Linguistic Markers of Population Replacement Conspiracy Theories in YouTube Immigration Discourse

Erik Bran Marino[1], Davide Bassi[2] and Renata Vieira[1]

[1]Universidade de Évora, CIDEHUS, Évora, Portugal
[2]Universidade de Santiago de Compostela, Santiago de Compostela, Spain

Abstract

This paper presents a linguistic analysis of YouTube comments related to immigration discourse, analyzing the contrasts between standard anti-immigration comments and those linked to Population Replacement Conspiracy Theories (PRCT). Using a dataset of 71,137 YouTube comments classified into three stance categories (PRO, NEUTRAL, CONTRA) and PRCT annotation, we analyze the linguistic features of each group through LIWC (Linguistic Inquiry and Word Count). Our findings reveal significant differences in the language patterns of PRCT comments, both in comparison to standard anti-immigration discourse (CONTRA) and to all other groups. These differences appear particularly in religious references, power dynamics, conflict framing, and emotional tone. The high linguistic overlap (89.7%) between conspiracy and non-conspiracy anti-immigration discourse reveals the subtle nature of these differences. These distinctive linguistic patterns provide valuable insights both for the understanding and the automatic detection of conspiracy theories in online discourse, contributing to the growing body of research on computational approaches to identifying harmful content online.

Keywords

Population Replacement Conspiracy Theory, Immigration discourse, YouTube comments, LIWC analysis, LLMs, Deepseek, Hybrid approach, Computational Social Sciences

1. Introduction

Immigration has become one of the central and most controversial topics in cultural and political debates across Western societies. The debate is increasingly influenced by *Population-Replacement Conspiracy Theories* (PRCTs) narratives that portray demographic change as an *élite* plot to replace native populations [1, 2]. Online, the mantra at the core of these narratives—the Great Replacement—has migrated from fringe blogs to mainstream platforms, reshaping how migration is framed and politicised [3].

The impact of PRCTs goes beyond mere rhetoric. Analyses of terrorist manifestos show that the Christchurch (2019) and Utøya (2011) attackers adopted the Great-Replacement frame as moral legitimation for violence [4, 5, 6]. Experimental work further demonstrates that exposure to PRCT claims heightens Islamophobia and support for extremist action [4]. These findings underscore the societal risks tied to PRCT diffusion [5].

Automatic moderation faces two intertwined challenges. First, PRCT cues are lexically sparse, domain-flexible, and embedded in high-volume comment streams, limiting rule-based filters. Second, existing supervised classifiers require large, domain-specific corpora that are rarely available for niche conspiracies [7, 8]. Even state-of-the-art large language models (LLMs) may struggle when prompted zero-shot on conspiracy detection tasks [9, 10].

This study offers a dual contribution:

1. **Methodological**: We provide, to our knowledge, the first systematic evaluation of an open-weight LLM (DeepSeek-v3) for PRCT detection in a few-shot setting. Performance is validated against a gold subset independently annotated by two experts (see §3).
2. **Psycholinguistic**: Using LIWC, we deliver the first fine-grained comparison of PRCT language with other stances in the immigration debate (PRO, CONTRA, NEUTRAL), illuminating differences in temporal focus, power rhetoric and conflict framing [9][1].

These aims translate into two research questions:

RQ1 Can DeepSeek-v3, with minimal in-context examples, reliably distinguish PRCT comments from non-PRCT content?

RQ2 Do PRCT comments exhibit psycho-linguistic patterns that differ systematically from other immigration stances?

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

erik.marino@uevora.pt (E. B. Marino); davide.bassi@usc.es (D. Bassi); renatav@uevora.pt (R. Vieira)
0009-0008-4757-7540 (E. B. Marino); 0000-0003-2025-6559 (D. Bassi); 0000-0003-2449-5477 (R. Vieira)

[1]Throughout this paper we use *psycholinguistic* in the computational-social-science sense: the study of how everyday language reflects basic social and personality processes [11].

2. Related Work

PRCTs comprise a family of narratives such as the *Great Replacement*, the *Kalergi Plan*, *White Genocide* and *Eurabia*. Recent scholarships track their strategic mainstreaming, whereby far-right actors blend demographic alarmism with cultural-defence rhetoric to broaden appeal [1, 2].

In terms of computational approaches to conspiracy detection, early systems combined rule-based extraction with bag-of-words classifiers [8]. More recent pipelines present an automated pipeline using BERT embeddings to discover narrative frameworks in conspiracy theories and conspiracies. Evaluated against expert data, it shows relation extraction recall of 83.7-82.9% for Pizzagate and Bridgegate [7].

Large Language Models offer new possibilities for this domain, promising zero-shot classification without costly annotation. Previous works shows that GPT-3.5 and LLaMA-2 outperform RoBERTa on generic conspiracy tasks but inflate false-positive rates [12, 13]. However, no prior study evaluates DeepSeek on PRCT specifically, leaving a clear research gap that we address.

From a linguistic perspective, corpus studies reveal that conspiracy texts favour future-oriented temporal frames, certainty language and out-group pronouns [8, 7]. Our work isolates PRCT language to test whether it is merely an intensification of generic anti-immigration talk or a qualitatively distinct register. In this context, LIWC remains a widely validated tool for psycholinguistic profiling. In extremist contexts it is able to capture cues pertinent to radical rhetoric [14]. Yet its capacity to discriminate between sub-types of anti-immigration discourse goes beyond its goals. By integrating LIWC with stance labels, we extend its interpretive utility.

