0000-0003-1169-0978 (B. Chulvi); 0009-0003-8827-0215 (X. Bonet-Casals); 0000-0003-0089-940X (M. Taulé); 0000-0002-8922-1242 (P. Rosso); 0000-0002-6583-3682 (F. Rangel)

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

towards in-group and out-group members [4, 5]. As a result, being considered a conspiracist when you are not could be a threat to your social identity. Once the subject is the target of this accusation, a way to repair this stigmatization is to join conspiracist groups that will give the social support needed to recover a positive social identity. This process is not unusual. As several authors from the field of social sciences suggest, a fully-fledged conspiratorial worldview is the final step in a progressive "spiritual journey" that sets out questioning social and political orthodoxies [6, 7, 8]. Accordingly, the distinction between conspiratorial and critical thinking is crucial for automated content moderation: without it, there is a significant risk of driving individuals towards conspiracy communities. Specifically, mislabeling a text as conspiratorial when it merely challenges mainstream perspectives could inadvertently steer individuals who are simply questioning into the arms of conspiracy groups.

Furthermore, in the area of computational linguistics, Korenčić et al. [1] have shown that conspiracist narrative and critical thinking are different due to their potential social effect on public opinion discourse, with the former being significantly more associated with violent words and expressions of anger. In their corpus, the authors have also labelled key elements in oppositional narratives (goals, effects, agents, and the two groups in conflict, facilitators of government decisions and campaigners against them), demonstrating that a greater level of intergroup conflict between facilitators and campaigners is associated especially with conspiracy narratives and correlates with a greater use of violent words and the emotional manifestation of anger.

Based on this recent research [1], the present task addresses two new challenges for the NLP research community: (1) to distinguish the conspiracy narrative from other oppositional narratives that do not express a conspiracy mentality (i.e., critical thinking); and (2) to identify the key elements of the oppositional narrative in online messages. As demonstrated [1], predictive NLP systems for these two tasks have value for computational social scientists who are interested in analyzing oppositional narratives. Therefore, it is of interest to push the performance on these tasks beyond the previously proposed NLP approaches [1]. This PAN task has attempted to achieve this goal.

For the two tasks described above, we provide the XAI-Disinfodemic corpus [1], a multilingual (English and Spanish) corpus consisting of 10,000 annotated Telegram messages that focus on oppositional narratives related to the COVID-19 pandemic. For each language, a training set of 4,000 messages has been provided to the participants, while the outputs of the systems were computed and evaluated using the testing set consisting of 1,000 messages. These messages contain oppositional non-mainstream views on the COVID-19 pandemic, classified into two categories: critical and conspiratorial messages. Messages have been annotated at the span level with a topic-agnostic schema that distinguishes the key elements of an oppositional narrative: objectives, negative effects, agents, victims, and facilitators and campaigners (the two groups in conflict). We also provide strong baseline solutions [1]. The train and test splits of the dataset, as well as the code of the baseline systems, are freely available[1].

The following sections of this paper describe the key aspects of this task. Section 2 summarizes the related work on the classification of conspiratorial narratives in NLP and on the span detection of different elements of these narratives. Section 3 presents the dataset used in this task. Section 4 describes the two subtasks proposed above, as well as evaluation measures and baseline solutions. Section 5 presents the systems used by the participants. Section 6 analyzes the results and the systems of the participants. Finally, Section 7 contains conclusions and directions for future work.

2. Related Work

A recent literature review by Mahl et al. [9] indicates a rising interest in conspiracy theories within online environments, particularly within the Social Sciences. Approximately 80% of the research focuses on written content, with about a third using automated content analysis methods. In this chapter, we review research from NLP area which are relevant to the present tasks.

[1]https://github.com/dkorenci/pan-clef-2024-oppositional

2.1. Conspiracy detection in NLP

The COVID-19 pandemic has been one of the topics that has garnered the most attention in the study of conspiracy narratives since 2020. The pandemic has been fertile ground for the expansion of conspiracy theories. Among the works oriented in this direction, Uscinski et al. [10] collected a dataset of letters sent to a mainstream US publication, and labeled them as either containing a conspiracy or not. Another available corpus dedicated to conspiracy theories is LOCO corpus [11] containing 96,743 texts from a diverse collection of mainstream and conspiracy outlets. The texts are enriched with website metadata and auto-generated topics. With more detail about the content of conspiracy theories, we find COCO, a corpus of 3,495 texts promoting COVID-19 conspiracies [12]. The texts were manually annotated in the COCO corpus with a fine-grained classification scheme encompassing conspiracy sub-topics.

The problem has often been approached as a binary classification task with the goal of distinguishing conspiratorial from non-conspiratorial text. A good example is the two recent MediaEval challenges. Focusing on the classification of conspiracy texts [13, 14], this task led to a number of approaches demonstrating that the state-of-the-art architecture is a multi-task classifier [15, 16, 17] based on CT-BERT [18].

More nuanced methodologies using fine-grained approaches, like multi-label or multi-class classifications, have provided a detailed understanding [19, 20, 13, 14] of the diffusion of conspiracies. For example, Moffitt et al. [20] developed a classifier of conspiracy tweets and used it for propagation analysis. COVID-19 origin conspiracy theory tweets using this method and then used social cybersecurity methods to analyze communities, spreaders, and characteristics of the different origin-related conspiracy theory narratives. This research found that tweets about conspiracy theories were supported by news sites with low fact-checking scores and amplified by bots who were more likely to link to prominent Twitter users than in non-conspiracy tweets.

Other research in computational linguistics has dealt with different aspects related to the characteristics of the disseminators of conspiracy narratives or has focused on the characteristics of the messages. Bessi [21] employed a text scaling method to map conspiratorial texts to personality traits and analyze these conspiracies. Giachanou et al. [19] used psychological and linguistic features to classify and analyze the social media users who spread conspiracies. Topic modeling techniques were used by other authors [22, 23] to extract and examine common themes within conspiracy texts. Levy et al. [24], taking an approach different from the problem of classifying humans texts, analyze the capacity of large language models to generate conspiracies.

However, present research fails to differentiate between critical thinking and conspiratorial thinking, which is the main goal of this task.

2.2. Span detection in conspiracy theories

In the field of conspiracy theories, several papers have addressed the challenge of span detection. Samory and Mitra [23] utilized syntactic parsing to identify "motifs" (agent-action-target triplets) and analyze the patterns of their occurrence. Introne et al. [25] propose a span-level scheme of six categories (event, actor, goal, action, consequence, target), and use it to analyze 236 messages from anti-vaccination forums. They distinguish between *conspiracy theories* and *conspiratorial thinking*, a category that implies only passive support for a conspiracy. This distinction is not based on annotations grounded in theory but on the requirement of all the categories being present in a given text. However, in practice, fewer elements can convey a conspiracy theory in a very strong manner. Although this research identifies different elements of discourse, it also fails to consider the role played by intergroup conflict in the conspiracy narrative, which is addressed in the XAI-DisInfodemic corpus [1].

Holur et al. [26] focus on oppositional elements in the conspirational narrative, detecting the so-called *insider* and *outsider* entities within conspiracy texts by automatically labelling noun phrases. This *insider* and *outsider* schema is based on the positive or negative sentiment that each user conveys for each entity. Although this research starts a path that could arrive at the consideration of the important role of intergroup conflict in conspirational narratives, it fails in the proper identification of this intergroup conflict because objects and other inanimate realities which are clearly out of the social framework are also identified as insiders or outsiders.

The importance of detecting intergroup conflict, as proposed by Korenčić et al. [1], relies on the growing and potentially violent participation of conspiratorial groups in political activities. This connection implies that CTs aim to strengthen group cohesion and facilitate coordinated actions [27]. Consequently, detecting crucial aspects of the narrative at the level of span, such as intergroup conflict, can provide significant insights for content moderation.

3. Dataset

This task uses the XAI-DisInfodemic corpus [1], which consists of 10,000 annotated Telegram messages, 5,000 in English and 5,000 in Spanish. These messages contain oppositional, non-mainstream views on the COVID-19 pandemic, and were obtained from public Telegram channels in which users tend to post messages which oppose the mainstream discourse about the pandemic. They are classified into two categories: critical messages and conspiratorial messages. For the creation of this corpus, the authors developed an annotation scheme to differentiate between texts hinting at the existence of a conspiracy and those criticizing mainstream views on COVID-19 but without suggesting the existence of a conspiracy.

Language	Avg.	Std. dev	Min.	Q1	median	Q3	Max.
Spanish	128	123	23	49	98	148	766
English	265	528	12	32	65	266	4,108

Table 1
Statistics of the text length, measured in number of words (whitespace separated tokens), for English and Spanish corpora: the average, the standard deviation, the minimum, the first quartile, the median, the third quartile, and the maximum.

In addition to the annotation into the two classes, the XAI-Disinfodemic corpus offers a second annotation that presents the key elements in oppositional narratives. The tagset includes six labels which can be applied both to messages containing a conspiracy theory and messages containing critical thinking: goals, effects, agents, facilitators (the group that collaborates with the mainstream authorities) and campaigners (the group that conveys the oppositional message).

