Mute Cods: A Multilingual Telegram Dataset with Benchmark Models for Conspiracy Theory Detection

Katarina Laken*[♣†], Erik Bran Marino*[♢], Paloma Piot[♡], Davide Bassi[†]
Søren Fomsgaard[♠], Michele Maggini[†], Renata Vieira[♢]
Marcos Garcia[†], Sara Tonelli[♣]
[♣]Fondazione Bruno Kessler, Trento, Italy
[†]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Santiago de Compostela, Spain
[♢]Universidade de Évora, Évora, Portugal
[♡]IRLab, CITIC Research Centre, Universidade da Coruña, Spain
[♠]Université de Caen Normandie, Caen, France
alaken@fbk.eu, erik.marino@uevora.pt

Abstract

The proliferation of conspiracy theories and hateful messages on social media poses significant challenges for content moderation and public discourse. Despite their societal impact, existing datasets for automated conspiracy detection remain limited in scope and language coverage. We present a multilingual dataset of conspiracy content on Telegram comprising 5,750 messages across English, Dutch, Italian, Spanish and Portuguese from 87 channels documented as disseminating conspiracist and extremist content. Domain experts annotated messages for conspiracist tone, population replacement conspiracy theories, vaccine conspiracies, and hate speech. We extensively report on difficulties and caveats when creating and annotating this type of dataset. We establish classification baselines by evaluating six models in zero-shot fashion and fine-tuning three encoder models, achieving F1 scores up to 0.800 for conspiracist tone, 0.846 for PRCT, 0.843 for vaccine-related conspiracy theories, and 0.734 for hate speech. Inter-annotator agreement was moderate, consistent with the complexity documented in similar annotation tasks.

Keywords: Conspiracy Theories, Social Media analysis, Telegram, Multilingual dataset, LLMs

1. Introduction

Conspiracy theories are narratives that attempt to explain social phenomena through systematic secret plots of powerful actors (Hofstadter, 1964; Douglas et al., 2019). Unlike legitimate skepticism or critical inquiry, conspiracy theories posit coordinated intentionality behind every complex phenomenon, blaming orchestrators who manipulate events for malevolent purposes while dismissing contrary evidence as part of the conspiracy itself (Popper, 1963; Brotherton, 2015). Such theories have proliferated across digital platforms, evolving from marginal discourse to mainstream political narratives with documented real-world consequences (Enders et al., 2024). Indeed, the spread of conspiracy content on social media has been linked to political polarization, public health crises, and acts of violence, making computational detection increasingly urgent for researchers and policymakers (Marino et al., 2024).

A particularly dangerous family of conspiracy discourse is represented by Population Replacement Conspiracy Theories (PRCTs). These false narratives claim deliberate plans to replace native populations through immigration, differential birth rates, and demographic manipulation (Ekman, 2022). The *Great Replacement*, popularised by Renaud Camus,[1] alleges demographic substitution via immigration and differential fertility (Bracke and Hernández Aguilar, 2023). The related *White Genocide* theory, rooted in white supremacist environments, frames multiculturalism and policy choices as existential threats. The *Kalergi Plan* is a misreading of pan-Europeist ideals (Clark and Hagen, 2020). Finally, the *Eurabia* concept casts Muslim migration as deliberate Islamization.[2] Related narratives involve the alleged *feminization* of Western men, supposedly weakening reproduction rates as part of this plot (Hernandez Aguilar, 2024). Research demonstrates that endorsement of replacement narratives correlates with hostile intergroup attitudes, violent intentions, and exclusionary policies, with several mass shootings explicitly motivated by these beliefs (Wojtasik, 2020; Obaidi et al., 2022). These ideas currently legitimize debates about *remigration* (a euphemism for mass deportation) as a solution to this imaginary threat.

* These authors contributed equally to this work.

[1] For the origin and use of this concept, see Renaud Camus, *Le Grand Remplacement, suivi de Discours d'Orange*, R. Camus, 2011.

[2] For the origin and use of this concept, see Oriana Fallaci, *La forza della ragione*, BUR, 2004; and Bat Ye'or, *Eurabia: The Euro-Arab Axis*, Fairleigh Dickinson University Press, 2005.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7345–7358
11–16 May 2026.

Vaccine-related conspiracy theories constitute another prominent category, particularly following the COVID-19 pandemic (Gallotti et al., 2020). These narratives range from claims about vaccine-induced infertility and microchip implantation to broader assertions that vaccines serve as instruments of population control (Ghasemizade and Onaolapo, 2024). As also occurs in other forms of conspiracy language (Marino et al., 2025), it is more accurate to view vaccine skepticism not as a set of distinct categories, but as a spectrum. On one end lies general hesitancy, rooted in safety concerns; on the other, conspiracy-driven opposition, which presupposes systematic deception and is largely invulnerable to factual correction. Vaccine skepticism, and the conspiracy theories related to it, are closely related to other types of skepticism and conspiracist attitudes towards science in general, in which scientists are perceived as a powerful and competent, but not necessarily moral, elite group, that is potentially hostile towards 'the people' (Rutjens et al., 2021; Rutjens and Većkalov, 2022).

In light of the increasing concerns about the dangers of conspiracy theories, it is crucial to develop an in-depth understanding of how they are expressed in different online communities, and how effectively they can be detected automatically. We contribute to this by introducing **Mute Cods**,[3] a novel **Mu**ltilingual **Te**legram **Co**nspiracy **D**ata**s**et comprising over 477,000 tokens, spread over 5,750 messages in five languages (English, Dutch, Italian, Spanish, Portuguese) annotated for different dimensions related to conspiracy, namely: (i) *Conspiracist Tone*, to capture general conspiracy framing; (ii) support for *PRCTs*, given their prevalence in contemporary extremist discourse, documented links to violence and lack of specialized annotated datasets for its detection; (iii) support for *Vaccine Conspiracies*, which saw massive proliferation during the COVID-19 pandemic; (iv) *Hate speech*, a phenomenon closely related to conspiracy theories, as their tendency to posit certain social group as secretly powerful and controlling fosters intergroup hate and prejudice. This often translates directly to hate speech (Bilewicz, 2024).

