Recurring Techniques, Shifting Meanings: A Longitudinal Study of Divisive Rhetorical Techniques in Immigration-Related YouTube Comments

Anonymous ACL submission

Abstract

Divisive rhetorical techniques are a pervasive feature of contemporary online political discourse, yet computational work has primarily studied them in institutional settings, leaving peer-to-peer social network interactions comparatively understudied. We introduce a span-annotated dataset of 3,275 immigration-related YouTube comments labeled with 24 divisive rhetorical techniques and span-aware detection models for peer-to-peer discourse. Deploying these resources in a longitudinal study covering a full U.S. presidential cycle, we find that divisive rhetoric constitutes a largely shared and stable repertoire: most techniques are deployed at comparable rates regardless of channel ideology, with temporal fluctuations tracking broader cycles of political salience. Ideological differentiation emerges instead at the content level. A subset of argumentatively configured techniques exhibits episodic shifts in referential content tied to salient geopolitical events, with left-leaning discourse more readily integrating transnational frames and right-leaning discourse remaining anchored to domestic narratives. These findings suggest that online discussions are sustained less through divergent rhetorical repertoires than by strategic content selection [1].

1 Introduction

Rhetorical devices such as slogans, name calling and appeals to fear, authority or hypocrisy are a pervasive feature of contemporary online political discourse. These practices have traditionally been studied in institutional political communication, where they function as persuasive tools deployed by political elites and media actors to influence public opinion (Zeitzoff, 2023). However, they are equally present in peer-to-peer social network interactions, where they function as issue-positioning mechanisms through which users manage social identity and stake out group boundaries (Bassi et al., 2025a).

Beyond persuasion, these techniques shape the structure of public debate itself. When deployed in public debate, these techniques tend to simplify the complexity of contested issues, delegitimize opposing viewpoints, and systematically constrain interlocutors' intellectual autonomy, ultimately pushing disagreement toward destructive rather than cooperative exchange (Deutsch, 1973). Zompetti (2015) captures this function by referring to them as *divisive rhetorical techniques* (DRT) - a label we adopt here to foreground their role in structuring antagonistic interaction rather than their persuasive intent per se (Godber and Origgi, 2023).

Social networks (SNs) further amplify these dynamics by disproportionately rewarding highly expressive and status-seeking actors, thereby distorting users' perceptions of the broader political landscape and rendering moderate voices comparatively less visible (Bail, 2022). This makes such platforms critical sites for studying how DRT are deployed, distributed across ideological lines, and semantically adapted over time. Understanding these patterns is directly relevant to research on misinformation, radicalization, and hate speech (Krykoniuk et al., 2025): divisive rhetoric does not simply reflect underlying political polarization but actively constitutes it, by normalizing adversarial frames and narrowing the space available for deliberative exchange (Schumann and Oswald, 2024; Moghaddam et al., 2007).

Yet, the computational study of DRT has developed almost exclusively in institutional settings (e.g. news articles (Piskorski et al., 2023b), structured political debates (Goffredo et al., 2023) and advertisement (Meguellati et al., 2026)) offering limited insight into how these techniques operate

[1] Scripts and data used in this paper can be found at this anonymous repository: https://anonymous.4open.science/r/Tech_Comment_Detection-0763. They will be made publicly available upon acceptance.

in the fragmented, informal, and interaction-driven nature of SN peer-to-peer discourse (Blitvich et al., 2013). Moreover, existing work has focused on detection as an end goal, with little attention to how these techniques are distributed across political contexts and how their semantic realization evolves over time: these questions are essential to understand the dynamics through which online discussions become divisive.

To fill this gap, we introduce a dataset and span-aware detection models for DRT in large-scale, naturalistic SN interactions, and deploy them in a longitudinal study of how such techniques are used and semantically adapted over time. We focus on immigration-related discussions in YouTube comments - a domain where divisive rhetoric is particularly pervasive, intersecting with multiple contested issues (national identity, security, economic redistribution, foreign policy) and carrying tangible real-world implications for marginalized groups (Bassi et al., 2025a).