Overall, the literature lacks (i) validated LLM approaches for PRCT detection and (ii) systematic linguistic characterisation that separates PRCT from non-conspiratorial rhetoric. Our study addresses both gaps, laying empirical foundations for future detection pipelines and theory-driven analyses of demographic conspiracy talk. Furthermore, Hernaiz [15] theorizes that conspiracy theories operate within the same *secular rational frame* as mainstream explanations, suggesting that linguistic differences between conspiracy and non-conspiracy discourse may be more subtle than categorical, warranting empirical investigation of their shared and distinct features.

3. Methodology

3.1. Dataset

Our analysis is based on a dataset comprising 71,137 unique YouTube comments related to immigration. Specifically, we expanded the dataset described in Bassi et al. [16] by crawling a total of 15 videos about immigration (see Table 7 in the appendix for complete video list). Following the methodology established in the referenced study, which demonstrated that parent comment contextual information is crucial for accurate stance detection in YouTube comments, we employed the same hybrid pipeline to reconstruct conversation chains and preserve parent-child relationships between comments.

For stance classification, we utilized GPT-4o with contextual information from reconstructed comment chains to detect the stance of the comments. The vast majority of comments mention migration. The classification scheme distinguished between three primary categories:

- **CONTRA**: expressing anti-immigration views
- **NEUTRAL**: expressing neutral, unclear or unrelated perspectives towards immigration
- **PRO**: expressing pro-immigration views

A detailed performance evaluation of GPT-4o for immigration-related stance labelling is provided in [16]. The model achieved a $macro - F1 = 78.7\%$ on a manually labelled subset, demonstrating sufficient accuracy to enable automated annotation across the entire dataset.

Subsequently, the comments were further analyzed using DeepSeek v3 in a few-shot learning approach to identify those containing Population Replacement Conspiracy Theory elements, resulting in the PRCT annotation. The classification process employed carefully structured prompts that included reference examples extracted directly from the existing labeled dataset (5 PRCT examples and 5 Non-PRCT examples) to guide the model's understanding. Representative PRCT and Non-PRCT examples for the few-shot prompt were drawn from the training pool via stratified random sampling across the 15 videos, balancing length, topic, and stance. The five PRCT instances include both explicit markers (e.g. explicit mention of "Great Replacement") and implicit cues (coded dog-whistles such as "demographic engineering"); likewise, the five Non-PRCT examples span policy-oriented, economic, and security-focused objections free of conspiratorial framing. The prompts featured explicit definitions of PRCT content, encompassing specific conspiracy narratives such as "Great Replacement Theory", "White Genocide Theory", "Eurabia", and "Kalergi Plan", as well as broader indicators like demographic warfare narratives, terms such as "invasion", "replacement", and "remigration", and claims of orchestrated population change. Non-PRCT examples were defined to include policy discussions, border security debates, integration challenges, and economic impact analysis without conspiracy elements. The model was configured with temperature=0 to ensure deterministic and reproducible classifications, and was explicitly instructed to respond strictly with either

"PRCT" or "Non-PRCT", avoiding ambiguous classifications. To ensure the reliability of our PRCT classification, we validated DeepSeek v3's performance using a manually annotated gold standard dataset of 500 YouTube comments, evenly split between PRCT and Non-PRCT classifications[2]. Each comment was independently reviewed by two expert annotators following detailed annotation guidelines that provided clear criteria for identifying PRCT content. The inter-annotator agreement demonstrated high reliability with Gwet's AC1 = 0.891 and PABAK = 0.804, indicating substantial agreement particularly for PRCT identification (Positive Agreement Rate: 0.947). DeepSeek v3 achieved 94.5% accuracy on this gold standard, with balanced precision and recall, demonstrating robust detection capabilities across different PRCT manifestations.

This methodology allowed us to create a comprehensive dataset that distinguishes between standard anti-immigration discourse and discourse specifically containing population replacement conspiracy theories. Given the nature of our study, we proceeded by removing duplicated comments and applying a word count filter to retain comments between 5 and 1000 words, ensuring sufficient content for meaningful analysis while excluding extremely short or excessively long comments. Table 1 describes the final distribution of stance and PRCT annotations in our dataset.

Category	Count (%)
Stance	
CONTRA	37,531 (52.76%)
NEUTRAL	22,190 (31.19%)
PRO	11,416 (16.05%)
PRCT	
Non-PRCT	65,915 (92.66%)
PRCT	5,221 (7.34%)
Total Dataset	**71,137 (100.00%)**

Table 1
Distribution of stance categories and PRCT annotations in the dataset

Within the CONTRA stance category, 4,905 comments (13.07%) contained PRCT elements, while 32,625 comments (86.93%) were standard anti-immigration discourse without conspiracy theories. This distinction forms the basis of our comparative linguistic analysis.

3.2. LIWC Analysis

To analyze the linguistic characteristics of each comment category, we utilized the Linguistic Inquiry and Word Count (LIWC) tool. LIWC is a text analysis software that calculates the percentage of words pertaining to specific dictionaries falling into specific psychological and linguistic categories [17].

We processed all comments through LIWC, focusing on the following key dimensions:

Temporal focus: refers to the extent to which individuals characteristically direct their attention to the past, present, and future [18]. LIWC derives temporal focus scores by counting the frequency of time-related words in text. For example, past focus includes words like "ago" or "did;" present focus captures "today," "is," and "now," while future focus is based on "may," "will," and "soon"[19].