Our contributions are threefold: (i) we provide the first multilingual dataset for combined categories of Conspiracist tone, PRCT, Vaccine-conspiracy and Hate Speech detection; (ii) we evaluate performance across multiple model architectures, from encoder classifiers to state-of-the-art large language models (LLMs); (iii) we document systematic annotation challenges that characterize real-world conspiracy detection, including implicit narratives, coded language, and cross-linguistic variation. This paper also addresses the choices we made in creating the annotation guidelines, and discusses the difficulties we faced during the annotation process.

The multilingual nature of Mute Cods offers distinct advantages over isolated monolingual datasets. First, applying a single, unified annotation scheme across five languages ensures cross-lingual consistency, which enables comparative studies. Second, it addresses the scarcity of resources for non-English languages. Finally, a unified multilingual resource facilitates the training of models capable of tracking transnational narratives and supporting coordinated, multi-country monitoring efforts, such as those increasingly required at the European level.

2. Related Work

Conspiracy datasets: Despite the societal importance of studying conspiracy content, existing datasets tend to be limited in their scope. Most datasets include exclusively English-language data (Schroeder et al., 2021; Langguth et al., 2023; Wischerath et al., 2025; Phillips et al., 2022; Miani et al., 2022). Exceptions are Steffen et al. (2023), whose German-language dataset focuses on antisemitic conspiracy narratives, and Korenčić et al. (2024b), who create a dataset of conspiracist and oppositional Telegram messages in both Spanish and English. Many datasets focus on a specific type of conspiracy, with COVID-19 being the most prevalent one (Langguth et al., 2023; Korenčić et al., 2024b). Furthermore, Schroeder et al. (2021) collect data on conspiracies linking the COVID-19 pandemic to 5G technology.

Data availability in conspiracy studies is an open issue. Twitter/X is the most commonly studied platform for online conspiracist content (Schroeder et al., 2021; Langguth et al., 2023; Phillips et al., 2022), but its API only allows for the publication of tweet IDs, not the actual content. As many undesirable and extremist tweets are deleted, these datasets become more lacking over the years. Moreover, extracting the original tweets requires permissions that are increasingly difficult to obtain (Murtfeldt et al., 2024). Many papers investigating conspiracy content online do not use datasets annotated at the post level to study the phenomenon. Indeed, several datasets of conspiracist content are annotated at the level of the sender, for example Gambini et al. (2024) (Twitter users, English) or Laken et al. (2025) (Telegram channels, English and Italian). The LOCO corpus by Miani et al. (2022) uses predefined lists of trusted websites and websites known to spread conspiracist content as the basis of their corpus. They blindly annotate a small subset of their data to see how well these labels reflected the actual content of the

[3] Access to our dataset can be requested at https://zenodo.org/records/18848644

domains. Other datasets are not annotated but are rather collections of posts or articles from websites or platforms known to be conspiracist, such as Wischerath et al. (2025) (British conspiracist newspaper) and Aliapoulios et al. (2021) ('free speech' social media platform Parler). Röchert et al. (2022) manually annotate videos as conspiracist or not and look at the comments under those videos. Yet another approach is to use classifiers to estimate whether a post promotes conspiratorial thinking (Moffitt et al., 2021). Finally, there are datasets of misinformation that include tags for conspiratorial content (Burel et al., 2024).

Finally, datasets created for the purpose of training classifiers need to make a choice about what to include as the negative class. Ideally, the negative class is as similar as possible to the positive class in terms of topic and style; otherwise, the classifier can easily learn to rely on irrelevant but effective features. Miani et al. (2022) also use keywords to topic-match the mainstream and the conspiracy parts of their dataset. Phillips et al. (2022) collect tweets on four predefined topics using keyword searches, then manually label them for conspiratoriality, ensuring balanced topic distribution across classes. Korenčić et al. (2024b) create the classes of *conspiracist* and *oppositional* thinking. This is important, because oppositional thinking is similar to conspiracist thinking in some regards, but censoring it would be harmful; their dataset thus allows for the training of classifiers that make this fine-grained distinction.

Conspiracy detection: Several studies attempt to automatically detect conspiracy theories in online content using text classification approaches. The EVALITA shared task on Italian conspiracy data (Russo et al., 2023) reports prompt-based approaches performing best, with fine-tuning approaches also reaching high accuracy. Moreover, data augmentation proves to be a promising strategy as well.

Subsequent research shows the usefulness of emotional and psycholinguistic features for the classification of this type of content (Liu et al., 2024b; Giachanou et al., 2023; Maggini et al., 2025). Peskine et al. (2023) experiment with prompting LLMs using different strategies, showing the influence of using clear, straightforward definitions of the constructs at hand. Liu et al. (2024b) use an emotion-based LLM to classify conspiracy theories in social media content, showing the importance of emotional valence as a feature of conspiratorial content. Another line of research studies the narrative structure of conspiracy content, showing that the way these texts construct narratives differs from other types of content (Shahsavari et al., 2020; Miani et al., 2022; Pilati et al., 2025).

Pustet et al. (2024) compare fine-tuned BERT-based models on the dataset by Steffen et al. (2023) to zero-shot and few-shot classification with prompted models. Zero-shot and fine-tuned approaches perform comparatively in their experiments, but few-shot approaches produce suboptimal outcomes. Building on this work, we advance the field by providing the first multilingual dataset that combines multiple conspiracy dimensions with hate speech detection.

3. Dataset Creation

3.1. Data collection

We use data from Telegram, a messaging platform known for its proliferation of fringe content (Schulze et al., 2022). Using keyword search and snowball sampling (through the 'related channels' function in the application), we manually identified 62 conspiracist and white supremacist channels in English, Dutch, Italian, Spanish and Portuguese, following the same procedure for each language. We used the Telegram API to extract messages from August 2018 up to September 2024. For each language, we identified five mainstream channels to serve as a control group alongside the conspiracist channels. We selected mainstream channels with topics related to the conspiracist ones, such as news, science, and politics. Finding channels without extremist content proved more difficult than finding extremist channels, particularly in Dutch and Portuguese. This likely reflects the position of Telegram in the Netherlands and Portugal, where it appears to function as an echo chamber with fringe political views dominating the entire platform (Cinelli et al., 2022).