Specifically, we contribute with:

(*i*) A large-scale annotated dataset of 3,275 YouTube comments on immigration-related videos, labeled with 24 DRT following established annotation guidelines (Piskorski et al., 2023b, 2025).

(*ii*) Span-aware computational models for detecting such techniques in YouTube comments, including strong baselines at both the comment and span level, leveraging a hybrid taxonomy that consolidates overlapping techniques and mitigates annotation sparsity.

(*iii*) A longitudinal case study of immigration-related YouTube discussions examining how DRT are distributed across political leanings and how their semantic realization evolves over time.

2 Related Work

The computational study of rhetorical and persuasion techniques in political discourse has a well-established tradition in NLP, with foundational work focusing on propaganda and persuasion detection in news media (Da San Martino et al., 2019; Piskorski et al., 2023b,a, 2025) and fine-grained technique classification in structured political debates (Goffredo et al., 2023) and political advertising (Meguellati et al., 2026). These settings offer longer texts, clearer argumentative structure, and more “stable” discourse conventions, conditions that differ substantially from the fragmented, informal, and interaction-driven nature of SN peer-to-peer discourse (Bassi et al., 2024; Blitvich et al., 2013).

More recent work has extended fine-grained rhetorical analysis beyond institutional settings, targeting social media platforms including Arabic tweets (Alam et al., 2022), multilingual Facebook memes (Dimitrov et al., 2021, 2024), Italian X posts (Ramponi et al., 2025), and multilingual political tweets (Moral et al., 2024). While these contributions demonstrate the feasibility of fine-grained technique detection in SNs, they treat detection as an end goal and largely focus on short, broadcast-oriented content (memes and tweets as isolated posts) rather than the sustained, conversational peer-to-peer interactions characteristic of comment threads (Blitvich et al., 2013). In this regard, YouTube has evolved into a major site of peer-to-peer political discussion (Andersson, 2024; Blitvich, 2010), yet remains comparatively under-explored in this literature (Röchert et al., 2020). Addressing this gap, Bassi et al. (2025b) proposes a framework for fine-grained rhetorical analysis of YouTube comments, modeling constructive and destructive dialogical moves through a set of DRT, while Bassi et al. (2025a) examine how DRT relate to stance, topic controversiality, and identity management in YouTube controversial discussions.

In this work, we extend this line of research by treating DRT not only as detection targets but also as analytical tools for studying the dynamics of online political discourse. In particular, we investigate how their use varies across political contexts and over time, as well as how the semantic content used to realize them evolves. We ask:

RQ1 How does the use of DRT vary over time across different political leanings?

RQ2 Does the use of techniques vary depending on the political leaning of the channel where comments are posted?

RQ3 How does the semantic content used to realize DRT evolves over time?

RQ4 How does the semantic realization of DRT differ across channel political leanings over time?

3 Experimental Setting

3.1 Dataset

We built our dataset using Social Blade[2] to identify the top 25 U.S.-based YouTube channels by subscriber count within the "News and Politics" category. For each channel, we assigned a binary

[2] https://socialblade.com/

Source & Leaning	Videos	Comments
Content – Left	705	889,041
Content – Right	414	501,470
News – Left	2,693	1,411,869
News – Right	2,360	1,767,295
Total	**6,172**	**4,569,675**

Table 1: Dataset by source type and political leaning.

political orientation label (Left or Right) using AllSides Media Bias[3] and Media Bias Fact Check[4] and manually categorized channels as either "Legacy Media" or independent "Content Creators." Data collection spanned from 2020 to the end of 2024 to capture an entire U.S. presidential cycle, during which immigration-related debates were particularly salient. To avoid the recency bias and reproducibility issues documented for the YouTube API search endpoint (Rieder et al., 2025; Efstratiou, 2025), and to overcome the 20,000-video cap of the UploadPlaylist method (insufficient for channels exceeding 100,000 uploads), we relied on the yt-dlp library for comprehensive channel crawling (see Appendix 6 for details).