Pronoun usage: Pronoun use highlights whether attention is on others—third-person singular/plural (he/she, they), on ourselves as distinct entities—first-person singular pronouns (I), or ourselves embedded within a social relationship—first-person plural (we) and second-person (you) [20].

Cognitive processes: This dictionary comprises over 1,000 entries that identify active information-processing; it yields six sub-scores (insight, causation, discrepancy, tentativeness, certainty and differentiation) [21]. These dimensions capture the depth and style of mental elaboration, indicating whether individuals are reasoning analytically (causation, insight), expressing uncertainty or confidence (tentativeness, certainty), or making distinctions and comparisons (differentiation, discrepancy).

Emotional dimensions: LIWC distinguishes between broad sentiment and specific emotions [22]. The affect category encompasses both positive tone (e.g., "good," "love," "happy") and negative tone (e.g., "bad," "hate," "hurt") words, which reflect general sentiment. The emotion categories are more targeted, focusing on specific emotion labels such as positive emotion (e.g., "joy," "excited"), negative emotion (e.g., "sad," "angry"), and discrete emotional states including anxiety (e.g., "worry," "fear"), anger (e.g., "mad," "frustrated"), and sadness (e.g., "disappointed," "cry") [19]. These dimensions capture both the valence and intensity of emotional expression in text.

Social dynamics: this dictionary captures references to interpersonal relationships and social behaviors, including social referents (e.g., "you," "we"), prosocial behavior (e.g., "help," "care"), conflict (e.g., "fight," "argue"), and communication acts (e.g., "said," "tell"). The framework also measures power-related language reflecting awareness of social hierarchies and clout, which captures confidence or leadership displayed through language [19, 20].

Linguistic style: captures stylistic markers (such as usage of exclamation and question marks, or periods) which can reflect formality or communicative intent [19].

For each category, we averaged LIWC scores and conducted comparative analyses to identify significant

[2]Detailed annotation criteria for the PRCT validation task are publicly available at https://zenodo.org/records/16605519.

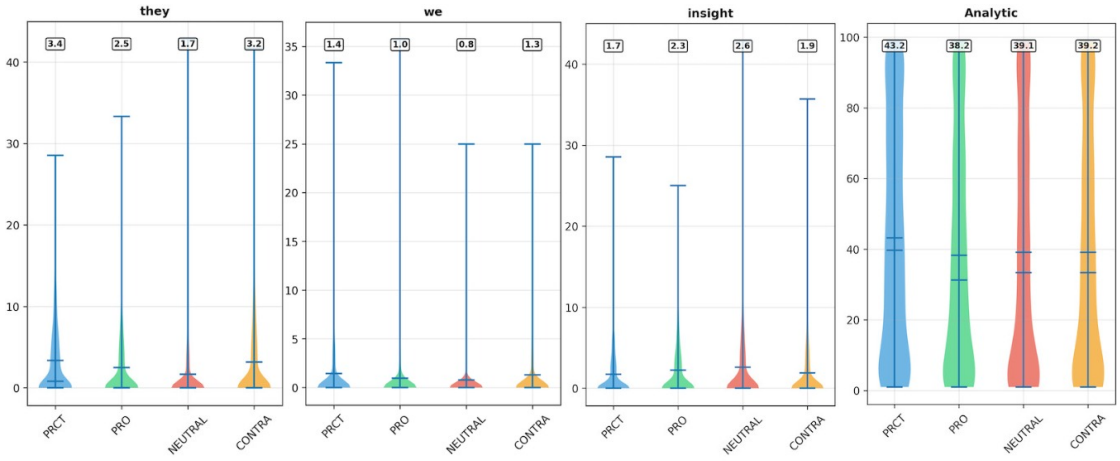

Figure 1: Dimensions in which both anti-immigration and PRCT groups differ from the other stances

differences, particularly between CONTRA-PRCT (the 4,905 merged class) comments and other categories. We adopted an exploratory approach, running the complete LIWC dictionary and retaining all variables for analysis. Figure 3 displays the subset that reached $|d| > 0.2$ after multiple-comparison correction; these include both single-word scores (e.g. *religion*) and composite categories (e.g. *analytic*).

3.3. Statistical Analysis

Statistical Test Selection: given the large sample sizes and non-normal distributions typical of linguistic data, for each dimension, we assessed normality conditions through Shapiro-Wilk and homogeneity of variance using Levene's test. Normality assumption was violated in all 39 cases, hence we recurred to Kruskal-Wallis test.

Multiple Comparison Correction: Given the exploratory nature of our research (comparison of multiple LIWC dimensions across 4 different groups), we applied multiple comparison corrections. Specifically, False Discovery Rate (FDR), and Bonferroni Correction to identify most robust effects.

Effect Size: for each significant difference, we calculated Cohen's d. In this regard, we highlight how usually effect sizes $0.2 \leq |d| \geq 0.5$ are considered small, however we considered effect sizes of $|d| > 0.2$ as substantial, in line with field-specific benchmarks for linguistic research [23, 24].

Two-Phase Analysis: Our analytical approach comprised two phases: (1) a comprehensive four-group comparison (CONTRA-PRCT, CONTRA, NEUTRAL, PRO) to establish general immigration discourse patterns, and (2) a focused binary analysis (CONTRA-PRCT vs CONTRA) to identify features specifically distinguishing conspiracy content from general anti-immigration rhetoric. The binary comparison directly addresses whether conspiracy theories represent fundamentally different discourse or an intensification of existing patterns.

PRCT-Specific Feature Classification: We categorized the LIWC dimensions as either *PRCT-specific* (statistically significant after FDR correction with $|d| \geq 0.2$) or *shared features* ($|d| < 0.2$). The overlap percentage was calculated as the proportion of shared features relative to total features analyzed.