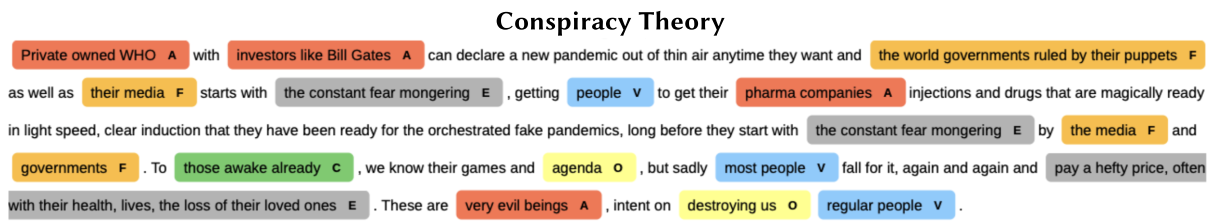

Figure 1: A *Conspiracy* message annotated with elements of oppositional narrative: *Agents* (A), *Facilitators* (F), *Campaigners* (C), *Victims* (V), *Objectives* (O), *Negative Effects* (E).

Korenčić et al. [1] identified the following six categories of narrative elements (see Figure 1 for an example annotation of a *Conspiracy* message, and Figure 2 for an example annotation of a *Critical* message.):

1. *Agents* (A): Those responsible for the actions and/or negative effects described in the comment. In *Conspiracy*, it could be the hidden power that pulls the strings (in Figure 1, "*Private owned WHO*", "*investors like Bill Gates*", "*pharma companies*" and "*very evil beings*"). In *Critical*, it could be the actors that design the mainstream public health policies (in Figure 2, "*White House chief medical*

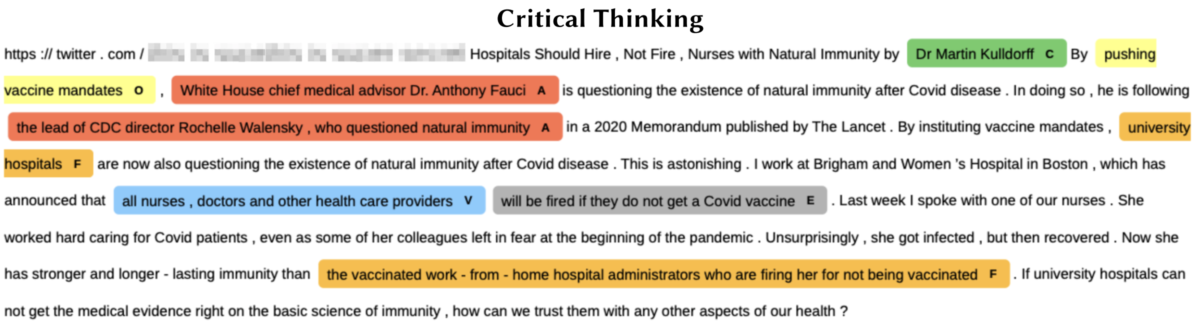

Figure 2: A *Critical* message annotated with elements of oppositional narrative: *Agents* (A), *Facilitators* (F), *Campaigners* (C), *Victims* (V), *Objectives* (O), *Negative Effects* (E).

		A	F	C	V	O	E
ES	All	3,329 (14.0%)	2,688 (11.3%)	4,231 (17.8%)	5,260 (22.2%)	622 (2.6%)	7,150 (30.2%)
	Conspiracy	1,361 (9.8%)	1,184 (8.6%)	2,133 (15.4%)	3,543 (25.6%)	23 (0.2%)	5,326 (38.5%)
	Critical	1,968 (20.0%)	1,504 (15.2%)	2,098 (21.3%)	1,717 (17.4%)	599 (6.1%)	1,824 (18.5%)
EN	All	6,411 (22.4%)	3,462 (12.1%)	6,416 (22.4%)	4,433 (15.5%)	2,073 (7.2%)	5,565 (19.4%)
	Conspiracy	3,333 (21.1%)	1,336 (8.5%)	3,839 (24.4%)	2,734 (17.3%)	615 (3.9%)	3,708 (23.5%)
	Critical	3,078 (23.9%)	2,126 (16.5%)	2,577 (20.0%)	1,699 (13.2%)	1,458 (11.3%)	1,857 (14.4%)

Table 2
Statistics for the gold span-level annotations of the narrative elements. Absolute number and percentage of spans are given for each of the binary text classes and for all texts, and for each of the six narrative categories: *Agents* (A), *Facilitators* (F), *Campaigners* (C), *Victims* (V), *Objectives* (O), *Negative Effects* (E).

advisor Dr. Anthony Fauci" and "*the lead of CDC director Rochelle Walensky, who questioned natural immunity*").

2. *Facilitators* (F): Those who collaborate with the agents and contribute to the execution of their goals. In *Conspiracy*, they could be governments or institutions which, either intentionally or unwittingly, collaborate with the conspirators and help the conspiracy move forward (in Figure 1, "*the world governments ruled by their puppets*", "*their media*", "*the media*" and "*governments*"). In *Critical*, the facilitators could be healthcare workers, mass media or authority figures who abide by governmental instructions (in Figure 2, "*university hospitals*" and "*the vaccinated work - from - home hospital administrators who are firing her for not being vaccinated*").
3. *Campaigners* (C): Those who oppose the mainstream narrative. In *Conspiracy*, those who know the truth and expose it to society at large (in Figure 1, "*those awake already*"). In *Critical*, those who oppose the enforcement of laws and/or refuse to follow health-related instructions from the authorities (in Figure 2, "*Dr Martin Kulldorff*").
4. *Victims* (V): Those who suffer the consequences of the actions and decisions of the agents and/or the facilitators. In *Conspiracy*, the people who are deceived by those in power, and suffer, become ill, lose their freedom, or die as a result of a hidden plan (in Figure 1, "*people*", "*most people*" and "*regular people*"). In *Critical*, the people who receive the negative consequences of the actions and the decisions made by those in power, and also suffer, lose their freedom, become ill, or die as a result of incorrect decisions (in Figure 2, "*all nurses, doctors and other health care providers*").
5. *Objectives* (O): The intentions and purposes that the agents are trying to achieve. In *Conspiracy*, the goals of the conspirators (in Figure 1, "*agenda*" and "*destroying us*"). In *Critical*, the goals of public authorities, pharmaceutical companies, organizations, etc. (in Figure 2, "*pushing vaccine mandates*").
6. *Negative Effects* (E): The negative consequences suffered by the victims as a result of the actions and decisions of those in power and/or their collaborators (in Figure 1, "*the constant fear mongering*"

and "*pay a hefty price, often with their health, lives, the loss of their loved ones*"; in Figure 2, "*will be fired if they do not get a Covid vaccine*").

Table 2 shows the amount and the percentages of spans in the GS that have been annotated with each label for each category (*Conspiracy* or *Critical*).

4. Task Setup

For each language, the corresponding dataset of 5,000 texts was divided into train and test sets using stratified sampling. The train set consisted of 4,000 messages while the test set consisted of 1,000 messages. The participants had access to the train set from the start of the task, and prior to the evaluation deadline they were provided with the unlabeled test set and asked to submit their predictions. Each team was allowed to submit up to two predictions for each combination of subtask and language.

The dataset, the code for building and applying the baseline systems, as well as the evaluation code and task instructions, are made available[2].

Distinguishing Between Critical and Conspiratorial Messages (Subtask 1) This is a binary classification task differentiating between (1) critical messages, i.e. those that question major decisions in the public health domain, but do not promote a conspiracist mentality [1]; and (2) conspiratorial messages, i.e. those that view the pandemic or public health decisions as a result of a malevolent conspiracy by secret, influential groups [1]. Input data consists of a set of messages, each of which associated with one of two categories: either *CONSPIRACY* or *CRITICAL*. The evaluation metric used for this subtask is Matthews Correlation Coefficient (MCC) [28].

Detecting Elements of Oppositional Narratives (Subtask 2) This is a token-level classification task aimed at recognizing text spans corresponding to the key elements of oppositional narratives [1]. Input data consists of a set of messages, each of which is accompanied by a (possibly empty) list of span annotations. Each annotation corresponds to a narrative element, and is described by its borders (start and end characters), as well as its category. There are six distinct span categories: *AGENTS*, *FACILITATORS*, *VICTIMS*, *CAMPAIGNERS*, *OBJECTIVES*, *NEGATIVE_EFFECTS*. The evaluation metric used for this subtask is macro-averaged span-F1 [29].

4.1. Evaluation Measures

As the main criterion for evaluation in Subtask 1 , we used the MCC [28]. MCC serves the same purpose as the macro-averaged F1 measure – it aggregates performance across both classes. We opted for the MCC measure since it works well on imbalanced datasets, while being reliable and less optimistic than the macro-averaged F1 [30], and comparing favorably to other alternatives [28].

For evaluation in Subtask 2 , we used the span-F1 measure [29], which is an adapted version of the F1 measure and accounts for partially correct predictions by looking at span overlap. Specifically, a predicted span is not required to exactly match a gold standard span in terms of start and end characters. Instead, the proportion of overlapping characters is used to calculate precision and recall [29]. This approach offers a fairer evaluation in tasks with long spans, and with inherent subjectivity of the span boundaries. For tasks like traditional, non-nested Named Entity Recognition (NER), where named entities are shorter and are expected to have well-defined boundaries, exact matching is a reasonable method of evaluation.

As the main criterion for evaluation we used macro-averaged span-F1, i.e., span-F1 averaged over all six span labels corresponding to six elements of oppositional narratives described in Section 3.

