One Model to Detect Them All? Comparing LLMs, BERT and Traditional ML in Cross-Platform Conspiracy Detection

Anonymous ACL submission

Abstract

The proliferation of Population Replacement Conspiracy Theories (PRCTs) on social media platforms poses significant challenges for content moderation systems and societal cohesion. This paper conducts a comparative analysis of various approaches for detecting PRCT content, with particular focus on their generalization capabilities across platforms and languages. We evaluate several distinct methodologies: pure few-shot learning utilizing Large Language Models (such as Deepseek-V3 and GPT-4o), BERT-based models fine-tuned for this task, and traditional machine learning models. Through analysis of 56,085 YouTube comments and evaluation using a manually annotated gold standard, we found a superior performance of few-shot learning, achieving 94.5% accuracy with DeepSeek and 91.0% with GPT-4o, though DeepSeek showed worse generalization power with higher performance drops in different contexts. These results significantly outperform traditional methods and show robust cross-platform and cross-lingual generalization when tested on multilingual Telegram data. To support reproducibility, both gold-standard datasets and annotation guidelines are made publicly available.

1 Introduction

The rise of Population Replacement Conspiracy Theories (PRCTs) on social media platforms represents a growing challenge for content moderation systems. These theories, which falsely posit orchestrated demographic manipulation through immigration, have evolved from fringe beliefs to widespread narratives that can influence public discourse and policy (Marino et al., 2024). These narratives lay at the core of the extreme right ideology (Ekman, 2022; Bracke and Aguilar, 2024). On platforms like YouTube, where comment sections often serve as breeding grounds for extremist content, PRCTs present unique detection challenges due to their sophisticated use of coded language, implicit rhetoric, and ability to blend with legitimate immigration discourse. Traditional content moderation approaches, primarily relying on keyword matching or machine learning techniques, often struggle to identify these evolving forms of extremist discourse.

Previous work in conspiracy theory detection has largely focused on traditional machine learning approaches, rule-based systems, or BERT-based models, all of which - to different extents - fail to capture the nuanced ways in which these narratives are expressed. While recent advances in Large Language Models (LLMs) offer new possibilities for detecting such content through their enhanced understanding of context and rhetoric, their comparative effectiveness against established methods remains understudied.

Our work aims to advance the field through several contributions to the understanding and detection of PRCTs. We present a comprehensive comparison of detection approaches, ranging from pure few-shot learning with state-of-the-art LLMs (DeepSeek V3 and GPT-4o) to specialized BERT-based models and traditional machine learning approaches including SVM, Random Forest, Logistic Regression, and KNN. This comparison is supported by a novel dataset of 56,085 YouTube comments from immigration-related content, with a manually annotated gold standard for evaluation. Furthermore, we conduct cross-platform validation on golden-labeled Telegram data to assess generalization capabilities across social media platforms.

Our empirical results demonstrate the superior effectiveness of few-shot learning approaches, with DeepSeek and GPT-4o achieving 94.5% and 91.0% accuracy respectively on our gold standard dataset, while maintaining robust performance (respectively, 84.4% and 83.8%) on previously unseen multilingual Telegram data. From an NLP

perspective, these results are particularly significant given the unique challenges PRCTs present through their use of sophisticated rhetorical devices, coded language, and context-dependent messaging. Our findings advance the field by demonstrating effective techniques for detecting implicit ideological content and cross-lingual generalization in conspiracy detection, with implications extending to broader NLP challenges in few-shot learning and the understanding of implicit content.

2 Related Work

Recent advances in computational analysis of conspiracy theories have focused on three main areas: developing detection frameworks, creating annotated datasets, and analyzing linguistic patterns. While these works provide valuable foundations, significant gaps remain in detecting migration-related conspiracy theories and leveraging modern LLM capabilities.

Several studies have developed hierarchical models for conspiracy theory analysis. For instance, Ghasemizade and Onaolapo (2024) created a tree-structured taxonomy of conspiracy theories using transformer-based classifiers (RoBERTa) and clustering techniques, achieving 87% F1-score. Although comprehensive, their approach concentrates on general conspiracy detection rather than specific narratives like population replacement. Similarly, Gambini et al. (2024) analyzed behavioral patterns of conspiracy theorists through a novel Twitter dataset, revealing key differences in terminology and engagement patterns compared to mainstream users; their classifier achieved 94% F1-score using linguistic and behavioral features, yet — like most prior work — focuses on platform-specific detection without cross-platform validation. Along the same lines, Maggini et al. (2024) explored prompting strategies for hyperpartisan detection using Llama3-8b, highlighting how model performance can vary significantly based on specific domain and dataset characteristics.