4. Results

Our analysis revealed distinct linguistic patterns in immigration-related discourse, with significant differences between stance groups while highlighting substantial overlap between conspiracy and non-conspiracy anti-immigration rhetoric.

4.1. General Immigration Discourse Patterns

The comprehensive four-group comparison (CONTRA-PRCT, CONTRA, NEUTRAL, PRO) revealed systematic linguistic differences across immigration stances. After applying FDR correction for multiple comparisons, the majority of LIWC dimensions showed significant differences ($p_{\text{FDR}} < 0.05$).

Anti-Immigration vs Pro-Immigration Discourse. As shown in Figure 1, both anti-immigration groups (CONTRA-PRCT and CONTRA) demonstrated a similar depersonalised rhetoric, signalled by a higher usage of third-person plural pronouns (*they*), reflecting out-group focus, and first-person plural pronouns (*we*),

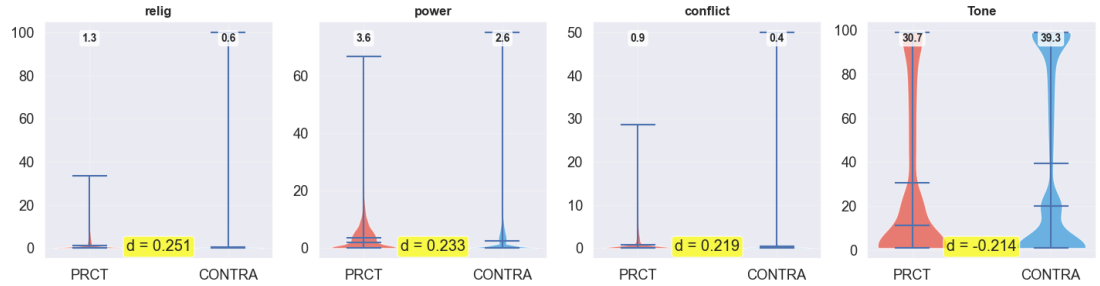

Figure 2: PRCT-Specific Features: Violin plots showing the distribution of the four dimensions that distinguish conspiracy discourse from standard anti-immigration rhetoric

signalling in-group consolidation, compared to PRO and NEUTRAL comments.Specifically, "They" pronouns: CONTRA-PRCT (3.36), CONTRA (3.17) vs PRO (2.51) vs NEUTRAL (1.68); and "We" pronouns: CONTRA-PRCT (1.43), CONTRA (1.30) vs PRO (0.97) vs NEUTRAL (0.77). Additionally, PRCT discourse exhibited distinct cognitive processing patterns. PRCT comments showed the highest analytic thinking scores (43.2) compared to all other groups (PRO: 38.2, NEUTRAL: 39.1, CONTRA: 39.2), suggesting more structured, logical reasoning style. Conversely, PRCT comments demonstrated lower insight language usage (1.7) compared to PRO (2.3) and NEUTRAL (2.6) groups, indicating less expression of sudden understanding or realization. This pattern can indicate that while PRCT discourse employs analytical framing, it may rely more on predetermined interpretive frameworks rather than exploratory or discovery-oriented thinking.

4.2. PRCT-Specific Linguistic Signature

To isolate features unique to conspiracy discourse from general comments against immigration, we conducted a focused binary comparison between CONTRA-PRCT (n=4,905) and CONTRA non-PRCT (n=32,625) comments. This analysis revealed a striking finding: 89.7% of linguistic features showed negligible differences (Cohen's d < 0.2) between conspiracy and non-conspiracy anti-immigration discourse, suggesting that anti-immigration discourse, regardless of conspiracy content, shares fundamental characteristics of outgroup construction and authoritative positioning. As shown in Figure 2, only four dimensions exceeded the meaningful effect size threshold.

As shown in Figure 3, four dimensions demonstrated meaningful effect sizes ($d \geq 0.2$) with statistical significance after FDR correction:

Religion ($d = 0.251$, $p_{\text{FDR}} < 0.001$; CONTRA-PRCT: 1.274 vs CONTRA: 0.591): PRCT discourse shows 115.6% higher usage of religious language, reflecting the framing of demographic change as a spiritual or civilizational threat.

Power Language ($d = 0.233$, $p_{\text{FDR}} < 0.001$; CONTRA-PRCT: 3.621 vs CONTRA: 2.560): PRCT discourse shows 41.4% higher usage of power-related language, reflecting emphasis on elite control and orchestrated manipulation.

Conflict Framing ($d = 0.219$, $p_{\text{FDR}} < 0.001$: CONTRA-PRCT: 0.853 vs CONTRA: 0.437): Conspiracy discourse frames immigration as active conflict/warfare with 95.2% higher conflict language usage.

Tone ($d = -0.214$, $p_{\text{FDR}} < 0.001$; CONTRA-PRCT: 30.674 vs CONTRA: 39.347): PRCT comments exhibit significantly more negative tone, with 22.0% lower positive sentiment scores than standard anti-immigration discourse.

5. Discussion

The linguistic patterns identified in our analysis offer significant insights into the nature of PRCT discourse and its distinction from standard anti-immigration rhetoric. Our findings reveal that while conspiracy and non-conspiracy anti-immigration discourse share 89.7% of their linguistic features, they differ significantly in four key dimensions: religious references, power dynamics, conflict framing, and emotional tone.