[2]https://github.com/dkorenci/pan-clef-2024-oppositional

4.2. Baseline Solutions

Baselines for both subtasks are based on the approaches from Korenčić et al. [1], where more details can be found. For each subtask, we took as a baseline the version based on the transformer model which resulted in the lowest performance in Korenčić et al. [1]. Hyperparameters were not changed, the models were trained on the entire train set, and then applied to the test set.

Distinguishing Critical and Conspiratorial Messages (Subtask 1) The approach for this binary classification task is based on fine-tuning the BERT transformer model [31] from the Hugging Face[3] repository, using the case-sensitive "base" version. The BETO [32] version of BERT was used for the Spanish dataset. The number of tokens was set to 256. We tuned the models for three epochs using the AdamW optimizer, learning rate of $2e^{-5}$, slanted triangular LR scheduler with a 10% warm-up period, a batch size of 16, and a weight decay of 0.01. All the layers of the transformers were fine-tuned. The dropout rate for the classification head was 0.1.

Detecting Elements of Oppositional Narratives (Subtask 2) The baseline for this sequence labeling task is based on fine-tuning a transformer model with added token classification heads. To account for the possibility of overlapping spans with different categories, we used six separate per-category heads that performed BIO sequence tagging. We employed multi-task learning [33] by connecting the per-category taggers to the same transformer backbone. Multi-task learning has several advantages, such as improved regularization and implicit data augmentation [33], and the described approach was successfully deployed for a similar task of span-level skill extraction [34]. We used the same configuration and hyperparameters as in the case of Subtask 1 . The exception was the number of epochs, which we increased to 10 in order to accommodate for the increased task complexity. The BERT model [31] was used as the base transformer for the English dataset, while for the Spanish dataset the BETO version of BERT [32] was used.

5. Participating Systems

A total of 82 teams submitted their solution for at least one of the tasks. The approaches included pre-neural NLP models, small transformers such as BERT [31], and Large Language Models [35]. Techniques such as Ensemble Methods [36] and Data Augmentation [37] were also used to improve performance. Another important factor was the data on which the chosen transformer models were pretrained – participants experimented with both domain-specific models such as CT-BERT [18] and multilingual models such as mBERT [38].

Most of the approaches relied on fine-tuning BERT-like transformers [31]. This is not surprising since these models yield strong results for both classification [31] and sequence labeling [31], and since baselines based on this approach were provided to the participants.

To describe the approaches based on transformer models [39] we shall use the abbreviation *SLM ("Small" Language Models)* to describe transformers with fewer than one billion parameters. For the transformers with more than one billion parameters, we shall use the standard abbreviation *LLM (Large Language Models).*

Working Notes Submissions A total of 17 participating systems had their working notes papers accepted. Huertas-García et al. [40] tackled Subtask 1 , experimenting with a range of SLMs and with the commercial LLM Claude[4]. Vallecillo-Rodríguez et al. [41] experimented with the fine-tuning of two LLMs: LLaMA3-8B-instruct [42] and GPT-3.5 [43]. Hu et al. [44] used SLMs with an added BiGRU LSTM layer [45] to tackle both tasks. Damian et al. [46] approached both tasks using ensembles of mono- and multi-lingual SLMs. Sánchez-Hermosilla et al. [47] focused on Subtask 1 using a range of SLMs, data

[3]https://huggingface.co/models
[4]https://www.anthropic.com/claude

augmentation, and ensembling techinques. Zrnić [48] experimented with mono- and multilingual SLMs in order to tackle both tasks. Sahitaj et al. [49] approached Subtask 1 using SLMs and a LLM-based data augmentation technique. Gómez-Romero et al. [50] used an approach based on OpenAI Embeddings and a deep feedforward network for Subtask 1 and, in addition, did entity masking in order to increase the models' generality. Mahesh et al. [51] experimented with SLMs and non-neural approaches on Subtask 1 . Zeng et al. [52] employed mono- and multi-lingual SLMs for both Subtask 1 and Subtask 2 . Huang et al. [53] used SLMs for both tasks, and employed ensembling for Subtask 1 . Tulbure and Coll Ardanuy [54] experimented with SLMs boosted by data augmentation and ensembling, and for Subtask 2 split the input texts into sentences. Liu et al. [55] experimented with a range of LLMs using zero-shot chain-of-thoughts prompts to tackle Subtask 1 , and used a SLM approach for Subtask 2 . Mhalgi et al. [56] approached Subtask 1 using data augmentation, non-neural classifiers, SLMs and LLMs, as well as model ensembles.

Several participants basically repeated what had been done in the baseline solution, i.e., fine-tuned and applied one or several SLMs [57, 58, 59].

Teams that did not submit working notes accounted for 65 submissions and provided a short description of their approaches. Many of these submissions were minor modifications of the provided baseline, i.e., changing of an SLM to be fine-tuned. However, a number of these teams achieved competitive results or provided useful datapoints using, for example, ensembling techniques, data and feature augmentation techniques, and non-neural NLP approaches.

6. Results and Analysis

6.1. Distinguishing Critical and Conspiracy Texts (Subtask 1)

Table 6.1 displays the results of the most successful teams on Subtask 1 – the teams with performance equal to or greater than the provided baseline.

English		Spanish	
TEAM	MCC	TEAM	MCC
IUCL [56]	0.8388	SINAI [41]	0.7429
AI_Fusion	0.8303	auxR	0.7205
SINAI [41]	0.8297	RD-IA-FUN [40]	0.7028
ezio [44]	0.8212	Elias&Sergio	0.6971
hinlole [53]	0.8198	AI_Fusion	0.6872
Zleon [48]	0.8195	zhengqiaozeng [52]	0.6871
virmel	0.8192	virmel	0.6854
inaki [47]	0.8149	trustno1	0.6848
yeste	0.8124	Zleon [48]	0.6826
auxR	0.8088	ojo-bes	0.6817
Elias&Sergio	0.8034	tulbure [54]	0.6722
theateam	0.8031	sail [50]	0.6719
trustno1	0.7983	nlpln [55]	0.6681
DSVS [46]	0.7970	*baseline-BETO*	*0.6681*
ojo-bes	0.7969		
sail [50]	0.7969		
RD-IA-FUN [40]	0.7965		
baseline-BERT	*0.7964*		

Table 3
Performance of top teams, in terms of Matthews Correlation Coefficient (MCC), on Subtask 1 – binary classification of text as either conspiracy or critical.

Results for English The top IUCL team [56] employed the DeBERTa model [60] fine-tuned on an augmented dataset comprising the Subtask 1 dataset and the conspiracy-labeled examples from the LOCO corpus [11] (cca. 16,000 examples were selected). The AI_Fusion team came a close second, simply by relying on the fine-tuned ELECTRA model [61]. A close third was the SINAI team [41], which used the fine-tuned LLaMA3-8B-instruct LLM [42] as a solution. Additionally, their experiments demonstrated that fine-tuned LLMs outperform the LLM-based zero-shot approaches by a large margin [41].

The rest of the top-performing models on English based their approaches on SLMs, with several teams using techniques such as ensembling and data augmentation. The Covid-twitter-BERT [18], used by the teams ezio [44], hinlole [53], Zleon [48], and inaki [47], seems to be a successful transformer model for this use-case. Some teams with competitive results used standard transformer models: the theateam, trustno1, and ojo-bes teams used standard RoBERTa [62], while the virmel team used BERT [31] and the yeste team relied on the ELECTRA model [61].

Two fully multilingual approaches performed competitively, those of the auxR and RD-IA-FUN [40] teams. Both approaches were based on a multilingual transformer trained on joint English and Spanish data. The auxR team employed the Twitter-XLM-RoBERTa-large model, a derivative of the XLM-RoBERTa model [63] domain-adapted using Twitter data, while the RD-IA-FUN [40] team used the multilingual-e5-large model [64], a derivative of XLM-RoBERTa. The Elias&Sergio team used monolingual RoBERTa, but fine-tuned the model using the Spanish dataset translated to English (in addition to the English dataset).

Notably different was the approach of the sail team [50], who used OpenAI Embeddings[5] in combination with a deep feed-forward neural network for fine-tuning. Additionally, they pre-processed the texts by replacing named entities with entity classes such as 'PERSON', in order to "enhance the model's generalization capabilities" [50]. They showed that, for Subtask 1 , the masked model performs better than the non-masked one.

Results for Spanish Many of the teams that did well on Spanish also achieved top results on English. For these teams, we will briefly describe the differences between the two approaches, and we refer the reader to the English section of Subtask 1 for details.

Top performance was obtained by the SINAI team [41], which relied on LLMs. In contrast to what happened in English, the fine-tuned GPT-3.5 model [43] outperformed LLaMA3-8B-instruct [42] by a large margin, yielding the best overall solution.

The second and third positions are held by the two fully multilingual approaches of the auxR and RD-IA-FUN teams [40], which also performed well on English.

Interestingly, five out of the six following teams (Elias&Sergio, AI_Fusion, zhengqiaozeng, virmel, trustno1, Zleon) employed standard SLM fine-tuning with PlanTL-GOB-ES/roberta-base-bne [65] as the base model. The exception is the zhengqiaozeng team [52], which relied on the multilingual XLM-RoBERTa model. The tulbure team [54] relied on an ensemble of three Spanish SLMs.

The sail team [50] used the same approach as for English, based on multilingual OpenAI Embeddings.