The creation of specialized corpora has enabled more precise conspiracy theory research. For example, Langguth et al. (2023) developed a COVID-19 conspiracy tweet dataset with 42,000 manual labels across 12 conspiracy categories, demonstrating BERT's effectiveness for multi-task detection. In a similar vein, Pogorelov et al. (2021) focused specifically on 5G-COVID misinformation by creating a 10,000-tweet corpus with conspiracy/non-conspiracy distinctions. Notable work on cross-platform generalization includes Straton (2023), who developed a computational model for COVID vaccine stigma that successfully generalized across Reddit, Twitter, and YouTube, achieving a 0.794 F1-score through deep learning approaches. Similarly, Kim et al. (2023) demonstrated successful cross-lingual generalization in vaccine misinformation detection across three middle-income countries, showing improvements of up to 15.9 percentage points through domain-specific pre-training. While valuable, these datasets primarily target pandemic-related conspiracies rather than demographic narratives. Furthermore, while the 88-million-word LOCO corpus (Miani et al., 2022) includes related themes, such as topic "k300.66" on "immigration, borders, refugees, and migrants", it lacks specific annotations for population replacement theories.

In parallel, recent work has investigated distinguishing features of conspiratorial discourse. Korenčić et al. (2024) developed a multilingual annotation scheme to differentiate conspiracy theories from legitimate criticism, identifying intergroup conflict and violent rhetoric as key discriminators. Their analysis of 10,000 Telegram messages revealed that conspiracy narratives contain 47% more conflict-related language than critical discourse. Similarly, Meuer et al. (2023) compared 72 real-world conspiracy/non-conspiracy articles, finding that conspiratorial explanations rely more on emotional appeals (Cohen's d=1.2) and less on falsifiable claims. In the vaccine domain, Cheatham et al. (2022) demonstrated the effectiveness of transformer models for stance detection, though their approach required periodic retraining to handle language drift. Despite these contributions, three critical gaps remain. First, prior datasets and models focus overwhelmingly on COVID-19/5G conspiracies (COCO, WICO) or general theories (LOCO), with less specific dedicated resources for migration-related narratives. Second, while some studies have shown promising results in cross-platform and cross-lingual generalization for vaccine-related content, most conspiracy detection work (e.g., Gambini et al. (2024)) remains confined to single platforms. Third, the majority of current state-of-the-art methodologies rely on BERT variants or traditional machine learning, without focusing enough

on LLMs' potential for few-shot conspiracy detection. Our work addresses these limitations by providing a large-scale analysis of Population Replacement Conspiracy Theories (PRCTs) across YouTube and Telegram, systematically comparing LLM approaches (DeepSeek, GPT-4o) against traditional ML and BERT variants, and introducing a novel dataset (56K comments) with cross-platform validation and substantiated by a manual annotation achieving 0.891 Gwet's AC1 agreement.

3 Methodology

Our methodological framework encompasses distinct approaches for PRCT detection, evaluated across two social media platforms. This section details our data collection, preprocessing steps, implementation of detection approaches, and evaluation methodology.

3.1 YouTube Dataset

Our primary dataset consists of 56,085 YouTube comments in English. Drawing from (Bassi et al., 2025) we collected 15 videos discussing immigration-related topics. The data collection followed a systematic approach to ensure representativeness and minimize potential biases. Initially, we identified the 100 most-viewed English-language videos from U.S. sources (2013-2024) for immigration-related topics using the YouTube API, with a minimum threshold of 1,000 comments per video. These videos were then ranked based on their comment volume to identify those generating substantial discussions.

Research has shown that source partisanship significantly influences user engagement and comment civility (Su et al., 2018; Törnberg, 2022; Labarre, 2024; Yu et al., 2024). To address this potential bias, we carefully curated our final selection to include a balanced representation of viewpoints across different channel types. The selected content spans mainstream news networks providing border and policy coverage, independent news channels offering alternative perspectives, and specialized content creators focusing on immigration topics. This diversity in source types helps capture different discourse styles and audience interactions. Our final dataset comprises six videos expressing support for immigration through coverage of migrant experiences and policy discussions, six presenting opposing views through border crisis coverage and policy critiques, and three maintaining a neutral perspective through demographic reporting and policy analysis. Table 4 provides a detailed breakdown of these sources and their categorization.

The preprocessing phase was designed to clean the data while preserving its semantic integrity and contextual elements essential for interpretation. We removed comments shorter than 15 characters, which eliminated 2,873 entries, and handled missing values, which removed one additional row. URLs and special characters were normalized to enhance text consistency. Throughout this process, we retained essential metadata including comment timestamps, engagement metrics, and thread structure, resulting in a final dataset of 53,211 comments.

3.2 Telegram Cross-Platform Dataset

To evaluate cross-platform generalization, we collected a multilingual dataset from Telegram channels spanning Italian, Spanish, English, Dutch, and Portuguese content. For our evaluation, we focused specifically on Spanish and Portuguese messages (as these are the languages the authors are most familiar with) creating a balanced test set of 160 messages (80 PRCT, 80 Non-PRCT) manually validated by two annotators with a Cohen's Kappa of 0.8609. For reproducibility purposes, both the anonymized YouTube gold standard and the Telegram test set, along with detailed annotation guidelines and classification prompts, are made publicly available through an anonymous repository[1]. The datasets have been preprocessed to remove personally identifiable information while maintaining their research utility.