3.2. Annotation guidelines

This section provides an overview of the annotation classes and their operational definitions.[4] Hate speech and conspiracy theories are complex social constructs that manifest differently across contexts and cultures (Matamoros-Fernández and Farkas, 2021). We are aware that binary classification is necessarily a simplification of this complexity and we address this through detailed annotation guidelines that explicitly define the scope and boundaries of each category, as follows. Beyond these four primary binary classes, annotators could also mark intermediate cases using additional options (*Doubt*, *Mentions but Opposes*, *Mentions No Judgement*) when messages did not clearly fit binary classification.

[4]The complete guidelines, together with the data, are available on Zenodo

Conspiracist tone: Annotators were asked to label each message as having a conspiracy tone or not. A message with a conspiracist tone is a message that alludes to or mentions a conspiracy. There must be some kind of intention: just stating that certain negative things are happening (whether they are true or not) is not enough if they are not posited as part of a conspiracy. However, it is not necessary for the message to outwardly state what the conspiracy is about, as for example in the following text:

> First Turkey blocks Ukraine from entering the NATO, and next thing you know there is an earthquake… But of course, we are supposed to believe that this is a mere coincidence! 🤡

Population Replacement Conspiracy Theories (PRCTs): This family of theories suggests that the culture, identity and demography of Western societies is directly under threat and a strong action is needed to counter this phenomenon. An example is included below. Our guidelines included lists of commonly used dog whistles and symbols associated with these theories.

Not all anti-migration or migration-skeptic posts are examples of PRCT, which posits cultural and racial purity as an absolute value; blending is always seen as bad as it somehow damages the (white) in-group. This is not the same as discussing current problems with migration politics.

> [name] thinks pro-White people should "pool resources" to save poor Muslim immigrants from "state-backed racial harassment." Does this guy even care about the Great Replacement?

Vaccine-related conspiracy: This label includes conspiracies about vaccines causing issues like infertility, sudden death, or autism, or being otherwise dangerous. We consider extreme vaccine skepticism as inherently conspiratorial when it alleges cover-ups of harmful effects contrary to scientific consensus. We know that this is not always the same exact thing, but we have no way to differentiate between the two based on just a message. Lighter vaccine skepticism is not necessarily conspiratorial. Posts that posit COVID-19 as fake or a diabolical plan entail extreme vaccine skepticism (at least against COVID vaccine) or a generally conspiratorial world view, so we also consider them as vaccine-related conspiracy theories. Sometimes, references to vaccine related conspiracies can be subtle, like in the following example:

> Yet another young, healthy person drops dead... Nothing to see here!!

Hate speech: This label concerns whether the post propagates hate against people or a specific person based on them belonging to some social group. Following earlier work (Davidson et al., 2017; ElSherief et al., 2018; Piot et al., 2024), we define hate speech as content that *expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group, based on identity-based attributes like gender, sexual orientation, ethnicity, race, disability, or religion*. Hate speech can be obscene and use swear words, but this is not always the case. Our dataset includes various nazist channels that use scientific language to spread false information about genetic differences between people. Conversely, an abusive or offensive post attacking individual people with no reference to demographic/social groups is not considered hate speech.

3.3. Annotation process

Our annotation was carried out by 8 expert annotators. All annotators were PhD students specializing in topics related to disinformation, hate speech, and NLP, apart from the Portuguese native speakers, who were Master's students. The annotation was carried out over seven rounds of annotation (but not all annotators participated in each round). Regular meetings were organized in which all annotators met to discuss doubts regarding the annotation guidelines and borderline cases. After all items were annotated, the first two authors jointly adjudicated 435 items with disagreement, discussing each of them. We used the disambiguated labels in our experiments.

Following standard practice in large-scale annotation projects, we computed agreement on multiple-annotated items while expanding dataset coverage through single annotation of additional instances. Table 1 shows the amount of annotations per item per language. We calculated inter-annotator agreement using Krippendorff's α (Castro, 2017), obtaining $\alpha = 0.46$ for conspiracist tone, $\alpha = 0.58$ for PRCT, $\alpha = 0.36$ for vaccine conspiracy, and $\alpha = 0.50$ for hate speech, aggregated across all languages and annotators. When calculating IAA, also adjudicated labels were included as a third label beside those assigned by two annotators.

While these values indicate moderate agreement, the corresponding percentual agreements were substantially higher: 72.07% for conspiracist tone, 93.80% for PRCT, 93.84% for vaccine conspiracy, and 86.21% for hate speech, comparable to

Langguth et al. (2023)'s reported range of 75.85%–97.45%. This discrepancy reflects the impact of class imbalance on chance-corrected metrics: Krippendorff's α accounts for agreement by chance, which is elevated when one class (typically negative) predominates in the data. The moderate α values are consistent with the complexity and inherent subjectivity documented in conspiracy theory annotation tasks (Hemm et al., 2024). In fact, since seven rounds of annotation and discussions were not enough to reconcile all disagreements, we can infer that different labels correspond to genuine human label variation depending on different interpretations by the annotators (Plank, 2022; Sandri et al., 2023).

Lang.	# Annotators		
	1	2	3+
EN	674	398	78
NL	698	355	97
IT	394	609	147
ES	646	381	123
PT	739	315	96

Table 1: Table showing the amount of annotators per item per language (1, 2, or 3 or more). Adjudication labels are counted as one annotation

4. Dataset Description

4.1. Data statistics

Our final dataset, which we call **MUTE CODS**, includes, for each language, 1,000 posts from conspiracist channels (annotated on a post level) and 150 posts from relatively mainstream channels. Table 2 shows the amount of channels with the mean and standard deviation of the amount of tokens and sentences (as per the NLTK sentence tokenizer function (Bird, 2006)). The standard deviation of both measures is much larger for conspiracist channels across all languages, indicating a greater variability in post length.