5.1. High Linguistic Overlap

A potential limitation is that the *Non-PRCT* comparison set, although explicitly anti-immigration, aggregates heterogeneous sub-registers (security, economic, assimilationist). This breadth may inflate the observed linguistic overlap. Nevertheless, the residual differences we detect—religious framing, power attribution, conflict, and tone—remain interpretable within Hernaiz [15]'s framework of shared rational frames, suggesting that

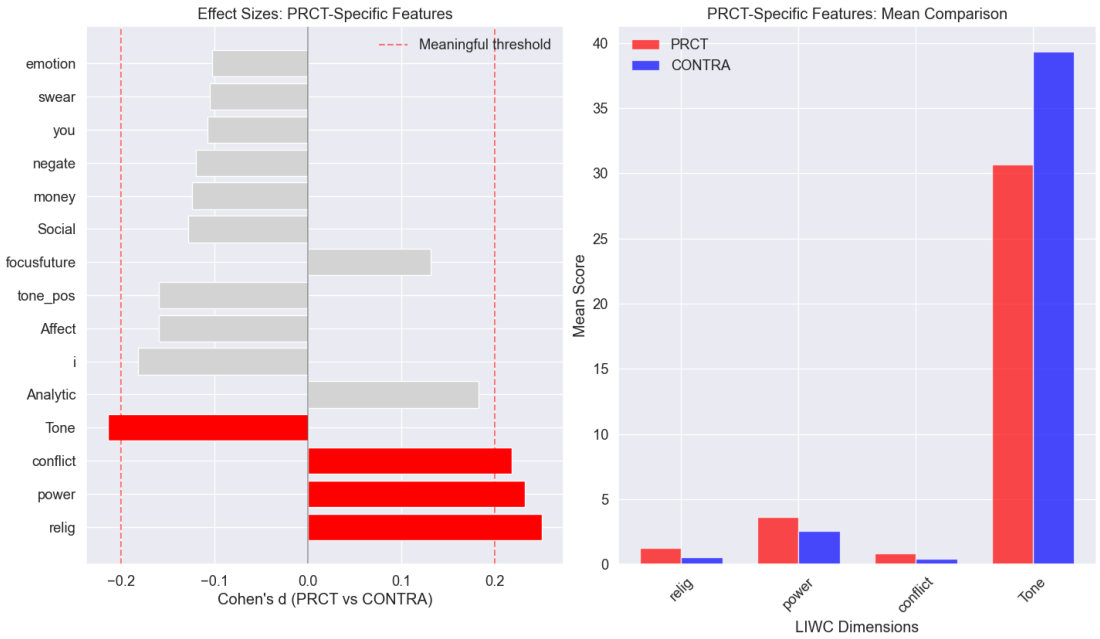

Figure 3: Overview of LIWC dimensions analysis. On the left: Effect sizes (Cohen's d) for the top 15 dimensions by magnitude (largest effect sizes) from the comprehensive LIWC analysis, with red bars indicating dimensions exceeding the meaningful threshold ($|d| \geq 0.2$). On the right: Mean comparison of the four significant dimensions between PRCT and standard anti-immigration (CONTRA) comments. Note: For the Tone dimension, higher values indicate more positive emotional expression.

PRCT discourse intensifies, rather than qualitatively departs from, mainstream anti-immigration rhetoric. The substantial overlap (89.7%) between PRCT and standard anti-immigration discourse, in fact, aligns with Hernaiz [15]'s theoretical framework of shared *secular rational frames*. Rather than representing fundamentally different discourses, conspiracy theories may intensify existing rhetorical patterns while operating within the same rational framework as mainstream explanations. Our finding of high ANALYTIC thinking combined with low INSIGHT language suggests that PRCT commenters employ analytical reasoning to validate existing beliefs rather than explore new understandings, potentially reflecting the confirmatory versus exploratory cognitive distinction [15]. This high overlap could pose challenges for automated detection systems but provides valuable insights for understanding how conspiracy narratives emerge from and relate to mainstream discourse.

5.2. PRCT-Specific Features

The four distinctive features of PRCT discourse, as visualized in Figure 3, provide interesting insights into its conceptual structure:

Religious language (d = 0.251): The significantly higher use of religious terminology in PRCT comments reflects the framing of immigration as not merely a political or economic issue, but as a threat to cultural and spiritual identity. This finding aligns with Hernaiz [15]'s observation that conspiracy theories operate within a hybrid framework, employing rational secular arguments while simultaneously appealing to notions of "faith" and "belief" that pair them with religious explanations. Like religious narratives, PRCT discourse ascribes demographic change to volitional agents with malevolent intent, transforming a social phenomenon into a spiritual or civilizational crisis. This supports previous findings that replacement conspiracy theories present demographic change as an existential threat to a civilization's core values [1]. This pattern manifests empirically in comments such as *"Have you heard of Islamic Jihad? that's most likely why.....Islamization!!"* (relig=27.27), where immigration becomes reframed as deliberate religious warfare rather than demographic movement, directly invoking the Eurabia conspiracy framework that portrays Muslim immigration as orchestrated civilizational replacement.

Power dynamics (d = 0.233): The emphasis on power-related language could reflect the classic conspiratorial view that demographic changes are orchestrated by powerful elites rather than resulting from natural social pro-

cesses. This finding corroborates studies showing that attribution of agency and intentionality to shadowy power centers is a defining characteristic of conspiracy thinking [10, 25]. The linguistic manifestation of this attribution appears in constructions such as *"Import third world → become third world"* (power=66.67), where the verb "import" transforms organic migration processes into deliberate elite manipulation. This deterministic arrow formulation removes agency from migrants themselves while implying the existence of powerful orchestrators capable of engineering demographic transformation, exemplifying how power-related language shifts explanatory frameworks from socio-economic to conspiratorial causation.