3.3 Detection Approaches

3.3.1 Pure Few-Shot Learning

For our few-shot learning experiments, we implemented two approaches using GPT-4o and DeepSeek. To ensure consistency, both models were provided with the same structured prompt. This prompt was carefully designed to guide the classification process by clearly defining the distinction between PRCT and legitimate discussions on immigration. Additionally, they included illustrative examples of both PRCT and Non-PRCT

[1] Anonymous repository: https://anonymous.4open.science/r/PRCT-Detection-Dataset-7748

content to provide the models with concrete reference points. Finally, the instructions explicitly required the models to classify each instance strictly as either "PRCT" or "Non-PRCT", avoiding ambiguous or uncertain classifications. The prompts included explicit PRCT indicators such as "Great Replacement Theory", "White Genocide Theory", "Eurabia", "Kalergi Plan", demographic warfare narratives, and claims of orchestrated population change. Non-PRCT examples encompassed policy discussions, border security concerns, and economic impact analysis without conspiracy elements.

3.3.2 Traditional Machine Learning

To complement our few-shot learning approach, we also developed four machine learning models trained on a balanced synthetic dataset. These models were designed to explore different classification strategies and assess their effectiveness in distinguishing PRCT from non-PRCT content. The first model utilized a Support Vector Machine (SVM) with a linear kernel, leveraging its ability to find optimal decision boundaries in high-dimensional spaces. The second approach applied Logistic Regression combined with TF-IDF vectorization, allowing the model to capture important textual patterns and term relevance. Additionally, we implemented a Random Forest Classifier, which introduced ensemble learning to improve robustness and reduce overfitting. Finally, we experimented with a K-Nearest Neighbors (KNN) model, relying on similarity-based classification to determine label assignments based on proximity to known examples.

The training data comprised 1,700 PRCT comments and 1,700 randomly sampled Non-PRCT comments (identified through few-shot classification - and tested on the gold standard), split 80%/20% for training and testing.

3.3.3 Fine-tuning of BERT Models

Beyond Machine Learning models, our experiments extend to three specialized BERT models, each selected for its unique pre-training and suitability for social media analysis. One of these is CT-BERT (Müller et al., 2023), which has been pre-trained on a vast corpus of COVID-19-related tweets. This model was chosen due to its demonstrated effectiveness in handling social media text, particularly in the context of misinformation and emotionally charged discussions.

Another model included in our study is HateBERT (Caselli et al., 2020), which was trained on data from banned Reddit communities. This specialized pre-training equips HateBERT with a heightened ability to recognize and model abusive and toxic language, making it a relevant candidate for tasks involving online discourse analysis.

As a multilingual baseline, we retain mBERT, a model with strong generalization capabilities across different languages. While all three models share the same fundamental architecture—consisting of 12 transformer layers, a hidden size of 768, and 12 attention heads — they differ significantly in their pre-training objectives and the nature of the data they were exposed to.

For fine-tuning, we apply a uniform training protocol across all BERT models. This involves using a learning rate of 2e-5 and a batch size of 32 while training for a maximum of six epochs. To prevent overfitting and ensure optimal performance, we employ an early stopping mechanism based on the validation set.

3.4 Evaluation Framework

3.4.1 Gold Standard Dataset

To ensure a reliable benchmark for model evaluation, we constructed a manually annotated gold standard dataset consisting of 500 YouTube comments, evenly split between PRCT and Non-PRCT classifications. Each comment was independently reviewed by two expert annotators following a detailed set of annotation guidelines. These guidelines provided clear criteria for identifying PRCT content, outlining key conspiracy narratives and common rhetorical strategies, including implicit dog whistles. The guidelines can be found in the anonymized repository.

3.4.2 Evaluation Metrics

To assess the performance of our classification models, we employed a comprehensive evaluation framework. Standard classification metrics such as accuracy, precision, recall, and F1-score were used to measure overall effectiveness. Beyond these core metrics, we conducted a detailed analysis of false positives and false negatives to identify common misclassification patterns. Additionally, we examined the models' ability to generalize across different platforms, assessing their robustness beyond the YouTube dataset. Finally, an error pattern analysis was carried out to detect system-

atic failures and refine our approach for future iterations.