Table 3 shows the distribution of class labels over the data (including both the conspiracist and mainstream part of the dataset). We see very large differences between languages, especially when it comes to the 'vaccine' and 'hate speech' categories. The English and Dutch samples are characterized by relatively high amounts of conspiracist tone (around twice as much as the amount of hate speech posts in the sample), with equal, relatively low numbers of vaccine and population replacement-related content. The Spanish sample is similar, but with relatively lower amount of posts with a conspiracist tone. The Italian channels show a large amount of posts with a conspiracist tone, relatively many posts regarding vaccines, and almost

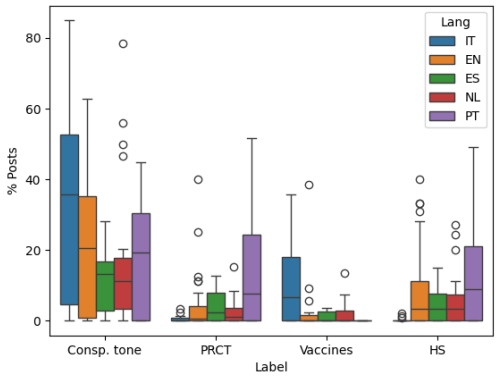

Figure 1: Percentage of posts per channel labeled as positive for each class, split out by language. Each data point is one channel. Circles represent outliers.

no hate speech. The Portuguese sample is almost the opposite, with no vaccine-related conspiracies and relatively large amounts of hate speech and population replacement theories. These statistics characterize our specific sample. Given the relatively small number of channels and the considerable variability in label distribution across individual channels, they should be interpreted as descriptive of our dataset composition rather than as broader cultural patterns, since our sample included relatively few channels.

4.2. Challenging annotation cases

MUTE CODS contains posts from channels known to spread conspiracist and otherwise harmful content. These concepts, however, are not as clear-cut as expected. We provide below a list of challenging annotation cases with examples.

Support for extremist ideologies: Several posts in our sample discuss issues relating to extremist ideologies (most notably white supremacy), without mentioning the parts of those ideologies that make them explicitly hateful and/or conspiracist, see the following example:

> Activists from the Netherlands and Germany have joined forces to spread a pro-white message, wishing people in the Netherlands a happy White 2023. 👌

We decided that, in accordance with our guidelines, these types of posts cannot be considered hate speech. Even though the social implications of being 'pro-White' are extremely different from the implications of being, for example, 'pro-Black' or 'pro-woman', we intentionally avoided making social

	Conspiracy			Mainstream			Total		
Lang.	**# Chann.**	**M (SD) Tokens**	**M (SD) Sent.**	**# Chann.**	**M (SD) Tokens**	**M (SD) Sent.**	**# Chann.**	**M (SD) Tokens**	**M (SD) Sent.**
EN	26	73.49 (102.30)	3.90 (5.05)	5	75.95 (60.29)	2.91 (2.43)	31	73.81 (97.84)	3.77 (4.80)
NL	14	63.24 (67.31)	3.98 (3.87)	5	42.21 (42.91)	2.79 (2.59)	19	60.49 (65.02)	3.82 (3.74)
IT	12	110.53 (115.62)	5.50 (5.33)	5	32.71 (24.23)	2.05 (1.61)	17	100.38 (111.29)	5.05 (5.14)
ES	10	66.77 (70.81)	3.60 (3.01)	5	110.58 (45.41)	4.89 (2.79)	15	72.48 (69.60)	3.11 (3.01)
PT	9	59.39 (63.59)	3.18 (3.11)	5	105.49 (114.72)	4.73 (5.08)	14	65.40 (73.91)	3.39 (3.46)

Table 2: Table showing, for each language and each type, the amount of channels, the average amount of tokens per post, and the average amount of sentences per post. M = mean, SD = standard deviation

	Conspir. Tone		Pop. Repl.		Vaccines		Hate Speech	
Lang.	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
EN	276	24.00	43	3.74	50	4.35	107	9.30
NL	324	28.17	79	6.87	89	7.74	142	12.35
IT	426	37.04	17	1.48	131	11.38	13	1.13
ES	148	12.87	49	4.26	14	1.22	58	5.04
PT	256	22.26	180	15.65	0	0.00	258	22.43
Total / Avg.	1430	24.87	368	6.40	284	4.94	578	10.05

Table 3: Number of posts (N) and percentage (%) associated with each class across languages. The bottom row shows the total number and the percentage average across all languages.

marginalization a criterion in our hate speech definition. Based on this logic, we also cannot classify as hateful messages that explicitly *support* a social group based on its social power, if the post does not explicitly put down other groups. We are however aware that this leads to problematic posts getting a 'nothing detected' label, and we encourage future dataset creators to give this problem careful consideration, making choices based on their dataset purpose and the specifics of their data.

Emerging narratives: Some messages take on a hateful message only in the context of the channel as a whole, that paints a conspiracist and/or hateful picture of the world. Each post in itself cannot clearly be considered problematic because the hateful message emerges by presenting information in a biased and incomplete way. For example, certain white supremacist channels post many studies about genetic differences between people from different ethnic backgrounds. In another context, these could be considered interesting factoids, but in the context of the channels they are used to consolidate a racist narrative (see example below). But as we operate at the post level, these posts were not labeled as being conspiracist or hateful.

> A 2023 study proposed that Eurasians and Sub-Saharan Africans genetically diverged over 100,000 years ago. Eurasians lived in the Saudi Peninsula, genetically isolated from at least 80,000 years ago, before expanding north around 54,000 years ago. This 30,000 year isolation period led to genetic adaptations in Eurasians relating to "the regulation of fat storage, neural development, skin physiology" potentially representing "selection for cold adaptation." [redacted]

Vaccine skepticism: The boundary of when vaccine skepticism becomes conspiracism is not always clear. We consider extreme skepticism as conspiratorial when it presupposes coverups or hidden agendas, but not when expressing concerns about safety or bodily autonomy.

Unimodality: Many posts originally included videos or images, but our dataset only preserves textual content, so crucial information is lacking. Including the multimodal data could change the interpretation of a post. Moreover, many posts contain URLs with potentially important context, but we decided to convert all URLs to a [URL] token in the preprocessing phase in order to avoid potential copyright infringement.

Post length: Some posts were very long, with the specific hate speech or conspiracist content being only a relatively small part of the post. We did not include any span annotation or split up posts in paragraphs. This means that labels apply to entire posts even when only a small portion contains the problematic content.