The nlpln team [55] made it over the baseline using an unconventional approach in the context of this challenge - zero-shot prompting based on LLMs and the chain-of-thought prompting technique [66]. We note that the same approach scored competitively on the English classification subtask, achieving an MCC of 0.7844 (see Table A). The nlpln team [55] tested a number of LLMs, including GPT, Claude, and Gemini, on the full training set. The DeepSeek V2 model [67], a large mixture-of-experts LLMs, achieved the best results. Surprisingly, the results on the test data proved this model to be relatively competitive with fine-tuned LLMs.

Analysis The results of the top teams suggest that the most successful English transformer-based models are the DeBERTa model [60], the ELECTRA model [61] and the large LLaMA3-8B-instruct LLM [42]. The Covid-twitter-BERT [18] model was used by a number of high-performing teams, suggesting

[5]https://platform.openai.com/docs/guides/embeddings

that pre-training on social media data probably influences performance. However, both BERT [31] and RoBERTa [62] were shown to be able to perform competitively. The performance edge obtained by the IUCL team [56] suggests that the LOCO conspiracy corpus [11] is a useful resource for boosting conspiracy-related classifiers for other use-cases.

In Spanish, the choice of a model seems to be more important, and many of the best teams used the Spanish 'Maria' RoBERTa model [65], trained exclusively on the data crawled from the web, while none of the top teams employed either the BETO [32] or BERTIN [68] models. Moreover, the top three teams employed either fine-tuned LLMs [41] (GPT-3.5 [43]) or multilingual models [40, 63]. These teams, especially the top one based on LLMs, outperformed the others by a significant margin. Interestingly, none of the participants used RoBERTuito [69], a model pretrained on Spanish social media text.

It would be interesting to perform ablation studies in both languages in order to measure the influence of both architectural improvements and the choice of the pretraining dataset on performance.

As for the application of the LLMs [35], the results on English show no big difference between fine-tuned LLMs and fine-tuned SLMs. Therefore, we hypothesize that the superiority of fine-tuned GPT-3.5 [43] on Spanish is due to the pre-training data (GPT-3.5 has probably "seen" much more texts from then social media then the Spanish SLMs). The results of the nlpln team [55] demonstrate the competitiveness, in both languages, of the DeepSeek V2 model [67], in combination with chain-of-thoughts prompting [66]. Therefore, this approach seems to be a good way to quickly bootstrap a conspiracy vs. critical classifier for other use-cases and other supported languages. The approach of Sahitaj et al. [49], which was based on using LLM-based elaboration on text's context and argumentation as additional input for classification, might prove beneficial for improving LLM-based zero-shot prompting.

A number of teams opted to use non-neural text classifiers, such as LinearSVM [70] or Random Forest [71] in combination with tf-idf- or n-gram-based features. The average score of these approaches is 0.7080 MCC for English, and 0.5814 MCC for Spanish.

The baseline systems [1] were based on BERT [31] and BETO [32], respectively, for the English and Spanish dataset. These models were chosen as the baseline as they yielded the weakest performance in Korenčić et al. [1]. The best performance, corresponding to the state-of-art before this challenge, was obtained for DeBERTaV3 [72] and 'BERTIN' RoBERTa [68] models. When these models were applied to the train-test split of the challenge, the MCC scores of 0.8259 and 0.6681 were obtained, respectively, for English and Spanish. The score of DeBERTaV3 represents an improvement in relation to BERT. Even with this improvement, the participants managed to improve upon the state-of-art performance.

6.2. Detecting Elements of the Oppositional Narratives (Subtask 2)

Table 6.2 contains the results of the most successful teams on Subtask 2 – the teams with performance equal to or greater than that of the provided baseline.

Results for English The most successful team, tulbure [54], relied on a combination of preprocessing techniques and data augmentation. While the provided baseline used multi-task learning to account for overlapping spans of different categories [1], Tulbure and Coll Ardanuy [54] opted to use a single model for all the span categories and modified the data accordingly. Additionally, each Telegram text was segmented into sentences which were used as examples for learning. This solved the problem of texts longer than the maximum length supported by a transformer. Data augmentation was performed by "replacing words in the texts by synonyms or semantically-related words", and the RoBERTa model was used [62].

As the remaining teams mostly relied on modifying the multi-task sequence labeling approach of the baseline [1], this will be the assumed default approach. Only if another approach was used will the difference be described.

The second-placed team, Zleon [48], used a large variant of RoBERTa [62] and increased the model's maximum sequence length to 512. The third-placed team, hinlole [53], used Covid-twitter-BERT [18] as the base model.

English TEAM	span-F1	Spanish TEAM	span-F1
tulbure [54]	0.6279	tulbure [54]	0.6129
Zleon [48]	0.6089	Zleon [48]	0.5875
hinlole [53]	0.5886	AI_Fusion	0.5777
oppositional_opposition	0.5866	virmel	0.5616
AI_Fusion	0.5805	CHEEXIST	0.5621
virmel	0.5742	miqarn	0.5603
miqarn	0.5739	DSVS [46]	0.5529
TargaMarhuenda	0.5701	TargaMarhuenda	0.5364
ezio [44]	0.5694	Elias&Sergio	0.5151
zhengqiaozeng [52]	0.5666	hinlole [53]	0.4994
Elias&Sergio	0.5627	*baseline-BETO*	*0.4934*
DSVS [46]	0.5598		
CHEEXIST	0.5524		
rfenthusiasts	0.5479		
ALC-UPV-JD-2	0.5377		
baseline-BERT	*0.5323*		

Table 4
Performance of top teams, in terms of span-F1 metric [29] (macro-averaged over span labels), on Subtask 2 – token classification of span-level narrative elements.

The oppositional_opposition team used the DistilBERT model [73] in combination with Conditional Random Fields [74]. Interestingly, the same type of model was used for Subtask 2 in Spanish, but achieved a very low result (see Table 10 in Appendix A), as if overfitting or failing to converge. The AI_Fusion team used the RoBERTa model [62] and chose the best model over the 50 fine-tuning epochs. The virmel team used the RoBERTa model with the maximum sequence length set to 512. The zhengqiaozeng team [52] employed the RoBERTa model, while the ALC_UPV_JD_2 team relied on the small ALBERT model [75].

The miqarn team used the multilingual mBERT model [38], trained on datasets in both languages. This approach also performed well on the Spanish dataset.

The TargaMarhuenda team used the RoBERTa model, and added pre-computed POS tags as input by concatenating them to the model's token embeddings to construct input to the initial layer of the transformer. The Elias&Sergio team used a similar approach, but concatenated one-hot POS vectors with the token representations of the final layer of the transformer to construct input to the token classification head.

The ezio team [44] modified the multi-tasking approach using "BiGRU LSTM", a bidirectional LSTM network based on gated recurrent units [45]. Instead of using simple per-task classification heads, each task was assigned both a task-specific LSTM network and a task-specific classification head. Covid-twitter-BERT [18] was used as the base model.

The DSVS [46] team created an ensemble of token classifiers based on different SLMs such as BERT, RoBERTa and ELECTRA, and performed "logit averaging" to obtain their final predictions.

The CHEEXIST team used the Fake-News-Bert-Detect model, a domain-adapted version of RoBERTa. Additionally, they replaced the final classification layer with a shallow neural network.

The rfenthusiasts team used the DeBERTaV3 model [72] and did a data augmentation by replacing characters in text. The same approach, when used in combination with the XLM-RoBERTa model [63], did not work well on the Spanish dataset.

Results for Spanish All of the teams that achieved top results on the Spanish dataset did the same on the English dataset. Therefore, here we will only briefly describe the differences, which mostly pertain to a different choice of transformer model. Similarly as for English, the majority of the approaches relied on the multi-task sequence labeling approach of the baseline [1].

The same two teams - tulbure and Zleon - took the first and second place, as on the English dataset. Both relied on the same respective approach that they used on English, with the difference of using the Spanish 'Maria' RoBERTa model [65].

The AI_Fusion team, placed third, relied on the XLM-RoBERTa model [63], while the virmel team relied on Spanish 'BERTIN' RoBERTa model [68]. The CHEEXIST team used the 'Maria' RoBERTa model [65].

The miqarn team used a single mBERT [38] model fine-tuned on both datasets, and achieved good results on Spanish. The DSVS [46] team's ensemble approach also achieved good results in the case of the Spanish dataset. The ensemble consisted of a number of Spanish and multilingual models [46].

Two approaches based on using POS tags as additional input to the model, used by the Targa-Marhuenda and Elias&Sergio teams, relied on the Spanish RoBERTa model. The hinlole team [53] relied on the Spanish BETO model [32].

Analysis The system that clearly outperformed the others in both languages was the one of the tulbure team [54]. Its sentence-level processing of texts shows that signals for the inference of the elements of oppositional narrative are largely sentence-local. It would be interesting to perform ablation studies to determine how much data augmentation influences performance in contrast to sentence segmenting. Further improvements might be achieved by way of using multi-task learning and transformers other than RoBERTa, as well as other data augmentation techniques, possibly based on LLMs.

The competitive results of the Zleon team [48] and several other teams relying on the multi-task baseline approach show its effectiveness in combination with an improved choice of the backbone SLM and increased maximum sequence length. Covid-twitter-BERT [18], used by the second- and third-placed teams, seems to be a successful choice for English.