3.4.3 Inter-annotator Agreement

The analysis of inter-annotator agreement required careful consideration due to the inherent class imbalance in our YouTube validation set, with 473 true-positives versus 27 false-positives for the first annotator, and 446 versus 54 for the second. While traditional metrics showed moderate agreement with Cohen's Kappa at 0.348, these measures are known to be problematic for highly imbalanced datasets. We therefore employed metrics specifically designed for imbalanced data, achieving a Prevalence-Adjusted Bias-Adjusted Kappa (PABAK) of 0.804 and a Gwet's AC1 score of 0.891, indicating substantial to very strong agreement. The class-specific agreement analysis revealed a positive agreement rate of 0.947, a negative agreement rate of 0.395, and an F1-like agreement of 0.947. The observed overall agreement of 90.2%, combined with the strong PABAK and Gwet's AC1 scores, demonstrates good reliability in our annotations, particularly for the crucial PRCT cases that are our primary focus. The lower negative agreement rate reflects the inherent complexity in definitively identifying non-PRCT cases.

Metric	Score
Observed Agreement	90.2%
PABAK	0.804
Gwet's AC1	0.891
Positive Agreement Rate	0.947
Negative Agreement Rate	0.395
F1-like Agreement	0.947

Table 1: Inter-annotator agreement metrics for Youtube dataset showing consistency in PRCT identification

4 Results

4.1 Comparative Performance

Our evaluation reveals significant performance differences across approaches, with LLMs demonstrating superior capabilities in PRCT detection across both platforms, as expected.

4.1.1 YouTube Performance

On the YouTube dataset, DeepSeek achieved the highest performance (94.5% accuracy), followed closely by GPT-4o (91.0%). Both LLMs showed balanced precision and recall, indicating robust detection capabilities across different PRCT manifestations. Traditional ML approaches performed notably lower, with SVM leading at 82.8% accuracy, followed by Logistic Regression (81.1%), Random Forest (78.0%), and KNN (75.9%).

4.1.2 Cross-Platform Generalization

Testing on the Telegram dataset revealed a significant disparity between approaches. Large Language Models demonstrated robust generalization capabilities, with DeepSeek and GPT-4 achieving 84.4% and 83.8% accuracy respectively. In contrast, traditional machine learning methods showed substantial performance degradation, with KNN reaching 56.9% accuracy while SVM, Random Forest, and Logistic Regression performed near chance level at 49.4%, 49.4%, and 48.8% respectively. These results highlight the fundamental advantage of LLMs in cross-platform generalization tasks.

4.2 Error Analysis

Since PRCT narratives are often conveyed implicitly, classifying them is challenging even for human annotators. A qualitative examination of misclassified instances reveals that this difficulty is particularly evident in traditional ML models and, to some extent, in BERT-based models. However, state-of-the-art LLMs demonstrate a stronger ability to capture these narratives effectively. To illustrate these challenges, we present three eloquent examples from our dataset.

In the first example, a Portuguese Telegram comment reads as follows:

> Perda de população portuguesa na última década representa quase 4% do PIB nacional. Portugal perdeu 653 mil Portugueses em idade ativa. Desses 653 mil, quase 200 mil são licenciados. Em contrapartida, Portugal recebe indianos, brasileiros e africanos de baixas qualificações, ocupando cargos como entregador de Uber Eats, limpeza ou restaurantes, como foi anteriormente noticiado no canal. Isto quando não terminam como reclusos. Que futuro podemos esperar quando espantamos os nossos melhores e importamos os piores?[2]

[2]Loss of Portuguese population in the last decade represents almost 4% of the national GDP. Portugal lost 653 thou-

Here, the author does not employ overt terms such as "substituição demográfica" or "invasão", but instead contrasts the loss of a native population with the influx of low-skilled migrants. This implicit framing constructs a narrative of population replacement. Both LLMs and fine-tuned BERT models successfully identified the underlying PRCT narrative, whereas traditional machine learning models, dependent on token-based matching, often failed to flag this kind of comment.

A second example from Youtube utilizes vivid metaphorical language:

> @[USER] People need to take the red pill and realize it is being done on purpose by the left to transform the nation into a one party socialist nation, the border, the crime, it's all on purpose. They will not stop this arson on the nation until they are voted out and stripped of power but time is running out to do so, their plan is to transform the nation so they are never voted out.

This message avoids explicit PRCT keywords and instead leverages metaphors like "red pill" and "arson on the nation" to suggest a deliberate, covert plan to alter national identity. Both LLMs, with their enhanced contextual reasoning, succeeded in detecting the underlying conspiratorial narrative. In contrast, traditional ML models—relying solely on surface-level token patterns—failed to associate these metaphorical cues with PRCT content.

Finally, consider the following comment:

> Presidente do Reino Unido de origem indiana inaugura acordo que promove a entrada de indianos no Reino Unido meras semanas após assumir o cargo. Quanto mais 'diversificado' um país europeu se torna, mais se assemelha a um estado sob ocupação e administração estrangeira.[3]

In this instance, the metaphor of a country transforming into "um estado sob ocupação" is employed to insinuate a loss of national identity due to increasing diversity. Notably, only the LLM-based methods (DeepSeek and GPT-4o) were able to infer the conspiratorial intent embedded in this figurative language, while both the BERT-based models and traditional ML approaches did not capture the implicit PRCT cue.