Conflict framing (d = 0.219): PRCT discourse shows nearly double the rate of conflict terminology compared to standard anti-immigration comments (0.85% vs 0.44%), representing a 95.2% relative increase. This signals how conspiracy theories transform social issues into existential struggles between groups [26]. This Manichean framing can serve to legitimize more extreme responses, as demonstrated by Bracke and Aguilar [6]. The militarization of discourse materializes in statements like *"aggressively defending our borders from invaders"* (conflict=25.00), where immigration policy becomes reconceptualized as warfare requiring defensive military action. The lexical choice of "invaders" transforms migrants from policy subjects into military threats, while "aggressively defending" positions exclusionary responses as legitimate self-defense, illustrating how conflict framing escalates immigration discourse from policy debate to existential combat.

Negative tone (d = -0.214): The markedly more negative emotional tone of PRCT discourse, with 22.0% lower positive sentiment scores, shows the affective dimension of conspiracy theories. This emotional negativity may function as a mobilizing mechanism, generating moral outrage and urgency [4]. This heightened negativity appears in apocalyptic formulations such as *"Most of Europe has been destroyed because of illegal immigrants"* (tone_neg=30.00), where the verb "destroyed" escalates beyond policy criticism to civilizational annihilation. The continental scope ("Most of Europe") and direct causal attribution ("because of") exemplify how PRCT discourse employs catastrophic language to transform demographic statistics into existential crisis narratives, intensifying emotional engagement through linguistic extremity.

These four linguistic markers offer insights for both socio-psychological understanding of conspiracy discourse and the development of computational detection systems, providing empirically grounded features that could enhance automated identification of PRCT content online. While this study isolates Population-Replacement Conspiracy Theories, the four linguistic dimensions we identify—religious sacralization, elite power attribution, conflict framing and negative affect—map closely onto defining features documented in other conspiracy families (e.g., QAnon, anti-vaccination, or Great Reset narratives). Future work can test whether these markers generalize across domains, turning the present fine-grained analysis into a broader framework for detecting conspiratorial escalation in online discourse.

5.3. Socio-Linguistic Mechanisms in PRCT Discourse: Theoretical Perspectives

The linguistic patterns identified in our analysis invite broader theoretical reflections on the socio-linguistic mechanisms underlying PRCT discourse. While acknowledging the limitations of drawing definitive conclusions from a single study with an English-language YouTube dataset, the distinctive features we observed suggest several promising avenues for theoretical exploration. The high linguistic overlap (89.7%) between PRCT and standard anti-immigration discourse suggests what might be conceptualized as a rhetorical continuum rather than a categorical distinction. This finding resonates with the concept of the Overton window [27] - the range of politically acceptable discourse at a given time. Rather than emerging as entirely separate discourses, conspiracy narratives may represent incremental shifts along this continuum, potentially facilitating the mainstreaming of fringe ideas through gradual rhetorical transformations. Within this continuum, we observe that the significantly higher use of religious terminology in PRCT comments (+115.6%) might reflect the so-called *sacralization of collective identity* - a process through which political issues are transformed into matters of existential and moral value [28]. While our data cannot establish causality, this linguistic pattern aligns with Girard's (2020) theory of sacred differentiation, where boundaries between in-group and out-group acquire quasi-religious significance. The emphasis on power-related language (+41.4%) in PRCT discourse further connects to what Hofstadter [30] termed the paranoid style in political rhetoric - the perception of systematic, malevolent orchestration behind social phenomena. This linguistic pattern may reflect the construction of alternative relevance structures through which events are reframed as evidence of hidden designs [31]. Equally notable is the substantial increase in conflict terminology (+95.2%), suggesting a potential militarization of the interpretive frame that transforms political debate into existential struggle. This might create what Bauman [32] characterizes as a *discursive state of emergency* in which exceptional responses become justified by the perception of imminent threat. Such framing represents not merely a rhetorical choice but a fundamental shift in how immigration discourse is conceptualized and processed. These theoretical perspectives collectively suggest several promising directions for

future research. Longitudinal studies could track the evolution of these linguistic markers over time to understand how discursive shifts occur. Comparative analyses across different languages and cultural contexts would test the generalizability of these patterns, while experimental studies might investigate how exposure to these specific linguistic features affects audience perceptions and beliefs. It is important to emphasize that these theoretical interpretations remain speculative based on our limited dataset. The patterns we observed offer intriguing correlations, but establishing causal relationships between these linguistic features and the social mechanisms described would require more extensive mixed-methods research combining computational and qualitative approaches. Nevertheless, these preliminary findings suggest that the subtle linguistic distinctions between conspiracy and non-conspiracy discourse may reveal deeper social and cognitive processes worthy of further investigation. Future research might investigate whether the transition from mainstream to conspiratorial discourse follows predictable linguistic trajectories, and how immigration discourse becomes embedded within broader civilizational or existential frames.

6. Conclusion

This study advances both methodological and theoretical fronts. **RQ1** asked whether DeepSeek-v3 can reliably detect PRCT content with minimal examples; our validation on a 500-comment gold set (§3) confirms 94.5 % accuracy (balanced precision/recall), demonstrating that a LLM in a few-shot regime is adequate for this task. **RQ2** examined whether PRCT comments exhibit distinct psycho-linguistic patterns; the comparison revealed four robust markers—religious references, power dynamics, conflict framing and negative tone—that systematically differentiate PRCT from standard anti-immigration discourse.