Small scale conspiracies: The channels in our sample post in support of big conspiracies that in-

volve practically the whole world. Individual posts often discuss specific events that are framed as conspiracies, without being explicitly linked to the 'overarching conspiracy'. The problem with these types of posts is that they are often not problematic in and of itself; the suggestion of small scale conspiracies (e.g. of a couple of people deliberately plotting something) is not necessarily a problem.

Fake news: Several posts presented 'facts' that appeared to be fake or severely distorted, but without being conspiratorial or hateful. We do not take into account the annotation of factuality, since it was outside the scope of this work.

5. Methodology

We evaluate whether manual annotation can be replicated accurately by automated classification. Each category (conspiracist tone, PRCT, vaccine conspiracy, and hate speech) is treated as an independent binary classification task, resulting in four separate classifiers. Classes are not mutually exclusive: messages may exhibit multiple categories simultaneously, with the constraint that vaccine conspiracy or PRCT necessarily imply conspiracist tone. For conspiracist tone and hate speech, classification is binary (Yes/No). For PRCT and vaccine conspiracy, we reduced the original annotation to binary classification by grouping positive instances (whether the message mentions or supports the theory) versus negative instances (does not mention). We split the full dataset into 4,599 training instances and 1,151 test instances using an 80/20 split stratified by channel.

We evaluate three sets of models: (i) API-based large language models (GPT-5 (OpenAI, 2025), DeepSeek v3 (Liu et al., 2024a)), (ii) locally-deployed generative models (Mistral-7B-Instruct-v0.3 (Mistral-AI, 2025) and Dolphin-Mistral-Nemo-12B (Hartford, 2025),[5] ConspEmoLLM (Liu et al., 2024b) and Phi4 (Abdin et al., 2024) via Ollama), and (iii) fine-tuned transformer encoders (CT-BERT-v2 (Müller et al., 2023), Toxic-BERT (Davidson et al., 2019), XLM-RoBERTa (Conneau et al., 2020)). All models are trained and evaluated on the full multilingual dataset without language-specific splits, which is indeed interesting but we leave as future work. Among the encoder models, Toxic-BERT covers four of our languages (English, Spanish, Italian, Portuguese, but not Dutch), while XLM-RoBERTa is fully multilingual and CT-BERT-v2 supports English only. Nonetheless we chose to keep this model to test whether conspiratorial language patterns exhibit cross-lingual transfer despite training exclusively on English data. All generative models were evaluated using an identical, structured zero-shot prompt. The prompt explicitly defined the task and required the output in a strict json format.[6]

For classification using API-based and locally-deployed LLMs, we implement a rigorous protocol following Atil et al. (2024) to address non-determinism in neural text generation. Each message receives five independent classifications using different random seeds (42, 123, 456, 789, 1024) with temperature set to 0, top-p to 1.0, max tokens to 500, and frequency/presence penalties to 0. Final predictions result from majority voting across the five runs, with confidence scores computed as the proportion of runs agreeing on each classification. Results show a mean Total Agreement Rate (TAR) of 1.0 for local models and >0.96 for API models. Concerning the encoder models, we fine-tune and evaluate them separately for each task. The training portion of the dataset is further divided into 80% for training and 20% for validation, resulting in approximately 3,679 training and 920 validation instances per task. We use a maximum sequence length of 128 tokens, batch size of 16 with gradient accumulation over 4 steps, learning rate of 2e-5 with 10% linear warmup, and weight decay of 0.01. Training includes a maximum of 5 epochs with evaluation every 100 steps. To address class imbalance across tasks, we compute balanced class weights using scikit-learn and incorporate them into a weighted cross-entropy loss function through a custom Trainer implementation. Best checkpoints are selected based on validation F1-score with early stopping enabled.

6. Results

Table 4 shows the results of our experiments. For each individual task, we report two metrics: the Macro-F1, calculated as the unweighted average of the F1-scores of its two constituent classes (positive and negative), and the Binary-F1, which evaluates performance exclusively on the positive minority class (a reporting format also adopted by Korenčić et al. (2024a)). Zero-shot LLMs consistently outperform fine-tuned encoder models across all tasks, with GPT5 and DeepSeek-v3 achiving the highest macro F1 scores. Notably, Phi4 demonstrates competitive performance despite its smaller size (14B parameters), achieving the best score for vaccine conspiracy detection (0.843 macro F1) and strong results across other tasks. This performance gap likely stems from LLMs' exposure to diverse content during pre-training and alignment, enabling them to recognize conspiratorial and hateful language patterns. In contrast, encoder models

[5] On NVIDIA A100-SXM4-40GB GPUs with CUDA 11.8

[6] The prompt structure was: **Instruction** → **Label Definitions** → **Input Text** → **Classification Task**.

Model	Conspiracist Tone		Hate Speech		PRCT		Vaccine	
	$F1_{MACRO}$	$F1_{BINARY}$	$F1_{MACRO}$	$F1_{BINARY}$	$F1_{MACRO}$	$F1_{BINARY}$	$F1_{MACRO}$	$F1_{BINARY}$
GPT-5	0.784	0.679	**0.734**	**0.530**	**0.846**	**0.709**	0.795	0.607
DeepSeek V3	**0.800**	**0.713**	0.716	0.505	0.792	0.613	0.826	0.667
Dolphin	0.709	0.531	0.585	0.344	0.735	0.518	0.812	0.642
Mistral	0.750	0.614	0.626	0.310	0.794	0.605	0.747	0.509
ConspEmoLLM	0.515	0.412	0.481	0.155	0.489	0.000	0.491	0.000
Phi4	0.752	0.618	0.730	0.517	0.812	0.646	**0.843**	**0.696**
CT-BERT	0.719	0.597	0.624	0.323	0.671	0.384	0.786	0.593
Toxic-Bert	0.701	0.570	0.680	0.430	0.701	0.442	0.698	0.427
XLM-RoBERTa	0.743	0.629	0.695	0.457	0.720	0.478	0.770	0.567

Table 4: Unified comparison across models on the Telegram test set. For each model, we show the macro and binary F1.

face the challenge of learning from a multilingual training set with relatively limited data per language, which may hinder their ability to capture language-specific features.