Together, these three qualitative examples underscore that while fine-tuned BERT models can capture some implicit cues, only LLM-based approaches consistently handle the most nuanced cases.

4.3 Detailed Performance Breakdown

Model	Acc.	Prec.	Rec.	F1
DeepSeek	**0.945**	**0.946**	**0.946**	**0.945**
GPT-4o	0.910	0.911	0.911	0.910
CT-BERT	0.838	0.865	0.833	0.833
HateBERT	0.808	0.845	0.802	0.800
mBERT	0.794	0.846	0.787	0.783
SVM	0.828	0.828	0.828	0.828
Log. Reg.	0.811	0.811	0.811	0.811
Rand. Forest	0.780	0.780	0.780	0.780
KNN	0.759	0.776	0.759	0.756

Table 2: Performance comparison of models on the YouTube dataset

Model	Acc.	Prec.	Rec.	F1
DeepSeek	**0.844**	**0.845**	**0.844**	**0.844**
GPT-4o	0.838	0.845	0.837	0.837
CT-BERT	0.719	0.742	0.719	0.712
HateBERT	0.700	0.724	0.700	0.692
mBERT	0.644	0.689	0.644	0.621
KNN	0.569	0.597	0.569	0.535
Rand. Forest	0.494	0.248	0.494	0.331
SVM	0.494	0.248	0.494	0.331
Log. Reg.	0.488	0.247	0.488	0.328

Table 3: Performance comparison of models on the Telegram dataset

These results demonstrate the robustness of LLMs approaches across different social media platforms. The consistent performance of DeepSeek and GPT-4o on both YouTube and Telegram data suggests their effectiveness in capturing

sand active-age Portuguese. Of these 653 thousand, almost 200 thousand are graduates. In contrast, Portugal receives Indians, Brazilians, and Africans with low qualifications occupying jobs as Uber Eats delivery drivers, cleaners, or restaurant workers, as previously reported on the channel, that is, when they do not end up as prisoners. What future can we expect when we drive away our best and import the worst?

[3]The President of the United Kingdom of Indian origin inaugurates an agreement that promotes the entry of Indians into the United Kingdom just weeks after taking office. The more 'diverse' a European country becomes, the more it resembles a state under occupation and foreign administration.

the fundamental characteristics of PRCT content, regardless of the specific platform conventions or communication styles.

5 Discussion

Our findings highlight the superior capability of LLMs for PRCT detection and discuss practical considerations for content moderation systems.

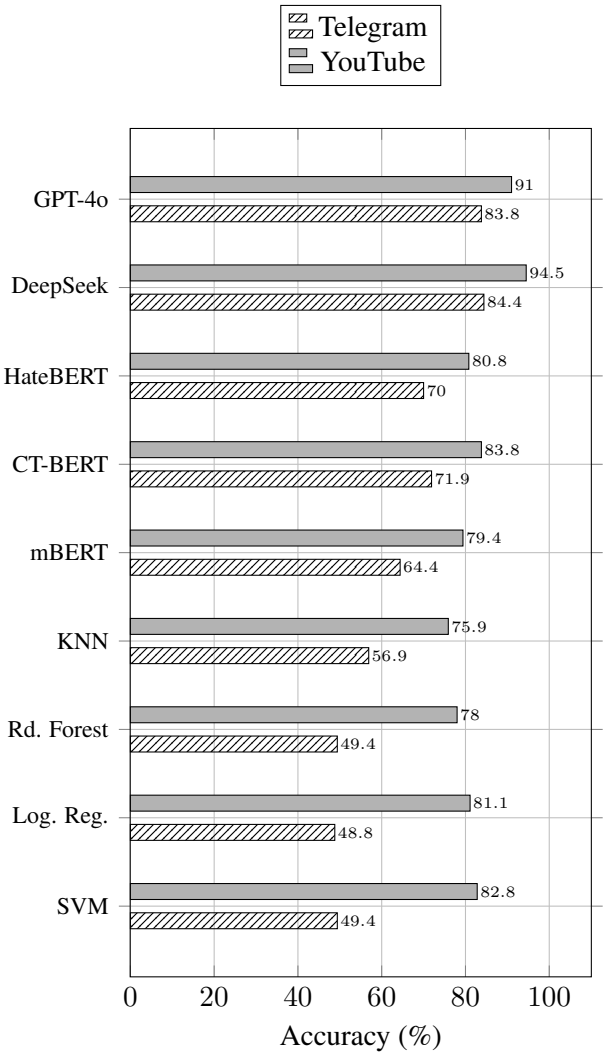

Figure 1: Cross-platform performance comparison between YouTube and Telegram datasets, with models ordered by generalization capability.