We collected all videos published between January 2020 and December 2024, excluding channels that disabled comments. Videos were filtered for immigration relevance by applying a binary classifier to titles and descriptions using *gpt-4.1-nano*[5], validated on a stratified sample of 100 videos (Accuracy: 96%). Comments associated with immigration-related videos were then retrieved via the YouTube Data API. Table 1 summarizes the final dataset of 6,172 videos and 4,569,675 comments; channel-level statistics are reported in the Appendix 6. Figure 1 shows the temporal distribution of comments disaggregated by source and channel leaning.

3.2 Dataset Annotation

Our annotation framework follows the fine-grained taxonomy of Piskorski et al. (2025), operationalized here via a coarser "hybrid" taxonomy obtained through deterministic post-hoc mapping.[6] This consolidation merges conceptually overlapping techniques with ambiguous decision boundaries (i.e. Doubt & Smears; Appeal to Values & Flag Waving) and collapses low-support techniques (i.e. Red Herring, Strawman, Whataboutism) for which separate modeling would yield unstable estimates.

[3] https://www.allsides.com/media-bias/ratings

[4] https://mediabiasfactcheck.com/

[5] https://platform.openai.com/docs/models/gpt-4.1-nano

[6] We release annotations under both the original and hybrid taxonomies. Full guidelines can be found in the repository.

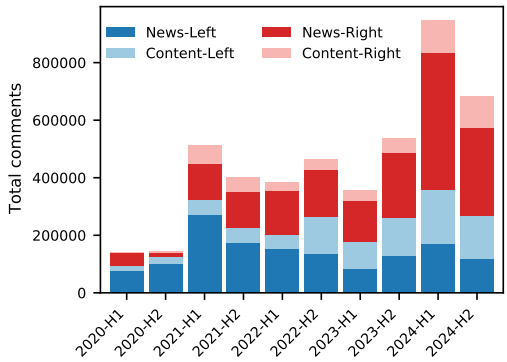


Figure 1: Comments Distribution per Semester

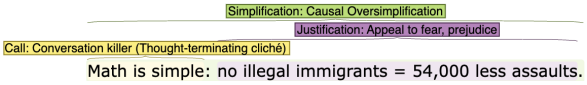


Figure 2: Annotation Example with Overlapping Spans

The annotation task required identifying persuasion techniques in anonymized YouTube comments and marking their text spans, formulated as multi-label span annotation with overlapping spans allowed (Figure 2). All annotations were conducted in INCEpTION (Klie et al., 2018). We hired from a private company two expert annotators with prior experience in persuasion technique annotation and at least C1-level English proficiency. The campaign followed a structured protocol: an initial calibration set of 100 jointly labeled comments was followed by independent annotation of distinct subsets, with quality checkpoints every 700 comments where both annotators labeled a shared subset of 85 comments; disagreements were resolved in consensus meetings to prevent annotator drift. This process yielded 525 double-annotated comments, split evenly into development and test sets.

Inter-annotator agreement was evaluated at two granularities: span-level γ for identification and classification (Mathet, 2017; Mathet et al., 2015), and comment-level Krippendorff's α (overall) and Cohen's κ (per technique). Overall scores were $\alpha = 0.57$, $\gamma = 0.50$ for span identification, and $\gamma_{\text{cat}} = 0.39$ for classification - consistent with prior work on fine-grained persuasion annotation (Ramponi et al., 2025). Agreement stability across sessions is shown in Appendix 6. Detailed per-technique instance counts and span length distributions are reported in the Appendix 6.

3.3 Techniques Detection: Comment Level

Single-Head Architecture We trained a baseline comment-level model using `DeBERTa-v3-Large` as encoder, with a multi-label classification head directly fine-tuned on full comment text, leveraging the strong performance of BERT-family architectures for persuasion techniques detection, especially on data from social networks (short texts extracted from memes) (Srba et al., 2026, p. 32). Training followed a two-phase strategy: first, domain adaptation via masked language modeling (MLM) on the full YouTube comment corpus; second, fine-tuning on the annotated dataset with hyperparameters optimized via Bayesian search with Optuna on the development set.