The performance of Subtask 2 seems to be less influenced by the choice of the transformer model, especially in the case of Spanish. Concretely, a larger variety of models appear among the top teams and, in the case of Spanish, all three families of models (BETO [32], BERTIN [68], and 'Maria' [65]) are represented.

The approach of the miqarn team, based on the multilingual mBERT model [38], worked well for both languages and could be a good approach for the task of inferring the elements of oppositional narrative in other languages, especially under-resourced ones.

The baseline systems [1] were based on BERT [31] and BETO [32] models, respectively, for the English and Spanish dataset. They were chosen since they yielded the weakest performance in Korenčić et al. [1]. Top performance, corresponding to the state-of-art before this challenge, was obtained for DeBERTaV3 [72] and BERTIN [68] models. When these models were applied to the train-test split of the challenge, the MCC scores of 0.5786 and 0.5369 were obtained, respectively, for English and Spanish. These scores represent an improvement in relation to the baseline, but even so the participants managed to significantly raise the state-of-art performance on the task.

7. Conclusions

The Oppositional Thinking Analysis PAN Task presented to the NLP community two subtasks: distinguishing between critical and conspiratorial messages, and detecting elements of oppositional narratives. These subtasks are of interest to computational social scientists interested in text-based analysis of oppositional thinking [1].

A total of 82 teams participated in the challenge, while 17 teams provided working notes papers. The teams devised a range of solutions, the most successful of which exceeded previous state-of-the-art [1] for both subtasks. The new solutions have the potential to facilitate researchers in applying the domain-agnostic annotation schemes proposed in Korenčić et al. [1] to new corpora.

For Subtask 1 the most successful submitted English system [56] relied on augmentation using the large news conspiracy corpus LOCO [11]. The best result for Spanish was achieved using a fine-tuned GPT-3.5 [41]. The multilingual approach of Huertas-García et al. [40] also proved competitive. An

LLM-based zero-shot approach of Liu et al. [55] achieved results competitive with supervised baselines on Subtask 1 and demonstrated a cost-effective way to bootstrap conspiracy vs. critical classifiers for new use-cases. The experiments also point to the need to create better small-scale transformer models for Spanish, as the solutions that work best on the Spanish dataset rely either on LLMs, or on multilingual SLMs.

For Subtask 2, the top system in both languages relied on a combination of data augmentation by word replacement and sentence-level processing [54]. Most of the other systems relied on improving the provided baseline solution by changing the underlying transformer model, or by modifying the training procedure.

There are many possible directions for creating even better-performing systems. Crafting new domain-specific SLMs would probably be beneficial, as demonstrated by the effectiveness of Covid-twitter-BERT [18] on both subtasks. Having in mind the difficulty of creating high-quality annotated data, further work on the LLM-based zero- and few-shot approaches would be beneficial for practitioners. Similarly, multi-lingual approaches adaptable to new languages with few annotated examples [76] would also be an interesting and potentially effective direction to pursue. If the topic-agnostic annotation scheme [1] used for this task is applied to create new labeled corpora, it would be interesting to use these corpora for benchmarking the approach of Gómez-Romero et al. [50], which focuses on the generalization capabilities of the models.

Acknowledgments

The shared task on Oppositional Thinking Analysis was organised in the framework of the XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (MICIN PLEC2021-007681), funded by MCIN/AEI/ 10.13039/501100011033 and by European Union NextGenerationEU/PRTR. The work of Damir Korenčić and Berta Chulvi was conducted while at Universitat Politècnica de València.

References

[1] D. Korenčić, B. Chulvi, X. Bonet, T. Mariona, A. Toselli, P. Rosso, What distinguishes conspiracy from critical narratives? A computational analysis of oppositional discourse, Expert Systems (2024). doi:10.1111/exsy.13671.

[2] K. M. Douglas, R. M. Sutton, What are conspiracy theories? A definitional approach to their correlates, consequences, and communication, Annual Review of Psychology 74 (2023) 271–298. URL: https://doi.org/10.1146/annurev-psych-032420-031329.

[3] H. Tajfel, J. C. Turner, An integrative theory of intergroup relations, Psychology of intergroup relations (1979) 33–47.

[4] R. Brown, Social identity theory: past achievements, current problems and future challenges, European Journal of Social Psychology 30 (2000) 745–778. doi:10.1002/1099-0992(200011/12)30:6<745::AID-EJSP24>3.0.CO;2-O.

[5] M. A. Hogg, Social identity theory (2016). doi:10.1007/978-3-319-29869-6_1.

[6] R. M. Sutton, K. M. Douglas, Rabbit hole syndrome: Inadvertent, accelerating, and entrenched commitment to conspiracy beliefs, Current Opinion in Psychology 48 (2022) 101462. URL: https://www.sciencedirect.com/science/article/pii/S2352250X2200183X. doi:https://doi.org/10.1016/j.copsyc.2022.101462.

[7] E. Funkhouser, A tribal mind: Beliefs that signal group identity or commitment, Mind & Language 37 (2022) 444–464. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mila.12326. doi:https://doi.org/10.1111/mila.12326. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/mila.12326.

[8] B. Franks, A. Bangerter, M. W. Bauer, M. Hall, M. C. Noort, Beyond “monologicality”? exploring

conspiracist worldviews, Frontiers in Psychology 8 (2017). URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00861. doi:10.3389/fpsyg.2017.00861.

[9] D. Mahl, M. S. Schäfer, J. Zeng, Conspiracy theories in online environments: An interdisciplinary literature review and agenda for future research, New Media & Society 0 (2022) 14614448221075759. URL: https://doi.org/10.1177/14614448221075759. doi:10.1177/14614448221075759. arXiv:https://doi.org/10.1177/14614448221075759.

[10] J. E. Uscinski, J. Parent, B. Torres, Conspiracy Theories are for Losers, 2011. URL: https://papers.ssrn.com/abstract=1901755, aPSA 2011 Annual Meeting Paper.

[11] A. Miani, T. Hills, A. Bangerter, Loco: The 88-million-word language of conspiracy corpus, Behavior research methods (2021) 1–24.

[12] J. Langguth, D. T. Schroeder, P. Filkuková, S. Brenner, J. Phillips, K. Pogorelov, Coco: an annotated twitter dataset of covid-19 conspiracy theories, Journal of Computational Social Science (2023) 1–42.

[13] K. Pogorelov, D. T. Schroeder, S. Brenner, J. Langguth, FakeNews: Corona Virus and Conspiracies Multimedia Analysis Task at MediaEval 2021, in: Working Notes Proceedings of the MediaEval 2021 Workshop Bergen, Norway and Online, 2021.

[14] K. Pogorelov, D. T. Schroeder, S. Brenner, A. Maulana, J. Langguth, Combining tweets and connections graph for fakenews detection at mediaeval 2022, in: Proceedings of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12-13 January 2023., 2023.

[15] Y. Peskine, G. Alfarano, I. Harrando, P. Papotti, R. Troncy, Detecting covid-19-related conspiracy theories in tweets, in: MediaEval 2021, MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop, 13-15 December 2021, 2021.

[16] Y. Peskine, P. Papotti, R. Troncy, Detection of COVID-19-Related Conpiracy Theories in Tweets using Transformer-Based Models and Node Embedding Techniques, in: Working Notes Proceedings of the MediaEval 2022 Workshop Bergen, Norway and Online, 2023.

[17] D. Korenčić, I. Grubišić, A. H. Toselli, B. Chulvi, P. Rosso, Tackling Covid-19 Conspiracies on Twitter using BERT Ensembles, GPT-3 Augmentation, and Graph NNs, in: Working Notes Proceedings of the MediaEval 2022 Workshop Bergen, Norway and Online, 2023. URL: https://2022.multimediaeval.com/paper8969.pdf.

[18] M. Müller, M. Salathé, P. E. Kummervold, Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter, Frontiers in Artificial Intelligence 6 (2023). URL: https://www.frontiersin.org/articles/10.3389/frai.2023.1023281. doi:10.3389/frai.2023.1023281.

[19] A. Giachanou, B. Ghanem, P. Rosso, Detection of conspiracy propagators using psycho-linguistic characteristics, Journal of Information Science 49 (2021) 3–17. doi:10.1177/0165551520985486.

[20] J. D. Moffitt, C. King, K. M. Carley, Hunting conspiracy theories during the covid-19 pandemic, Social Media + Society 7 (2021). doi:10.1177/20563051211043212.

[21] A. Bessi, Personality traits and echo chambers on facebook, Computers in Human Behavior 65 (2016) 319–324. URL: https://www.sciencedirect.com/science/article/pii/S0747563216305817. doi:10.1016/j.chb.2016.08.016.

[22] C. Klein, P. Clutton, V. Polito, Topic Modeling Reveals Distinct Interests within an Online Conspiracy Forum, Frontiers in Psychology 9 (2018). URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00189.

[23] M. Samory, T. Mitra, 'The Government Spies Using Our Webcams': The Language of Conspiracy Theories in Online Discussions, Proceedings of the ACM on Human-Computer Interaction 2 (2018) 1–24. URL: https://dl.acm.org/doi/10.1145/3274421. doi:10.1145/3274421.

[24] S. Levy, M. Saxon, W. Y. Wang, Investigating Memorization of Conspiracy Theories in Text Generation, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 4718–4729. URL: https://aclanthology.org/2021.findings-acl.416. doi:10.18653/v1/2021.findings-acl.416.