5.1 Advantages of LLM Approaches

The superior performance of DeepSeek and GPT-4o, which achieved accuracy rates of 94.5% and 91.0% respectively on YouTube, can be attributed to their advanced capabilities in processing and interpreting complex linguistic patterns. These models excel in understanding rhetoric that is highly context-dependent and often encoded in implicit language. Their ability to generalize across different forms of PRCT narratives allows them to detect variations of the same underlying conspiracy theories, even when expressed in distinct ways. Additionally, their adaptability to platform-specific communication styles enhances their detection accuracy in diverse online environments. A key factor in their strong performance is their multilingual training, which enables them to recognize patterns across languages and dialects.

This adaptability is particularly evident in the cross-platform evaluation. While traditional machine learning approaches suffered from significant performance degradation when applied to Telegram, both DeepSeek and GPT-4o maintained high accuracy rates (84.4% and 83.8% respectively), demonstrating their robustness in handling variations in discourse across different online spaces.

5.2 BERT Models Performance and Trade-offs

The performance of specialized BERT variants offers interesting insights into the value of domain-specific pre-training for conspiracy detection. CT-BERT, with its pre-training on COVID-19 conspiracy content, achieved the highest performance among BERT variants (83.8% on YouTube, 71.9% on Telegram), suggesting that exposure to conspiracy-related discourse patterns during pre-training provides meaningful advantages. However, the substantial gap between CT-BERT and LLMs (approximately 11 percentage points on YouTube) indicates that larger context windows and broader pre-training may be more valuable than domain specialization alone. HateBERT's performance (80.8% on YouTube, 70.0% on Telegram) demonstrates that pre-training on toxic content provides some transferable features for conspiracy detection, though not as effectively as conspiracy-specific pre-training. The baseline mBERT showed the lowest performance among transformer models (79.4% on YouTube, 64.4% on Telegram), indicating that general multilingual capabilities without domain specialization are insufficient for this task. Notably, all BERT variants exhibited more significant performance degradation in cross-platform evaluation compared to LLMs, with accuracy drops ranging from 10-15 percentage points. This suggests that the contextual understanding developed through BERT's pre-training may be more platform-specific than

the broader linguistic patterns captured by LLMs.

These findings highlight a fundamental trade-off: while specialized BERT models offer computational efficiency and decent performance on their target domain, they lack the robust generalization capabilities of LLMs. This is particularly evident in cross-platform scenarios, where the more comprehensive training of LLMs appears to better capture the underlying patterns of conspiratorial discourse across different communication contexts.

5.3 Practical Implications

The differences in performance across detection methods suggest that each approach is best suited for specific use cases. In high-stakes moderation scenarios, where false positives could have severe consequences, LLM-based approaches are preferable due to their superior precision and contextual understanding. On the other hand, BERT-basedmodels, despite their lower accuracy, may still be useful where computational efficiency is a priority for large-scale screening. Finally, the strong generalization capabilities of LLMs make them particularly valuable for cross-platform monitoring, as they can reliably track PRCT content across different digital environments, adapting to the linguistic and stylistic nuances of each platform.

5.4 Future Work

Future research should explore the development of hybrid approaches that strike a balance between accuracy and computational efficiency. A promising direction would be fine-tuning multilingual BERT models to perform emotional analysis as a preliminary step in PRCT detection. Given the established correlation between emotional content and conspiracy theory propagation, incorporating emotional features could enhance detection accuracy while maintaining computational efficiency. Given the continuously evolving nature of these narratives and their demonstrable potential to undermine democratic coexistence, future research must prioritize the development of few-shot learning strategies capable of adapting to emerging PRCT variants. Furthermore, expanding cross-platform validation to additional social media environments would further assess the generalizability of detection models beyond YouTube and Telegram. These efforts would help develop more scalable solutions while maintaining the high accuracy levels demonstrated in our current findings.

6 Conclusion

This study provides empirical evidence supporting the effectiveness of LLM-based approaches in detecting PRCT content across social media platforms. The significantly higher accuracy achieved by DeepSeek (94.5%) and GPT-4o (91.0%) compared to BERT-based models and traditional machine learning models highlights the importance of contextual understanding in identifying evolving conspiracy narratives. These results demonstrate that advanced language models are particularly adept at capturing subtle rhetorical patterns and implicit messaging strategies commonly found in PRCT discourse.

Our findings have direct implications for the design of content moderation systems. LLMs exhibit strong cross-platform generalization, making them valuable tools for detecting harmful content across different digital environments. At the same time, traditional machine learning models, despite their lower accuracy, can still help to reduce computational costs in large-scale detection pipelines.

Ongoing research with fine-tuned BERT models suggests potential for improving detection capabilities while maintaining computational efficiency. Future work should focus on developing scalable solutions that can sustain high accuracy levels while minimizing resource requirements, ensuring that detection systems remain both effective and practical for real-world applications.