Two-Heads Architecture To improve comment-level performance, we adopted a multitask architecture that jointly models token-level and comment-level signals. Given an input comment, the shared encoder produces both contextualized token representations and a global `[CLS]` embedding. A token-level head, supervised via span annotations, computes a span-aware comment representation through weighted pooling. This is fused with the `[CLS]` embedding via a learnable gating mechanism (details in Appendix 6) and passed to a comment-level classification head. Training jointly optimizes token-level and comment-level objectives with adaptive loss rescaling.

3.4 Techniques Detection: Token Level

Token-level detection is formulated as multi-label sequence tagging, building on the same domain-adapted `DeBERTa-v3-Large` encoder. Gold spans are projected onto tokenized inputs, yielding binary labels per token–technique pair, and training optimizes binary cross-entropy with optional per-label positive class weighting. Inference thresholds are tuned per-technique on the development set; contiguous predicted token sequences are merged into spans (see Appendix 6). Comment-level predictions are derived by flagging a technique as present if at least one span is detected.

Single-Head Architecture A single linear projection over encoder hidden states jointly predicts all techniques at each token position using sigmoid activations.

Multi-Head Architecture Following Ramponi et al. (2025), we also experiment with parallel per-technique classification heads over the shared encoder, each producing independent token-level logits. This decouples technique predictions and allows independent threshold tuning per head.

Table 2 reports performance at comment and span level averaged over three runs. The multitask comment-level architecture yielded consistent improvements in macro-F1 (+0.105) and recall (+0.055) over the single-head baseline, indicating improved detection of less frequent techniques. The single-head token-level model achieved the highest overall macro-F1 and was selected for all subsequent analyses; per-technique scores are reported in the Appendix 6.

4 Results

RQ1 - How does the use of DRT vary over time across political leanings? To address RQ1, we count distinct techniques detected per comment (i.e., techniques with at least one predicted span), average these at the video level and aggregate by semester and channel political leaning to reduce distortions from unequal comment volumes. To rule out comment length as a confound, we compared average word counts across channels (Appendix 3); Left- and Right-leaning comments are similar in length (24.6 vs. 22.9 words), suggesting that differences in technique counts reflect rhetorical patterns rather than verbosity.

Figure 3 reports semester-level trends for Left- and Right-leaning channels. Both groups exhibit consistently high technique usage (Left: 2.43; Right: 2.65 on average) and follow similar temporal trajectories: increasing from 2021, sustained through 2022–2023, and declining in 2024. Right-leaning channels consistently exhibit higher technique counts, with a steeper increase between 2020 and 2021.

RQ2 - Does the relative use of specific DRT differ across channel political leanings over time? For each technique and semester, we compute the proportion of comments containing at least one instance in Left- and Right-leaning channels and derive a leaning gap (Right minus Left). Figure 4 reports techniques with the strongest asymmetries alongside the median and interquartile range over all remaining techniques.

For most techniques, the leaning gap remains close to zero, indicating comparable employment rates across channel groups. A small subset displays more pronounced asymmetries: *Slogans* consistently exhibit positive gaps (max $+0.095$, mean

Architecture	Comment-Level			Span-Level		
	Prec.	Rec.	Macro-F1	Prec.	Rec.	Macro-F1
Comm_Single-Head	**62.82±0.6**	64.53±1.1	38.45±0.3	–	–	–
Comm_2Heads	62.02±1.4	70.03±0.99	48.98±1.5	–	–	–
Token_Single-Head	45.56±1.84	**61.65±1.22**	**50.37±1.45**	32.92±0.34	**56.96±1.32**	**40.16±0.52**
Token_Multi-Head	46.95±0.9	55.96±2.27	49.12±0.97	**33.22±2.16**	52.08±1.97	38.64±1.38

Table 2: Comparison of architectures at comment and span level (mean±std over three runs).