[25] J. Introne, A. Korsunska, L. Krsova, Z. Zhang, Mapping the Narrative Ecosystem of Conspiracy Theories in Online Anti-vaccination Discussions, in: International Conference on Social Media and Society, Association for Computing Machinery, 2020, pp. 184–192. URL: https://dl.acm.org/

doi/10.1145/3400806.3400828. doi:10.1145/3400806.3400828.
[26] P. Holur, T. Wang, S. Shahsavari, T. Tangherlini, V. Roychowdhury, Which side are you on? Insider-Outsider classification in conspiracy-theoretic social media, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4975–4987. URL: https://aclanthology.org/2022.acl-long.341. doi:10.18653/v1/2022.acl-long.341.
[27] P. Wagner-Egger, A. Bangerter, S. Delouvée, S. Dieguez, Awake together: Sociopsychological processes of engagement in conspiracist communities, Current Opinion in Psychology 47 (2022) 101417. URL: https://www.sciencedirect.com/science/article/pii/S2352250X22001385. doi:https://doi.org/10.1016/j.copsyc.2022.101417.
[28] 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) 13. URL: https://doi.org/10.1186/s13040-021-00244-z. doi:10.1186/s13040-021-00244-z.
[29] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, Fine-Grained Analysis of Propaganda in News Articles, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5636–5646. URL: https://aclanthology.org/D19-1565. doi:10.18653/v1/D19-1565.
[30] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (2020) 6. URL: https://doi.org/10.1186/s12864-019-6413-7. doi:10.1186/s12864-019-6413-7.
[31] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.
[32] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish Pre-trained BERT Model and Evaluation Data, 2023. URL: http://arxiv.org/abs/2308.02976. arXiv:2308.02976, arXiv:2308.02976.
[33] S. Ruder, An Overview of Multi-Task Learning in Deep Neural Networks, 2017. URL: http://arxiv.org/abs/1706.05098, arXiv:1706.05098.
[34] M. Zhang, K. Jensen, S. Sonniks, B. Plank, SkillSpan: Hard and soft skill extraction from English job postings, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 4962–4984. URL: https://aclanthology.org/2022.naacl-main.366. doi:10.18653/v1/2022.naacl-main.366.
[35] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models, 2023. URL: https://arxiv.org/abs/2303.18223. arXiv:2303.18223.
[36] T. G. Dietterich, Ensemble methods in machine learning, in: Multiple Classifier Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2000, pp. 1–15.
[37] C. Shorten, T. M. Khoshgoftaar, B. Furht, Text data augmentation for deep learning, Journal of big Data 8 (2021) 101.
[38] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual BERT?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4996–5001. URL: https://aclanthology.org/P19-1493. doi:10.18653/v1/P19-1493.
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/

3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[40] Á. Huertas-García, C. Martí-González, J. Muñoz, E. Ambite, Small Language Models and Large Language Models in Oppositional thinking analysis: Capabilities and Biases and Challenges, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[41] M. Vallecillo-Rodríguez, M. Martín-Valdivia, A. Montejo-Ráez, SINAI at PAN 2024 Oppositional Thinking Analysis: Exploring the fine-tuning performance of LLMs, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[42] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.
[43] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.
[44] Q. Hu, Z. Han, J. Peng, M. Guo, C. Liu, An Oppositional Thinking Analysis Method Using BERT-based Model with BiGRU, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[45] K. Cho, B. van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder–decoder approaches, in: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014, pp. 103–111.
[46] S. Damian, B. Herrera-Gonzalez, D. Vazquez-Santana, H. Calvo, E. Felipe-Riverón, C. Yáñez-Márquez, DSVS at PAN 2024: Ensemble Approach of Large Language Models for Analyzing Conspiracy Theories Against Critical Thinking Narratives, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[47] I. Sánchez-Hermosilla, A. Panizo Lledot, D. Camacho, A Study on NLP Model Ensembles and Data Augmentation Techniques for Separating Critical Thinking from Conspiracy Theories in English Texts, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[48] L. Zrnić, Conspiracy theory detection using transformers with multi-task and multilingual approaches, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[49] A. Sahitaj, P. Sahitaj, S. Mohtaj, S. Möller, V. Schmitt, Towards a Computational Framework for Distinguishing Critical and Conspiratorial Texts by Elaborating on the Context and Argumentation with LLMs, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[50] J. Gómez-Romero, S. González-Silot, A. Montoro-Montarroso, M. Molina-Solana, E. Martínez Cámara, Detection of conspiracy-related messages in Telegram with anonymized named entities, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[51] S. Mahesh, S. Divakaran, K. Girish, S. Lakshmaiah, Binary Battle: Leveraging ML and TL Models to Distinguish between Conspiracy Theories and Critical Thinking, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[52] Z. Zeng, Z. Han, J. Ye, Y. Tan, H. Cao, Z. Li, R. Huang, A Conspiracy Theory Text Detection Method based on RoBERTa and XLM-RoBERTa Models, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum,

CEUR-WS.org, 2024.
[53] J. Huang, Z. Han, R. Zhu, M. Guo, K. Sun, Conspiracy Theory Text Classification Based on CT-BERT and BETO Models, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[54] A. Tulbure, M. Coll Ardanuy, Conspiracy vs critical thinking using an ensemble of transformers with data augmentation techniques, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[55] B. Liu, Z. Han, H. Cao, An Approach to Classifying Conspiratorial and Critical Public Health Narratives, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[56] S. Mhalgi, S. Pulipaka, S. Kübler, IUCL at PAN 2024: Using Data Augmentation for Conspiracy Theory Detection, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[57] P. Balasundaram, K. Swaminathan, O. Sampath, P. Km, Oppositional Thinking Analysis: Conspiracy Theories vs Critical Thinking Narratives, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[58] A. Albladi, C. Seals, Detection of Conspiracy vs. Critical Narratives and Their Elements using NLP, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[59] D. Espinosa, G. Sidorov, E. Ricárdez-Vázquez, Using BERT to Identify Conspiracy Theories, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.
[60] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.
[61] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, 2020. URL: https://arxiv.org/abs/2003.10555. arXiv:2003.10555.
[62] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692, arXiv:1907.11692.
[63] 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. URL: https://arxiv.org/abs/1911.02116. arXiv:1911.02116.
[64] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A technical report, 2024. URL: https://arxiv.org/abs/2402.05672. arXiv:2402.05672.
[65] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural (2022) 39–60. URL: https://doi.org/10.26342/2022-68-3. doi:10.26342/2022-68-3.
[66] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.
[67] DeepSeek-AI, A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Xu, H. Yang, H. Zhang, H. Ding, H. Xin, H. Gao, H. Li, H. Qu, J. L. Cai, J. Liang, J. Guo, J. Ni, J. Li, J. Chen, J. Yuan, J. Qiu, J. Song, K. Dong, K. Gao, K. Guan, L. Wang, L. Zhang, L. Xu, L. Xia, L. Zhao, L. Zhang, M. Li, M. Wang, M. Zhang, M. Zhang, M. Tang, M. Li, N. Tian, P. Huang, P. Wang, P. Zhang, Q. Zhu, Q. Chen, Q. Du, R. J. Chen, R. L. Jin, R. Ge, R. Pan, R. Xu, R. Chen, S. S. Li, S. Lu, S. Zhou, S. Chen, S. Wu, S. Ye, S. Ma, S. Wang, S. Zhou, S. Yu, S. Zhou, S. Zheng, T. Wang, T. Pei, T. Yuan, T. Sun, W. L. Xiao, W. Zeng, W. An, W. Liu, W. Liang, W. Gao, W. Zhang, X. Q. Li, X. Jin, X. Wang, X. Bi,

X. Liu, X. Wang, X. Shen, X. Chen, X. Chen, X. Nie, X. Sun, X. Wang, X. Liu, X. Xie, X. Yu, X. Song, X. Zhou, X. Yang, X. Lu, X. Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zhao, Y. Sun, Y. Li, Y. Wang, Y. Zheng, Y. Zhang, Y. Xiong, Y. Zhao, Y. He, Y. Tang, Y. Piao, Y. Dong, Y. Tan, Y. Liu, Y. Wang, Y. Guo, Y. Zhu, Y. Wang, Y. Zou, Y. Zha, Y. Ma, Y. Yan, Y. You, Y. Liu, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Huang, Z. Zhang, Z. Xie, Z. Hao, Z. Shao, Z. Wen, Z. Xu, Z. Zhang, Z. Li, Z. Wang, Z. Gu, Z. Li, Z. Xie, Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024. URL: https://arxiv.org/abs/2405.04434. arXiv:2405.04434.

[68] J. D. l. Rosa, E. G. Ponferrada, M. Romero, P. Villegas, P. G. d. P. Salas, M. 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, number: 0.

[69] J. M. Pérez, D. A. Furman, L. A. Alemany, F. Luque, Robertuito: a pre-trained language model for social media text in spanish, 2022. URL: https://arxiv.org/abs/2111.09453. arXiv:2111.09453.

[70] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: C. Nédellec, C. Rouveirol (Eds.), Machine Learning: ECML-98, Springer Berlin Heidelberg, Berlin, Heidelberg, 1998, pp. 137–142.

[71] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[72] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, in: International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=sE7-XhLxHA.