Among encoder models, XLM-RoBERTa demonstrates the most robust performance thanks to its multilingual architecture, while CT-BERT's English-only training surprisingly shows competitive results, suggesting that conspiratorial language patterns may exhibit some cross-lingual transferability. ConspEmoLLM's poor performance likely stems from the zero-shot evaluation setup combined with task/schema mismatch: the model was fine-tuned on ConDID's specific label space (Covid-19 ternary classification) and must generalize to our different annotation framework without access to task-specific training data, unlike the fine-tuned encoder models (such as CT-BERT), which directly learn our label schema from the training set.

Our multi-seed evaluation protocol revealed interesting differences in prediction consistency among models' types. Locally-deployed models achieved near perfect determinism (TAR $= 1.0$), while API-based models showed substantial variability (TAR 0.973 for GPT-5, 0.964 for DeepSeek V3), with approximately 7% of messages receiving divergent classifications across the five runs despite deterministic parameters. This confirms Atil et al. (2024)'s findings on LLM non-determinism and suggests that single-run evaluations with API models, a common practice in prior works, risk reporting non-representative results.

The consistent gap between macro and binary F1 scores across all models reflects the impact of class imbalance inherent in real-world applications: models achieve higher overall accuracy by correctly predicting the majority class while struggling with minority class detection. As expected, larger state-of-the-art LLMs generally outperformed smaller models. However, Phi4 achieved competitive performance despite its smaller architecture (14B parameters compared to hundreds of billions in larger state-of-the-art LLMs), suggesting that while model size certainly helps, it does not solely determine effectiveness for such detection tasks. We observed cross-lingual performance variation, with classification of Spanish messages yielding the best results (79.41% correct predictions for conspiracist tone), followed by Portuguese (74.70%), English (72.12%), Dutch (67.54%), and Italian (56.18%). However, this ranking inversely correlates with conspiracy prevalence in each language's test set: Spanish exhibits the lowest conspiracist tone rate (12.87%) while Italian shows the highest (34.04%). This pattern, we believe, reflects models' conservative bias toward predicting the negative class rather than genuine language-specific differences: in languages with fewer positive instances, models achieve higher accuracy by predominantly predicting "does not mention". Conversely, in Italian, where positive instances are more frequent, the same conservative pattern results in more false negatives and lower overall accuracy.

6.1. Error analysis

Generally, the models are better at correctly classifying non-conspiracist content than at recognizing conspiracist messages, a pattern that matches human annotators. We manually inspected the instances that were most often classified wrongly by our models. Unsurprisingly, the false negatives category mainly concerned posts that were on the borderline of conspiracy. Other posts in this category discussed conspiracies, but did not seem to embrace them (see example below).

> AI Gangstalking refers to the belief that individuals are being constantly monitored and harassed by governments or organizations using artificial intelligence. Despite no scientific backing, some report physical symp-

toms they attribute to gangstalking.

Concerning false positives, we found 10 non-conspiracist posts that were labeled as conspiracist by most models. In all cases, these were strikingly short messages, spanning just a couple of sentences, suggesting that classifiers struggle in the identification of conspiracy content when context is limited. To further investigate whether language models tend to provide the same judgments on the same posts, we calculate IAA for the models for all classes using Krippendorff's α. We find it to be considerably lower than for human annotators for all classes (Tone = 0.39, PRCT = 0.39, HS = 0.31), except for vaccine conspiracy (α = 0.44). We then perform an ablation test by calculating IAA after removing one model at a time, and compare it against IAA with all models included, as a measure of how much each models predictions deviate from the others. Interestingly, ConspEmoLLM exhibits the largest deviation across all classes, with IAA improving most when this model was excluded. This may reflect its distinct generative architecture and emotion-oriented training compared to the encoder-only models in our comparison.

7. Conclusion

This paper introduced MUTE CODS, a multilingual dataset for the analysis and detection of online conspiracist content. Our annotation scheme uniquely combines classes for general conspiracist tone, PRCT, vaccines, and hate speech. We manually annotate 5,750 Telegram messages in English, Dutch, Italian, Spanish, and Portuguese.

Our classification experiments, which included zero-shot evaluation of six models and fine-tuning of three encoder-only models, achieved promising results, with F1 scores reaching 0.800 for conspiracist tone, 0.734 for hate speech, 0.846 for population replacement conspiracy theories, and 0.843 for vaccine-related conspiracy theories. These results establish a robust baseline for automated classification of conspiracy content and indicate the viability of computational approaches to analyse extremist discourse on social media. The dataset, the annotation guidelines and the code implemented to run the classification experiments are available on Zenodo.

8. Limitations

Our dataset exhibits moderate inter-annotator agreement, particularly for the 'vaccine' class. This reflects the inherent subjectivity of such annotation tasks, consistent with agreement ranges reported in similar works on conspiracy theory and hate speech detection. All of our annotators were experts in conspiracy theory research. Annotation across multiple languages meant that some annotators worked with languages they were not native speakers of, which may have affected the detection of culturally-specific linguistic patterns. But given the difficulty of the data (being posts from channels we know to be conspiracist), and the first authors' experience re-annotating disputed items, we hypothesize that a big part of this is due to an inherent ambiguity of both the measured constructs combined with the difficulty of our data, that includes many non-conspiracist posts from conspiracist channels.

We also want to point out that our annotations in and of itself also represent a rather biased view, since all annotators are white PhD or master's students based in Europe.

Another limitation is the distribution of labels in our dataset, that shows considerable differences between languages and categories. For example, our dataset includes zero instances of vaccine related conspiracies in Portuguese. While we hypothesize that this absence might reflect the actual prevalence of these specific narratives within the sampled regional Telegram communities, we lack definitive proof to confirm it. We did not artificially balance the dataset, which may explain the models' conservative bias and higher false negative rate. While future work could use resampling techniques or downsizing the negative class to improve minority class detection, we prioritized ecological validity. In real-world moderation, conspiratorial content is rare; training on artificially balanced data would over-sensitize the models, increasing false positives upon deployment.