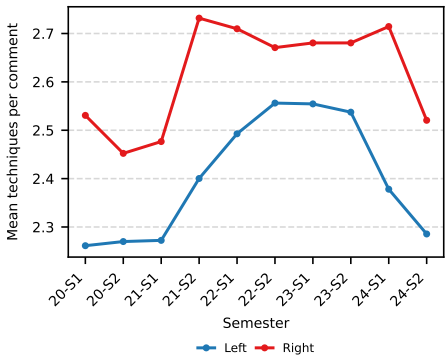


Figure 3: Mean number of distinct DRT per comment over time, aggregated at the video level and averaged by semester and channel political leaning.

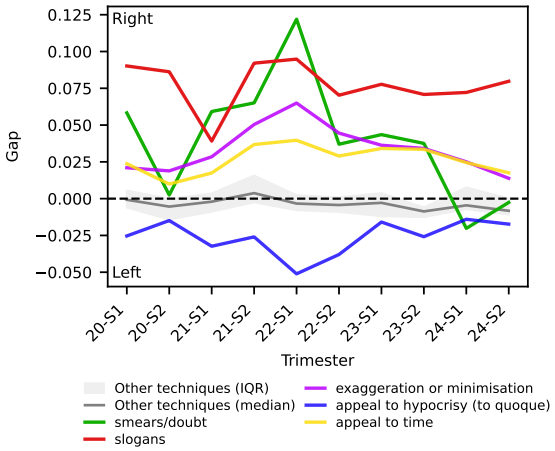


Figure 4: Comparison of technique-specific usage across political leaning channel groups.

+0.077), as do *Exaggeration or minimisation* (max +0.065, mean +0.034) and *Appeal to time* (max +0.040, mean +0.027), all reflecting higher relative usage in Right-leaning channels. *Appeal to hypocrisy (tu quoque)* shows consistently negative gaps (max −0.051, mean −0.026), indicating stable relative deployment in Left-leaning channels. *Smears/Doubt* exhibits the largest overall deviation (max gap +0.122, mean +0.040), but with a non-monotonic trajectory: near zero in late 2020, sharply positive during 2021–2022, and shifting negative in 2024.

RQ3 - How does the semantic content used to realize DRT evolve over time? For each predicted span, we extract its text and embed it with `all-MiniLM-L6-v2` (L2-normalized). Embeddings are aggregated into per-technique, per-semester centroids; semantic shift is defined as cosine distance between consecutive-semester centroids for the same technique. We tested whether observed shifts reflect generic topical drift by constructing a matched null baseline, sampling random spans from adjacent semesters without conditioning on technique labels and matching sample sizes to observed technique counts. We further conducted within-technique permutation tests, shuffling semester labels across spans 1,000 times and recomputing centroid distances.

Figure 5 reports techniques with average shift ≥ 0.01; techniques below this threshold show consistently low centroid displacement, indicating that semantic change is concentrated in a limited subset. For the strongest peaks, observed shift consistently exceeded the matched topical-drift baseline, and all peak-level permutation tests were significant ($p = .001$, empirical; see Repository for full results), supporting the interpretation that these shifts are technique-specific rather than a by-product of the broader news cycle.

To contextualize shift peaks, we complement centroid analysis with contrastive TF-IDF over span texts from adjacent semesters (stopword removal, lemmatized nouns/pronouns/adjectives via SpaCy), deriving per-term scores as the difference in mean TF-IDF weight between later and earlier semesters. Figure 6 reports the most pronounced shift peaks. The following examples are reproduced verbatim for analytical purposes and may contain offensive language.