[73] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL: https://arxiv.org/abs/1910.01108. arXiv:1910.01108.

[74] J. D. Lafferty, A. McCallum, F. C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001, p. 282–289.

[75] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, 2020. URL: https://arxiv.org/abs/1909.11942. arXiv:1909.11942.

[76] F. D. Schmidt, I. Vulić, G. Glavaš, Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10725–10742. URL: https://aclanthology.org/2022.emnlp-main.736. doi:10.18653/v1/2022.emnlp-main.736.

TASK 1 - ENGLISH POSITION	TEAM	MCC	F1-MACRO	F1-CONSPIRACY	F1-CRITICAL
1	IUCL [56]	0.8388	0.9194	0.8947	0.9441
2	AI_Fusion	0.8303	0.9147	0.8866	0.9429
3	SINAI [41]	0.8297	0.9149	0.8886	0.9412
4	ezio [44]	0.8212	0.9097	0.8792	0.9402
5	hinlole [53]	0.8198	0.9098	0.8811	0.9386
6	Zleon [48]	0.8195	0.9096	0.8804	0.9388
7	virmel	0.8192	0.9092	0.8793	0.9391
8	inaki [47]	0.8149	0.9072	0.8770	0.9374
9	yeste	0.8124	0.9057	0.8746	0.9368
10	auxR	0.8088	0.9043	0.8739	0.9347
11	Elias&Sergio	0.8034	0.9012	0.8687	0.9338
12	theateam	0.8031	0.8999	0.8650	0.9347
13	trustno1	0.7983	0.8991	0.8675	0.9307
14	DSVS [46]	0.7970	0.8985	0.8674	0.9296
15	sail [50]	0.7969	0.8978	0.8687	0.9268
16	ojo-bes	0.7969	0.8981	0.8648	0.9314
17	RD-IA-FUN [40]	0.7965	0.8977	0.8636	0.9317
	baseline-BERT	0.7964	0.8975	0.8632	0.9318
18	aish_team [58]	0.7917	0.8944	0.8580	0.9309
19	rfenthusiasts	0.7902	0.8948	0.8605	0.9291
20	Dap_upv	0.7898	0.8944	0.8593	0.9294
21	oppositional_opposition	0.7894	0.8935	0.8571	0.9300
22	miqarn	0.7881	0.8938	0.8593	0.9283
23	CHEEXIST	0.7875	0.8932	0.8576	0.9287
24	tulbure [54]	0.7872	0.8917	0.8536	0.9297
25	XplaiNLP [49]	0.7871	0.8922	0.8550	0.9294
26	TheGymNerds	0.7854	0.8923	0.8567	0.9278
27	nlpln [55]	0.7844	0.8922	0.8580	0.9263
28	ROLLoRico	0.7771	0.8879	0.8559	0.9198
29	LasGarcias	0.7758	0.8855	0.8447	0.9263
30	zhengqiaozeng [52]	0.7758	0.8866	0.8476	0.9256
31	ALC-UPV-JD-2	0.7725	0.8860	0.8491	0.9230
32	LorenaEloy	0.7713	0.8847	0.8455	0.9239
33	lnr-alhu	0.7708	0.8853	0.8488	0.9219
34	NACKO	0.7692	0.8838	0.8446	0.9230
35	paranoia-pulverizers	0.7680	0.8838	0.8462	0.9215
36	DiTana	0.7653	0.8806	0.8490	0.9123
37	FredYNed	0.7643	0.8806	0.8392	0.9220
38	dannuchihaxxx [59]	0.7643	0.8801	0.8377	0.9224
39	lnr-detectives	0.7631	0.8806	0.8472	0.9141
40	TargaMarhuenda	0.7617	0.8807	0.8424	0.9190
41	Trainers	0.7596	0.8797	0.8412	0.9182

Table 5
Results and rankings of the teams participating on Task 1 – binary classification of text as either conspiracy or critical, for English texts. Performance metrics are: Matthews correlation coefficient, macro-averaged F1, and per-class binary F1's.

A. Appendix: Detailed Results

TASK 1 - ENGLISH (cont.)

POSITION	TEAM	MCC	F1-MACRO	F1-CONSPIRACY	F1-CRITICAL
42	thetaylorswiftteam	0.7577	0.8755	0.8302	0.9208
43	locasporlnr	0.7575	0.8787	0.8399	0.9174
44	lnr-adri	0.7552	0.8759	0.8326	0.9192
45	TokoAI	0.7542	0.8767	0.8363	0.9172
46	ede	0.7539	0.8769	0.8384	0.9155
47	lnr-verdnav	0.7529	0.8746	0.8308	0.9185
48	lnr-dahe	0.7488	0.8736	0.8308	0.9163
49	epistemologos	0.7486	0.8742	0.8341	0.9143
50	lucia&ainhoa	0.7473	0.8733	0.8316	0.9150
51	pistacchio	0.7414	0.8678	0.8200	0.9155
52	lnr-BraulioPaula	0.7393	0.8658	0.8165	0.9152
53	Marc_Coral	0.7392	0.8663	0.8176	0.9150
54	Ramon&Cajal	0.7284	0.8633	0.8169	0.9096
55	lnr-lladrogal	0.7253	0.8603	0.8106	0.9100
56	lnr-fanny-nuria	0.7253	0.8594	0.8082	0.9106
57	MarcosJavi	0.7190	0.8583	0.8097	0.9069
58	lnr-cla	0.7168	0.8573	0.8085	0.9061
59	lnr-jacobantonio	0.7168	0.8573	0.8085	0.9061
60	MUCS [51]	0.7162	0.8538	0.7994	0.9082
61	lnr-aina-julia	0.7157	0.8574	0.8102	0.9046
62	LaDolceVita	0.7072	0.8519	0.8000	0.9037
63	alopfer	0.7056	0.8518	0.8012	0.9023
64	lnr-luqrud	0.7056	0.8518	0.8012	0.9023
65	LNR-JoanPau	0.7051	0.8426	0.7793	0.9058
66	lnr-carla	0.7000	0.8476	0.7932	0.9020
67	lnr-Inetum	0.6981	0.8328	0.7617	0.9039
68	lnr-antonio	0.6852	0.8300	0.7598	0.9002
69	LluisJorge	0.6784	0.8382	0.7830	0.8934
70	anselmo-team	0.6725	0.8341	0.7752	0.8930
71	lnr-pavid	0.5959	0.7974	0.7297	0.8651
72	LNRMADME	0.5469	0.7717	0.6914	0.8521
73	lnr-mariagb_elenaog	0.5069	0.7250	0.5966	0.8534
74	LNR_08	0.4429	0.6834	0.5276	0.8391
75	Kaprov [57]	0.3700	0.6240	0.4224	0.8255
76	lnr_cebusqui	0.0482	0.4760	0.1847	0.7674
77	jtommor	0.0403	0.5167	0.3312	0.7023
78	eledu	-0.4598	0.2350	0.2740	0.1960
79	david-canet	-0.6310	0.1632	0.1883	0.1381
80	lnr-guilty	-0.6595	0.1433	0.2247	0.0619
81	lnrANRI	-0.7551	0.1072	0.1474	0.0670
82	ROCurve	-0.8009	0.0884	0.1112	0.0656

Table 6
Results and rankings of the teams participating on Task 1 – binary classification of text as either conspiracy or critical, for English texts. Performance metrics are: Matthews correlation coefficient, macro-averaged F1, and per-class binary F1's.