9. Ethical Considerations

This dataset contains extremist content including white supremacist rhetoric, antisemitic conspiracy theories, and explicit hate speech targeting marginalized groups. While all data originates from publicly accessible Telegram channels, we acknowledge the sensitive nature of this material and implement several safeguards. The dataset will be released under a research-only license requiring institutional affiliation verification and explicit agreement not to use the data for training systems that generate or amplify harmful content. Moreover, we removed usernames as an anonymisation strategy, and the Telegram API does not allow to re-identify a user by searching for a comment, making tracking back posts and users difficult.

Annotators were exposed to prolonged contact with extremist narratives during the labeling process. We tried to mitigate this through regular debriefing meetings and explicit consent protocols. All annotators are PhD or Master's students with

prior research experience in extremism studies, though we acknowledge this does not eliminate psychological risks inherent to such work. Several annotators worked on languages they are not native speakers of, which may affect annotation quality for culturally-specific dog whistles or coded language. Future work should prioritize diverse annotator backgrounds and native speaker expertise for each language. The dataset enables detection systems that could assist platform moderation, but also carries dual-use risks. Malicious actors could potentially use conspiracy theory classifiers to identify and amplify such content, or to evade detection by understanding model weaknesses. We cannot prevent all misuse, but require explicit research-purpose declarations and prohibit commercial deployment without ethics review.

Acknowledgements

This work has been supported by the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351, by the UK Research and Innovation (UKRI) Horizon Europe funding guarantee - Grant Number: EP/X036758/1 and by the Galician Government (ED431G 2023/04 and ED431B 2025/16). Generative AI was utilized for language editing and proofreading. All content responsibility remains with the authors.

10. References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Max Aliapoulios, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. 2021. A large open dataset from the parler social network. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 943–951.

Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, et al. 2024. Non-Determinism of "Deterministic" LLM Settings. *arXiv preprint arXiv:2408.04667*.

Michał Bilewicz. 2024. How appraisal model allows to distinguish intergroup conspiracy theories from other forms of hate speech. *Psychological Inquiry*, 35(3-4):216–222.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Sarah Bracke and Luis Manuel Hernández Aguilar. 2023. *The Politics of Replacement: From "Race Suicide" to the "Great Replacement"*. Routledge, London.

Rob Brotherton. 2015. *Suspicious minds: Why we believe conspiracy theories*. Bloomsbury Publishing.

Grégoire Burel, Martino Mensio, Youri Peskine, Raphael Troncy, Paolo Papotti, and Harith Alani. 2024. Cimplekg: A continuously updated knowledge graph on misinformation, factors and fact-checks. In *International Semantic Web Conference*, pages 97–114. Springer.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Matteo Cinelli, Gabriele Etta, Michele Avalle, Alessandro Quattrociocchi, Niccolò Di Marco, Carlo Valensise, Alessandro Galeazzi, and Walter Quattrociocchi. 2022. Conspiracy theories and social media platforms. *Current Opinion in Psychology*, 47:101407.

Roland Clark and Nikolaus Hagen. 2020. Kalergi plan: The undying "white genocide" conspiracy theory.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international*

AAAI conference on web and social media, volume 11, pages 512–515.

Karen M Douglas, Joseph E Uscinski, Robbie M Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding conspiracy theories. *Political psychology*, 40:3–35.

Mattias Ekman. 2022. The great replacement: Strategic mainstreaming of far-right conspiracy claims. *Convergence*, 28(4):1127–1143.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Adam Enders, Casey Klofstad, and Joseph Uscinski. 2024. The relationship between conspiracy theory beliefs and political violence. *Harvard Kennedy School Misinformation Review*, 5(6).

Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. 2020. Assessing the risks of 'infodemics' in response to covid-19 epidemics. *Nature human behaviour*, 4(12):1285–1293.

Margherita Gambini, Serena Tardelli, and Maurizio Tesconi. 2024. The anatomy of conspiracy theorists: unveiling traits using a comprehensive twitter dataset. *Computer Communications*, 217:25–40.

Mohsen Ghasemizade and Jeremiah Onaolapo. 2024. Developing a hierarchical model for unraveling conspiracy theories. *EPJ Data Science*, 13(1):31.

Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2023. Detection of conspiracy propagators using psycho-linguistic characteristics. *Journal of Information Science*, 49(1):3–17.

Eric Hartford. 2025. Dolphin-2.9.3-mistral-nemo-12b. https://huggingface.co/dphn/dolphin-2.9.3-mistral-nemo-12b. Cognitive Computations. Instruction-fine-tuned model based on Mistral-Nemo-Base. Hugging Face model card.

Ashley Hemm, Sandra Kübler, Michelle Seelig, John Funchion, Manohar Murthi, Kamal Premaratne, Daniel Verdear, and Stefan Wuchty. 2024. Are you serious? handling disagreement when annotating conspiracy theory texts. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 124–132, St. Julians, Malta. Association for Computational Linguistics.

Luis M Hernandez Aguilar. 2024. Memeing a conspiracy theory: On the biopolitical compression of the great replacement conspiracy theories. *Ethnography*, 25(1):76–97.

Richard J. Hofstadter. 1964. The paranoid style in american politics. *Harper's Magazine*, pages 77–86.

Damir Korenčić, Berta Chulvi, Xavier Bonet-Casals, Mariona Taulé, Paolo Rosso, and Francisco Rangel. 2024a. Overview of the oppositional thinking analysis pan task at clef 2024.

Damir Korenčić, Berta Chulvi, Xavier Bonet Casals, Alejandro Toselli, Mariona Taulé, and Paolo Rosso. 2024b. What distinguishes conspiracy from critical narratives? a computational analysis of oppositional discourse. *Expert Systems*, 41(11):e13671.

Katarina Laken, Matteo Melis, Sara Tonelli, and Marcos Garcia. 2025. Multilingual analysis of narrative properties in conspiracist vs mainstream telegram channels. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 426–457, Vienna, Austria. Association for Computational Linguistics.

Johannes Langguth, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, Jesper Phillips, and Konstantin Pogorelov. 2023. Coco: an annotated twitter dataset of covid-19 conspiracy theories. *Journal of Computational Social Science*, 6(2):443–484.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Zhiwei Liu, Boyang Liu, Paul Thompson, Kailai Yang, and Sophia Ananiadou. 2024b. ConspEmoLLM: Conspiracy Theory Detection Using an Emotion-Based Large Language Model. *arXiv preprint arXiv:2403.06765*.