Figure 6a illustrates a shift in *Appeal to pity* following the Taliban offensive in Afghanistan. Earlier realizations invoke generic references to vul-

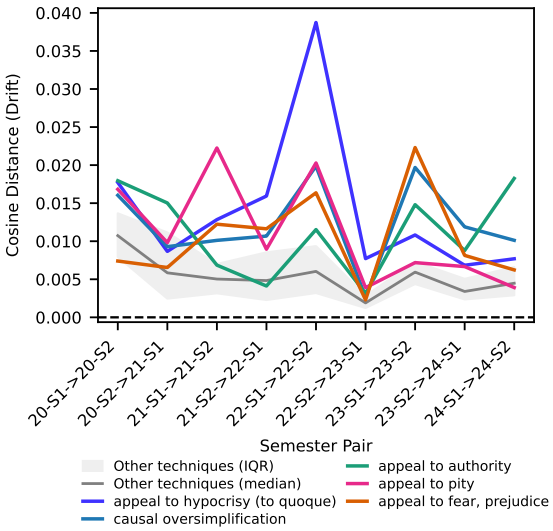


Figure 5: Average semantic shift of DRT across consecutive semesters, measured as cosine distance between technique-level span embedding centroids.

nerable individuals (e.g., *"we should welcome them, they are just providing for their starving families"*); in subsequent semesters, the Afghan humanitarian crisis is explicitly mobilized to argue for accepting refugees (e.g., *"these people are fleeing a war we helped create, turning them away is inhumane"*).

Figure 6b shows a comparable shift for *Appeal to hypocrisy*: earlier realizations invoke the war in Ukraine to expose inconsistencies in refugee policy (*"politicians claim to support refugees abroad, but suddenly borders matter when it comes to who we let in"*); later ones target domestic controversies around migrant bussing to sanctuary cities (*"the DC crowd opened the doors for them to enter UNTIL they appeared on their doorstep"*).

Figure 6c shows that *Appeal to fear* shifts from drug trafficking and cartel violence framing in early 2023 toward civilizational and religious stereotypes linked to the Israel/Palestine conflict in later semesters (e.g., *"Western people are too liberated and you cannot change their way of life, like the Gazans-extremely conservative, and you cannot change their mentality or else there's going to be jihad"*).

Overall, these examples show that DRT remain functionally stable while adapting their semantic content to salient geopolitical events, with foreign conflicts recurrently mobilized as reference frames for domestic immigration debates.

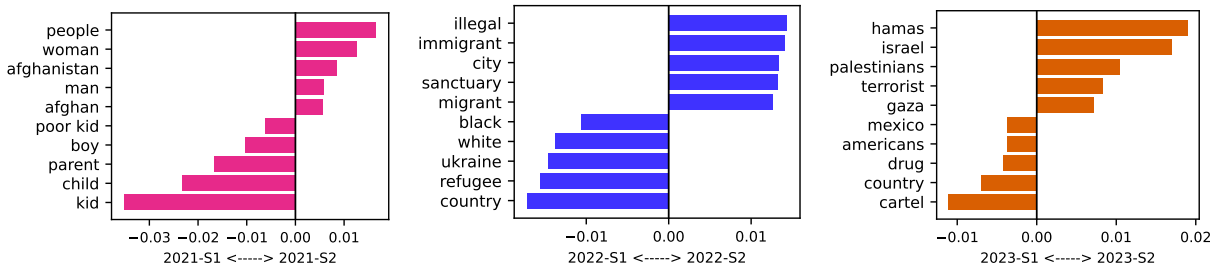
RQ4 - How does the semantic realization of DRT differ across channel political leanings over time? Following the procedure of RQ3, we compute separate per-technique, per-semester centroids for Left- and Right-leaning channels, and quantify cross-leaning semantic divergence as the cosine distance between them. To test whether these distances reflect technique-specific divergence rather than general Left–Right language differences, we computed a matched null baseline using random spans from Left- and Right-leaning channels in the same semester that were not labeled with the focal technique, again matching the original sample sizes. We also ran permutation tests by shuffling leaning labels across spans within techniques 1,000 times.