Beyond its technical contributions, this study also advances the theoretical understanding of online extremist discourse. By demonstrating the strengths and limitations of different detection approaches, our findings contribute to the broader effort of identifying and mitigating evolving forms of harmful content on social media platforms (Nannini et al., 2024).

7 Ethical Considerations

Our research on PRCT detection raises several important ethical considerations. First, while our work aims to assist in identifying harmful conspiracy content, we acknowledge the fine line between legitimate immigration discourse and conspiracy theories. Our detection systems must be carefully calibrated to avoid suppressing valid political speech or creating chilling effects on public discourse about immigration policy. The high accuracy of our LLM-based approaches (>90%) helps minimize false positives that could unfairly label

legitimate content as conspiratorial. However, we recognize that even a small percentage of misclassifications could impact individuals' ability to participate in important social discussions. To mitigate this risk, we recommend implementing these detection systems as advisory tools rather than automatic content removal systems, with human moderators making final decisions in ambiguous cases. We have also taken steps to ensure ethical data handling. All datasets were anonymized to protect user privacy, with personally identifiable information removed. The YouTube and Telegram data were collected in compliance with platform terms of service and research ethics guidelines. Our annotation process included clear guidelines to maintain consistency and reduce potential bias in labeling. Additionally, we acknowledge the potential dual-use nature of this technology. While intended for identifying harmful conspiracy content, similar techniques could potentially be misused for surveillance or censorship. We therefore emphasize the importance of transparent deployment, clear oversight mechanisms, and strict ethical guidelines for any real-world applications of these detection systems. Finally, our research contributes to the broader goal of maintaining healthy online discourse while protecting vulnerable communities from harmful conspiracy theories. The cross-platform applicability of our approach could help create safer online spaces without unduly restricting legitimate speech about immigration and demographic change.

8 Limitations

Our study has several key limitations. Our cross-platform evaluation was limited to a specific set of language pairs, which may not fully represent the diversity of multilingual PRCT discourse. Plus, the Telegram test set is small, as this is an area of study with limited available resources for the moment. Additionally, the temporal evolution of PRCT narratives presents an ongoing challenge, as these narratives continuously adapt and shift over time, potentially affecting the long-term reliability of our detection methods. Lastly, the high computational cost of LLM-based approaches, including processing power requirements, CO2 emissions, and API expenses, makes large-scale deployment significantly more resource-intensive compared to traditional machine learning methods or BERT-based approaches.

References

Davide Bassi, Michele Joshua Maggini, Renata Vieira, and Martín Pereira-Fariña. 2025. A pipeline for the analysis of user interactions in youtube comments: a hybridization of llms and rule-based methods. In *2024 Eleventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*.

Sarah Bracke and Luis Manuel Hernández Aguilar. 2024. The politics of replacement: from "race suicide" to the "great replacement". In *The Politics of Replacement*, pages 1–19. Routledge.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

S. Cheatham, P. E. Kummervold, L. Parisi, B. Lanfranchi, I. Croci, F. Comunello, M. C. Rota, A. Filia, A. E. Tozzi, C. Rizzo, and F. Gesualdo. 2022. Understanding the vaccine stance of italian tweets and addressing language changes through the covid-19 pandemic: Development and validation of a machine learning model. *Frontiers in Public Health*.

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

M. Gambini, S. Tardelli, and M. Tesconi. 2024. The anatomy of conspiracy theorists: Unveiling traits using a comprehensive twitter dataset. *Computer Communications*.

M. Ghasemizade and J. Onaolapo. 2024. Developing a hierarchical model for unraveling conspiracy theories. *EPJ Data Science*.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024. Large language models for propaganda span annotation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14522–14532, Miami, Florida, USA. Association for Computational Linguistics.

Pavan Holur, Tianyi Wang, Shadi Shahsavari, Timothy Tangherlini, and Vwani Roychowdhury. 2022. 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)*.

Jongin Kim, Byeo Rhee Bak, Aditya Agrawal, Jiaxi Wu, Veronika J Wirtz, Traci Hong, and Derry Wijaya. 2023. Covid-19 vaccine misinformation in middle income countries. In *Empirical Methods in Natural Language Processing 2023*, pages 3903–3915. Association for Computational Linguistics (ACL).

D. Korenčić, B. Chulvi, X. B. Casals, A. Toselli, M. Taulé, and P. Rosso. 2024. What distinguishes

conspiracy from critical narratives? a computational analysis of oppositional discourse. *Expert Systems*.

Julien Labarre. 2024. French fox news? audience-level metrics for the comparative study of news audience hyperpartisanship. *Journal of Information Technology & Politics*, pages 1–18.

J. Langguth, D. T. Schroeder, P. Filkuková, S. Brenner, J. Phillips, and K. Pogorelov. 2023. Coco: An annotated twitter dataset of covid-19 conspiracy theories. *Journal of Computational Social Science*.