While 89.7 % of linguistic features are shared between conspiracy and non-conspiracy anti-immigration comments, the four PRCT-specific dimensions remain stable and interpretable. These findings underscore that conspiracy narratives often intensify, rather than abandon, mainstream rhetorical frames, and they provide empirically grounded cues for automated moderation systems.

7. Limitations and Ethical Considerations

While our study reveals significant linguistic patterns in PRCT discourse, several limitations and ethical considerations warrant discussion. Our analysis focuses on English-language YouTube comments, which may limit generalizability to other platforms and languages where conspiracy discourse could manifest differently. The automatic classification process, though effective with high agreement scores, inevitably introduces some risk of misclassification that future work might address through additional validation approaches or multi-platform comparisons.

Regarding data handling, our research relies on user-generated content from public YouTube videos, raising important privacy considerations. We conducted this research in accordance with GDPR Article 9(2)(j) and Article 89, which permit processing of potentially sensitive data for research purposes with appropriate safeguards. Throughout our analysis, we removed personal identifiers from collected comments, focused on aggregate linguistic patterns rather than individual profiles, and maintained secure data storage with restricted access. Although the YouTube videos themselves remain publicly accessible, we do not publish the raw comment data openly to protect user privacy. Researchers interested in accessing the dataset for scientific purposes may contact the authors with appropriate research ethics documentation, with any data sharing conducted in compliance with GDPR and relevant national regulations.

This research also raises broader ethical questions about the study and identification of conspiracy theories online. While identifying linguistic markers of potentially harmful content could facilitate better content moderation, we recognize the complex balance between reducing harmful misinformation and protecting legitimate discourse. The high linguistic overlap (89.7%) between conspiracy and non-conspiracy anti-immigration discourse underscores the subtlety of these distinctions and the risks of over-moderation based solely on automated detection. Our findings should be interpreted as identifying patterns across large samples, not as definitive classifiers for individual comments. This complexity highlights the importance of human oversight in content moderation systems that might leverage these linguistic insights.

Acknowledgments

This research was conducted as part of a larger project focused on detecting disinformation such as conspiracy theories in online discourse. The authors would like to thank their supervisors and colleagues for their guidance and support throughout this research. We are particularly grateful to Katarina Laken for her valuable contributions and insightful advice. This work was supported by the HYBRIDS project, which has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351 and from the UK Re-

search and Innovation (UKRI) Horizon Europe funding guarantee (Grant Number: EP/X036758/1). The work is partially supported by the Portuguese Science Foundation as part of the projects CEECIND/ 01997/2017 and UIDP/00057/2025. The content of this work reflects only the authors' view and the funding agencies are not responsible for any use that may be made of the information it contains.

References

[1] M. Ekman, The great replacement: Strategic mainstreaming of far-right conspiracy claims, Convergence 28 (2022) 1127–1143.

[2] M. Sedgwick, The great replacement narrative: Fear, anxiety and loathing across the west, Politics, Religion & Ideology 25 (2024) 548–562. doi:10.1080/21567689.2024.2424790.

[3] E. B. Marino, J. M. Benitez-Baleato, A. S. Ribeiro, 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 (2024) 603.

[4] M. Obaidi, J. R. Kunst, S. Ozer, S. Y. Kimel, The "great replacement" conspiracy: How the perceived ousting of whites can evoke violent extremism and islamophobia, Group Processes & Intergroup Relations 25 (2021) 1675–1695. doi:10.1177/13684302211028293.

[5] M. Davis, Violence as method: The "white replacement", "white genocide", and "eurabia" conspiracy theories and the biopolitics of networked violence, Ethnic and Racial Studies (2024). doi:10.1080/01419870.2024.2304640, advance online publication.

[6] S. Bracke, L. M. H. Aguilar, The politics of replacement: from "race suicide" to the "great replacement", in: The politics of replacement, Routledge, 2023, pp. 1–19.

[7] S. Shahsavari, T. R. Tangherlini, B. Shahbazi, E. Ebrahimzadeh, V. Roychowdhury, An automated pipeline for the discovery of conspiracy and conspiracy-theory narrative frameworks, PLOS ONE 15 (2020) e0233879. doi:10.1371/journal.pone.0233879.

[8] M. Samory, T. Mitra, Conspiracies online: User discussions in a conspiracy community following dramatic events, in: Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018, AAAI Press, 2018, pp. 340–349. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17907.

[9] M. Hunter, T. Grant, Is linguistic inquiry and word count (liwc) reliable, efficient, and effective for the analysis of large online datasets in forensic and security contexts?, Applied Corpus Linguistics 5 (2025) 100118. doi:10.1016/j.acorp.2025.100118.

[10] A. Platt, J. Brown, A. Venske, Toward detecting conspiracy language in misinformation documents, in: Proceedings of the 2022 Computers and People Research Conference (SIGMIS–CPR '22), 2022. doi:10.1145/3510606.3551895.

[11] J. W. Pennebaker, The secret life of pronouns, New Scientist 211 (2011) 42–45.

[12] T. Vergho, J.-F. Godbout, R. Rabbany, K. Pelrine, Comparing gpt-4 and open-source language models in misinformation mitigation, arXiv preprint arXiv:2401.06920 (2024).

[13] A. Kumar, R. Sharma, P. Bedi, Towards optimal nlp solutions: analyzing gpt and llama-2 models across model scale, dataset size, and task diversity, Engineering, Technology & Applied Science Research 14 (2024) 14219–14224.