Michele Joshua Maggini, Dhia Merzougui, Rabiraj Bandyopadhyay, Gaël Dias, Fabrice Maurel, and Pablo Gamallo. 2025. Are LLMs Enough for Hyperpartisan, Fake, Polarized and Harmful Content Detection? Evaluating In-Context Learning vs. Fine-Tuning. ArXiv preprint arXiv:2509.07768.

Erik Marino, Jesus M. Benitez-Baleato, and Ana Sofia Ribeiro. 2024. The polarization loop: How emotions drive propagation of disinformation in online media—the case of conspiracy theories and extreme right movements in southern europe. *Social Sciences*, 13(603).

Erik Bran Marino, Davide Bassi, and Renata Vieira. 2025. Linguistic markers of population replacement conspiracy theories in youtube immigration discourse. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, Cagliari.

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & new media*, 22(2):205–224.

Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2022. Loco: The 88-million-word language of conspiracy corpus. *Behavior research methods*, 54(4):1794–1817.

Mistral-AI. 2025. Mistral-7b-instruct-v0.3 (model card). https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. Large language model, fine-tuned "instruct" version.

JD Moffitt, Catherine King, and Kathleen M Carley. 2021. Hunting conspiracy theories during the covid-19 pandemic. *Social Media+ Society*, 7(3):20563051211043212.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2023. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *Frontiers in artificial intelligence*, 6:1023281.

Ryan Murtfeldt, Naomi Alterman, Ihsan Kahveci, and Jevin D West. 2024. Rip twitter api: A eulogy to its vast research contributions. *arXiv preprint arXiv:2404.07340*.

Milan Obaidi, Jonas Kunst, Simon Ozer, and Sasha Y Kimel. 2022. The "great replacement" conspiracy: How the perceived ousting of whites can evoke violent extremism and islamophobia. *Group Processes & Intergroup Relations*, 25(7):1675–1695.

OpenAI. 2025. Gpt-5. https://chatgpt.com/en-EN/overview/. Large language model developed by OpenAI.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding GPT for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4054–4063, Singapore. Association for Computational Linguistics.

Samantha C Phillips, Lynnette Hui Xian Ng, and Kathleen M Carley. 2022. Hoaxes and hidden agendas: A twitter conspiracy theory dataset: Data paper. In *Companion Proceedings of the Web Conference 2022*, pages 876–880.

Federico Pilati, Tommaso Venturini, Pier Luigi Sacco, and Floriana Gargiulo. 2025. Pseudo-scientific versus anti-scientific online conspiracism: A comparison of the flat earth society's internet forum and reddit. *New Media & Society*, 27(9):5324–5341.

Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. Metahate: A dataset for unifying efforts on hate speech detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 2025–2039.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karl Popper. 1963. Conjectures and refutations: The growth of scientific knowledge. new york, harper.

Milena Pustet, Elisabeth Steffen, and Helena Mihaljevic. 2024. Detection of conspiracy theories beyond keyword bias in German-language telegram using large language models. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 13–27, Mexico City, Mexico. Association for Computational Linguistics.

Giuseppe Russo, Niklas Stoehr, and Manoel Horta Ribeiro. 2023. ACTI at EVALITA 2023: Automatic Conspiracy Theory Identification Task Overview. In *EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 3473, Parma. CEUR WS.

Bastiaan T Rutjens, Sander Van der Linden, and Romy Van der Lee. 2021. Science skepticism in times of covid-19. *Group Processes & Intergroup Relations*, 24(2):276–283.

Bastiaan T Rutjens and Bojana Većkalov. 2022. Conspiracy beliefs and science rejection. *Current Opinion in Psychology*, 46:101392.

Daniel Röchert, German Neubaum, Björn Ross, and Stefan Stieglitz. 2022. Caught in a networked collusion? homogeneity in conspiracy-related discussion networks on youtube. *Information Systems*, 103:101866.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Ježek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441.

Daniel Thilo Schroeder, Ferdinand Schaal, Petra Filkukova, Konstantin Pogorelov, and Johannes Langguth. 2021. Wico graph: A labeled dataset of twitter subgraphs based on conspiracy theory and 5g-corona misinformation tweets. In *ICAART (2)*, pages 257–266.

Heidi Schulze, Julian Hohner, Simon Greipl, Maximilian Girgnhuber, Isabell Desta, and Diana Rieger. 2022. Far-right conspiracy groups on fringe platforms: A longitudinal analysis of radicalization dynamics on telegram. *Convergence: The International Journal of Research into New Media Technologies*, 28(4):1103–1126.

Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R Tangherlini, and Vwani Roychowdhury. 2020. Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news. *Journal of computational social science*, 3(2):279–317.

Elisabeth Steffen, Helena Mihaljevic, Milena Pustet, Nyco Bischoff, María do Mar Castro Varela, Yener Bayramoglu, and Bahar Oghalai. 2023. Codes, patterns and shapes of contemporary online antisemitism and conspiracy narratives–an annotation guide and labeled german-language dataset in the context of covid-19. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1082–1092.

Darja Wischerath, Desislava Bocheva, Emily Godwin, Alberto Arletti, Olivia Brown, and Brittany I Davidson. 2025. Seeing The Light - looking into Britain's conspiracy Truthpaper. Dataset.

Karolina Wojtasik. 2020. Utøya–christchurch–halle. right-wing extremists' terrorism. *Security Dimensions. International and National Studies*, (33):84–97.

Appendix

Gender	Age	Native Lang.	# annotations per language					
			# EN	# NL	# IT	# ES	# PT	Total
Male	31	DK	519	699	0	0	0	1218
Female	30	NL	809	1104	101	238	0	2252
Male	27	IT	184	0	0	848	707	1739
Male	25	PT	10	0	0	0	578	588
Male	30	IT	99	0	1143	0	0	1242
Male		IT	119	0	1100	0	0	1219
Female	29	ES	49	0	0	849	46	944
Male	25	PT	6	0	0	0	374	380

Table 5: Overview of the annotator demographics and number of items annotated per language