Michele Joshua Maggini, Erik Bran Marino, and Pablo Gamallo Otero. 2024. Leveraging advanced prompting strategies in llama-8b for enhanced hyperpartisan news detection.

Erik Bran 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(11):603.

M. Meuer, A. Oeberst, and R. Imhoff. 2023. How do conspiratorial explanations differ from non-conspiratorial explanations? a content analysis of real-world online articles. *European Journal of Social Psychology*.

A. Miani, T. Hills, and A. Bangerter. 2022. Loco: The 88-million-word language of conspiracy corpus. *Behavior Research Methods*.

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.

Luca Nannini, Eleonora Bonel, Davide Bassi, and Michele Joshua Maggini. 2024. Beyond phase-in: assessing impacts on disinformation of the eu digital services act. *AI and Ethics*, pages 1–29.

K. Pogorelov, D. T. Schroeder, P. Filkuková, S. Brenner, and J. Langguth. 2021. Wico text: A labeled dataset of conspiracy theory and 5g-corona misinformation tweets. In *Proceedings of the Workshop on Open Challenges in Online Social Networks (OASIS)*.

Eugenia Ha Rim Rho, Gloria Mark, and Melissa Mazmanian. 2018. Fostering civil discourse online: Linguistic behavior in comments of metoo articles across political perspectives. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. 2023. Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*.

Nadiya Straton. 2023. Covid vaccine stigma: detecting stigma across social media platforms with computational model based on deep learning. *Applied Intelligence*, 53(13):16398–16423.

Leona Yi-Fan Su, Michael A Xenos, Kathleen M Rose, Christopher Wirz, Dietram A Scheufele, and Dominique Brossard. 2018. Uncivil and personal? comparing patterns of incivility in comments on the facebook pages of news outlets. *New Media & Society*, 20(10):3678–3699.

Petter Törnberg. 2022. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42):e2207159119.

Maxwell Weinzierl and Sanda Harabagiu. 2022. Vaccinelies: A natural language resource for learning to recognize misinformation about the covid-19 and hpv vaccines. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Xudong Yu, Magdalena Wojcieszak, and Andreu Casas. 2024. Partisanship on social media: In-party love among american politicians, greater engagement with out-party hate among ordinary users. *Political Behavior*, 46(2):799–824.

A Model Implementation Details

A.1 Machine Learning Models Configuration

Best parameters for each model:

- **SVM**: C=0.1, loss=squared_hinge
  - Best CV score: 0.694
  - Training accuracy: 0.935
  - Test accuracy: 0.590
- **Random Forest**: max_depth=15, min_samples_leaf=2, min_samples_split=5
  - Best CV score: 0.724
  - Training accuracy: 0.902
  - Test accuracy: 0.730
- **Logistic Regression**: C=1.0, penalty=l2
  - Best CV score: 0.696
  - Training accuracy: 0.948
  - Test accuracy: 0.560
- **Naive Bayes**: alpha=1.0
  - Best CV score: 0.709
  - Training accuracy: 0.840
  - Test accuracy: 0.720

Immigration Videos

Pro-Immigration

Chinese migrants fastest growing group crossing into U.S.
https://youtube.com/watch?v=M7TNP2OTY2g

Native American Shuts Down Immigration Protest
https://youtube.com/watch?v=2utsjsWOWUA

Migrants evade Texas floating barrier
https://youtube.com/watch?v=2i8n6jCH1S4

Denmark Leads Anti-Immigration Policies
https://youtube.com/watch?v=zpkBKEPxze4

Immigrant Left U.S. To Seek Asylum In Canada
https://youtube.com/watch?v=ONjCMzB_FPw

Venezuelan Immigrant Regrets Coming to U.S.
https://youtube.com/watch?v=3FPbZcVLTBI

Other/None

Migrant group attempts mass entry at border
https://youtube.com/watch?v=h_TqO9EqMhY

Norway's Muslim immigrants attend classes on women
https://youtube.com/watch?v=oKY600o3CXw

Why does Sweden reject immigrants?
https://youtube.com/watch?v=5CSUimZjiI0

Contra-Immigration

Sweden destroyed by Immigration Crisis
https://youtube.com/watch?v=rUw4cs2MHwc

Migrant crisis reaches boiling point in Staten Island
https://youtube.com/watch?v=-LDra78ksTo

"Deportation, not relocation!" Poland votes
https://youtube.com/watch?v=x4afwGepMkM

Obama Immigration Quote Sounds Racist?
https://youtube.com/watch?v=Vj9IxVlLRl0

Illegal immigrants crisis: Elon Musk visits Texas
https://youtube.com/watch?v=2_iYuiHyzKQ

Migrant beats resident, steals NY home flag
https://youtube.com/watch?v=FTXZmor6KBY

Table 4: YouTube Videos Dataset - Immigration Topics