TASK 1 - SPANISH POSITION	TEAM	MCC	F1-MACRO	F1-CONSPIRACY	F1-CRITICAL
1	SINAI [41]	0.7429	0.8705	0.8319	0.9091
2	auxR	0.7205	0.8572	0.8112	0.9032
3	RD-IA-FUN	0.7028	0.8497	0.8035	0.8960
4	Elias&Sergio	0.6971	0.8485	0.8087	0.8884
5	AI_Fusion	0.6872	0.8419	0.7931	0.8908
6	zhengqiaozeng [52]	0.6871	0.8417	0.7925	0.8909
7	virmel	0.6854	0.8426	0.8022	0.8831
8	trustno1	0.6848	0.8400	0.7895	0.8906
9	Zleon [48]	0.6826	0.8410	0.7955	0.8865
10	ojo-bes	0.6817	0.8395	0.8026	0.8764
11	tulbure [54]	0.6722	0.8293	0.7699	0.8887
12	sail [50]	0.6719	0.8299	0.7713	0.8884
13	nlpln [55]	0.6681	0.8339	0.7872	0.8806
	baseline-BETO	0.6681	0.8339	0.7872	0.8806
14	pistacchio	0.6678	0.8327	0.7822	0.8833
15	rfenthusiasts	0.6656	0.8255	0.7643	0.8868
16	XplaiNLP [49]	0.6622	0.8274	0.7708	0.8840
17	yeste	0.6609	0.8291	0.7770	0.8812
18	oppositional_opposition	0.6601	0.8274	0.7724	0.8825
19	epistemologos	0.6562	0.8264	0.7728	0.8801
20	miqarn	0.6562	0.8264	0.7728	0.8801
21	theateam	0.6557	0.8252	0.7695	0.8810
22	ezio [44]	0.6535	0.8242	0.7683	0.8801
23	lucia&ainhoa	0.6524	0.8260	0.7765	0.8754
24	TargaMarhuenda	0.6516	0.8240	0.7692	0.8787
25	TokoAI	0.6516	0.8240	0.7692	0.8787
26	paranoia-pulverizers	0.6494	0.8246	0.7762	0.8730
27	NACKO	0.6467	0.8232	0.7739	0.8726
28	ALC-UPV-JD-2	0.6467	0.8227	0.7705	0.8748
29	DSVS [46]	0.6462	0.8231	0.7753	0.8709
30	RD-IA-FUN	0.6445	0.8160	0.7523	0.8796
31	locasporlnr	0.6437	0.8216	0.7709	0.8723
32	DiTana	0.6377	0.8187	0.7677	0.8696
33	lnr-BraulioPaula	0.6358	0.8173	0.7731	0.8615
34	Dap_upv	0.6306	0.8115	0.7493	0.8737
35	TheGymNerds	0.6306	0.8106	0.7470	0.8743
36	MUCS [51]	0.6293	0.8060	0.7363	0.8756
37	LasGarcias	0.6247	0.8122	0.7594	0.8649
38	lnr-dahe	0.6196	0.8066	0.7437	0.8694
39	lnr-adri	0.6194	0.8060	0.7422	0.8698
40	hinlole [53]	0.6192	0.8048	0.7391	0.8706
41	RalloRico	0.6105	0.8018	0.7370	0.8666
42	lnr-aina-julia	0.6103	0.7978	0.7264	0.8692
43	lnr-verdnav	0.6101	0.7991	0.7298	0.8684
44	thetaylorswiftteam	0.6066	0.8025	0.7436	0.8613
45	lnr-alhu	0.6024	0.7991	0.7358	0.8624
46	lnr-luqrud	0.6010	0.7945	0.7237	0.8654
47	lnr-lladrogal	0.5967	0.7942	0.7256	0.8627
48	ede	0.5965	0.7967	0.7341	0.8593
49	Fred&Ned	0.5931	0.7940	0.7283	0.8597
50	LaDolceVita	0.5921	0.7818	0.6981	0.8656
51	LNR-JoanPau	0.5920	0.7916	0.7218	0.8614

Table 7
Results and rankings of the teams participating on Task 1 – binary classification of text as either conspiracy or critical, for Spanish texts. Performance metrics are: Matthews correlation coefficient, macro-averaged F1, and per-class binary F1's.

TASK 1 - SPANISH (cont.)

POSITION	TEAM	MCC	F1-MACRO	F1-CONSPIRACY	F1-CRITICAL
52	anselmo-team	0.5899	0.7860	0.7085	0.8634
53	Ramon&Cajal	0.5858	0.7916	0.7281	0.8552
54	lnr-fanny-nuria	0.5813	0.7874	0.7181	0.8567
55	lnr-antonio	0.5736	0.7816	0.7071	0.8561
56	LluisJorge	0.5690	0.7750	0.6929	0.8571
57	lnr-cla	0.5651	0.7788	0.7055	0.8520
58	lnr-jacobantonio	0.5651	0.7788	0.7055	0.8520
59	lnr-pavid	0.5569	0.7771	0.7089	0.8453
60	alopfer	0.5520	0.7727	0.6984	0.8470
61	LNRMADME	0.5490	0.7704	0.6937	0.8471
62	lnr-carla	0.5484	0.7686	0.6890	0.8482
63	LorenaEloy	0.5433	0.7621	0.6751	0.8492
64	CHEEXIST	0.5379	0.5995	0.5621	0.5456
65	lnr-guilty	0.5273	0.7620	0.6880	0.8360
66	eledu	0.5057	0.7263	0.6098	0.8429
67	lnr-mariagb_elenaog	0.4966	0.7325	0.6271	0.8379
68	dannuchihaxxx [59]	0.4727	0.7310	0.6382	0.8238
69	lnr-detectives	0.4029	0.6734	0.6509	0.6960
70	LNR_08	0.0608	0.4771	0.2000	0.7542
71	jtommor	0.0105	0.5051	0.3813	0.6288
72	lnr-Inetum	0.0000	0.3880	0.0000	0.7760
73	Marc_Coral	0.0000	0.2679	0.5359	0.0000
74	MarcosJavi	-0.0389	0.3887	0.0054	0.7720
75	lnr_cebusqui	-0.4112	0.2481	0.3466	0.1496
76	david-canet	-0.5058	0.2114	0.3029	0.1199
77	lnrANRI	-0.6146	0.1766	0.1939	0.1593
78	ROCurve	-0.6457	0.1628	0.1770	0.1485

Table 8
Results and rankings of the teams participating on Task 1 – binary classification of text as either conspiracy or critical, for Spanish texts. Performance metrics are: Matthews correlation coefficient, macro-averaged F1, and per-class binary F1's.

TASK 2 - ENGLISH POSITION	TEAM	span-F1	span-P	span-R	micro-span-F1
1	tulbure [54]	0.6279	0.5859	0.6790	0.6120
2	Zleon [48]	0.6089	0.5537	0.6881	0.5856
3	hinlole [53]	0.5886	0.5243	0.6834	0.5571
4	oppositional_opposition	0.5866	0.5347	0.6586	0.5344
5	AI_Fusion	0.5805	0.5585	0.6082	0.5437
6	virmel	0.5742	0.5235	0.6477	0.5540
7	miqarn	0.5739	0.5184	0.6462	0.5325
8	TargaMarhuenda	0.5701	0.5161	0.6477	0.5437
9	ezio [44]	0.5694	0.5229	0.6340	0.5389
10	zhengqiaozeng [52]	0.5666	0.5122	0.6485	0.5421
11	Elias&Sergio	0.5627	0.5149	0.6364	0.5248
12	DSVS [46]	0.5598	0.5332	0.6012	0.5287
13	CHEEXIST	0.5524	0.4767	0.6845	0.5299
14	rfenthusiasts	0.5479	0.5381	0.5666	0.5408
15	ALC-UPV-JD-2	0.5377	0.4643	0.6562	0.4956
	baseline-BERT	0.5323	0.4684	0.6334	0.4998
16	Dap_upv	0.5272	0.4617	0.6297	0.4973
17	aish_team [58]	0.5213	0.4181	0.7456	0.2571
18	SINAI [41]	0.4582	0.5553	0.4279	0.4571
19	Trainers	0.3382	0.5124	0.2609	0.2858
20	nlpln [55]	0.3339	0.5286	0.3303	0.2710
21	ROCurve	0.2996	0.3154	0.3031	0.3425
22	TokoAI	0.2760	0.1870	0.6119	0.2677
23	DiTana	0.2756	0.5259	0.1947	0.2599
24	TheGymNerds	0.2070	0.2076	0.2127	0.2329
25	epistemologos	0.1709	0.1286	0.3244	0.1201
26	theateam	0.1503	0.1401	0.1652	0.0387
27	LaDolceVita	0.0726	0.2040	0.0453	0.0630
28	kaprov [57]	0.0150	0.0261	0.0165	0.0600

Table 9

Results and rankings of the teams participating on Task 2 – token classification of span-level narrative elements, for English texts. Performance metrics are: span-F1 (macro-averaged over span labels), span-precision, span-recall, and micro-averaged span-F1 [29].

TASK 2 - SPANISH POSITION	TEAM	span-F1	span-P	span-R	micro-span-F1
1	tulbure [54]	0.6129	0.6159	0.6129	0.6108
2	Zleon [48]	0.5875	0.5439	0.6474	0.5939
3	AI_Fusion	0.5777	0.5437	0.6189	0.5843
4	CHEEXIST	0.5621	0.5379	0.5995	0.5456
5	virmel	0.5616	0.4963	0.6584	0.5620
6	miqarn	0.5603	0.5117	0.6273	0.5618
7	DSVS [46]	0.5529	0.5384	0.5785	0.5323
8	TargaMarhuenda	0.5364	0.5128	0.5710	0.5385
9	Elias&Sergio	0.5151	0.4864	0.5533	0.5231
10	hinlole [53]	0.4994	0.4530	0.5740	0.4890
	baseline-BETO	0.4934	0.4533	0.5621	0.4952
11	Dap_upv	0.4914	0.4555	0.5474	0.4917
12	zhengqiaozeng [52]	0.4903	0.4507	0.5494	0.4874
13	ALC-UPV-JD-2	0.4885	0.4509	0.5458	0.4683
14	ezio [44]	0.4869	0.4623	0.5229	0.4947
15	nlpln [55]	0.4672	0.5174	0.4426	0.2961
16	rfenthusiasts	0.4666	0.5104	0.4341	0.4697
17	SIANI	0.4151	0.4630	0.4054	0.4781
18	TheGymNerds	0.3984	0.3621	0.4483	0.5024
19	DiTana	0.3004	0.4490	0.2362	0.3117
20	ROCurve	0.2649	0.2706	0.2627	0.3562
21	TokoAI	0.1878	0.1189	0.5659	0.1739
22	epistemologos	0.1657	0.1906	0.1864	0.1534
23	LaDolceVita	0.1056	0.1158	0.0975	0.1321
24	theateam	0.0994	0.1051	0.0962	0.0358
25	oppositional_opposition	0.0037	0.0349	0.0022	0.0014

Table 10
Results and rankings of the teams participating on Task 2 – token classification of span-level narrative elements, for Spanish texts. Performance metrics are: span-F1 (macro-averaged over span labels), span-precision, span-recall, and micro-averaged span-F1 [29].