[14] A. Etaywe, K. Macfarlane, M. Alazab, A cyberterrorist behind the keyboard: An automated text analysis for psycholinguistic profiling and threat assessment, Journal of Language Aggression and Conflict (2024).

[15] H. A. P. Hernaiz, Competing explanations of global evils: Theodicy, social sciences, and conspiracy theories, AGLOS: journal of area-based global studies 2 (2011) 27.

[16] D. Bassi, M. J. Maggini, R. Vieira, M. Pereira-Fariña, A pipeline for the analysis of user interactions in youtube comments: A hybridization of llms and rule-based methods, in: 2024 11th International Conference on Social Networks Analysis, Management and Security (SNAMS), 2024, pp. 146–153. doi:10.1109/SNAMS64316.2024.10883781.

[17] Y. R. Tausczik, J. W. Pennebaker, The psychological meaning of words: Liwc and computerized text analysis methods, Journal of Language and Social Psychology 29 (2010) 24–54. doi:10.1177/0261927X09351676.

[18] S. J. Barnes, Stuck in the past or living in the present? temporal focus and the spread of covid-19, Social Science & Medicine 280 (2021) 114057. doi:https://doi.org/10.1016/j.socscimed.2021.114057.

[19] R. L. Boyd, A. Ashokkumar, S. Seraj, J. W. Pennebaker, The development and psychometric properties of liwc-22, Austin, TX: University of Texas at Austin 10 (2022) 1–47. URL: https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf.

[20] E. Kacewicz, J. W. Pennebaker, M. Davis, M. Jeon,

A. C. Graesser, Pronoun use reflects standings in social hierarchies, Journal of Language and Social Psychology 33 (2014) 125–143. doi:https://doi.org/10.1177/0261927X13502654.
[21] R. L. Moore, C.-J. Yen, F. E. Powers, Exploring the relationship between clout and cognitive processing in mooc discussion forums, British Journal of Educational Technology 52 (2021) 482–497. doi:https://doi.org/10.1111/bjet.13033.
[22] K. K. Aldous, J. An, B. J. Jansen, Measuring 9 emotions of news posts from 8 news organizations across 4 social media platforms for 8 months, Trans. Soc. Comput. 4 (2022). URL: https://doi.org/10.1145/3516491. doi:10.1145/3516491.
[23] L. Plonsky, F. L. Oswald, How big is "big"? interpreting effect sizes in l2 research, Language learning 64 (2014) 878–912.
[24] R. Wei, Y. Hu, J. Xiong, Effect size reporting practices in applied linguistics research: A study of one major journal, Sage Open 9 (2019) 2158244019850035.
[25] R. Brotherton, Suspicious minds: Why we believe conspiracy theories, Bloomsbury Publishing, 2015.
[26] M. Barkun, A culture of conspiracy: Apocalyptic visions in contemporary America, volume 15, Univ of California Press, 2013.
[27] N. J. Russell, An introduction to the overton window of political possibilities, Mackinac Center for Public Policy 4 (2006).
[28] E. Durkheim, Suicide: A study in sociology, Routledge, 2005.
[29] R. Girard, Il capro espiatorio, Adelphi Edizioni spa, 2020.
[30] R. Hofstadter, The paranoid style in American politics, Vintage, 2012.
[31] E. Goffman, Frame analysis: An essay on the organization of experience., Harvard University Press, 1974.
[32] Z. Bauman, Retrotopia, Revista Española de Investigaciones Sociológicas (REIS) 163 (2018) 155–158.

Appendix

YouTube Videos Used in Dataset Collection

- **Chinese migrants are fastest growing group crossing into U.S. from Mexico**
  youtube.com/watch?v=M7TNP2OTY2g
- **Native American Shuts Down Immigration Protest**
  youtube.com/watch?v=2utsjsWOWUA
- **Migrants evade Texas floating barrier**
  youtube.com/watch?v=2i8n6jCH1S4
- **Denmark Is Leading Europe's Anti-Immigration Policies**
  youtube.com/watch?v=zpkBKEPxze4
- **This Immigrant Left the U.S. To Seek Asylum In Canada And Regrets It**
  youtube.com/watch?v=ONjCMzB_FPw
- **Venezuelan Immigrant: 'I Regret Having Come to the United States'**
  youtube.com/watch?v=3FPbZcVLTBI
- **Migrant group attempts mass entry into US at Mexico border**
  youtube.com/watch?v=h_TqO9EqMhY
- **Norway's Muslim immigrants attend classes on western attitudes to women**
  youtube.com/watch?v=oKY600o3CXw
- **Why does Sweden no longer wants immigrants?**
  youtube.com/watch?v=5CSUimZjiI0
- **How Sweden is Destroyed by the Immigration Crisis**
  youtube.com/watch?v=rUw4cs2MHwc
- **Migrant crisis reaches boiling point on Staten Island**
  youtube.com/watch?v=-LDra78ksTo
- **"Deportation, not relocation!" Poland votes on illegal migration**
  youtube.com/watch?v=x4afwGepMkM
- **Students Say Obama Immigration Quote Is Racist... When They Think It's From Trump**
  youtube.com/watch?v=Vj9IxVlLRl0
- **US' illegal immigrants crisis: Elon Musk visits Texas**
  youtube.com/watch?v=2_iYuiHyzKQ
- **Migrant beats resident, steals flag from NY home**
  youtube.com/watch?v=FTXZmor6KBY