Figure 7 reports the resulting trajectories, with the largest distances clustering in early 2022 and late 2023 and divergence outside these periods more attenuated and concentrated in a small subset of techniques. Peak divergences remained above the matched cross-leaning baseline, and all peak-level permutation tests were significant ($p = .001$, empirical; see Repository for full results), indicating that the largest observed gaps are not reducible to background channel-level linguistic differences.

Figure 8a focuses on *Appeal to Values / Flag Waving* in early 2022, the largest observed semantic distance. Right-leaning channels instantiate it through nationalistic and sovereignty-related references (*"ILLEGAL IMMIGRATION IS NOT AN AMERICAN CONSTITUTIONAL RIGHT"*), while Left-leaning channels ground it in geopolitical events - here the invasion of Ukraine (*"I wouldn't fight for a country that is discriminating against me either. That doesn't make me a coward!"*).

Figure 8b shows that in the same semester, *Smears/Doubt* is realized through references to foreign political actors in Left-leaning channels (*"You can't believe the Putin media"*), and toward domestic political figures in Right-leaning ones (*"How much cartel money is Hunter getting in art sales?"*).

Figure 8c shows that in the second semester of 2023, *Causal Oversimplification* diverges along similar lines: Left-leaning channels invoke the Israel/Palestine conflict to explain migration dynamics (*"If you keep destabilizing entire regions, like Gaza, people will keep fleeing"*), while Right-leaning channels anchor the same technique to domestic policy and administration-level responsibility (*"More fentanyl has been confiscated at the border after Trump"*).



(a) Appeal to Pity (b) Appeal to Hypocrisy (c) Appeal to Fear

Figure 6: Contrastive TF–IDF top terms (later minus earlier semester) illustrating semantic reconfiguration across time.

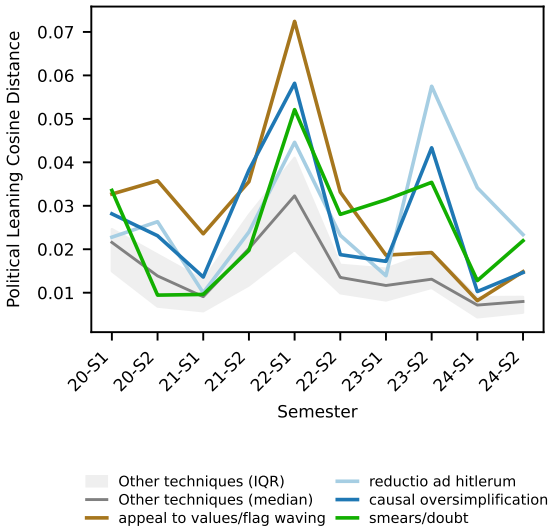


Figure 7: Semantic divergence of DRT across political leanings over time (cosine distance of span embeddings).

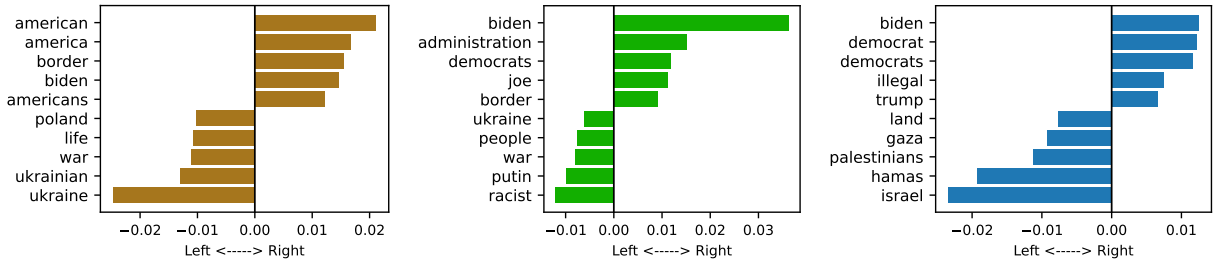
5 Discussion

Our results converge on a consistent picture: DRT constitute a largely shared, stable repertoire across time and political leanings, with systematic variation emerging at the level of intensity and semantic instantiation.

RQ1 shows that immigration consistently elicits high levels of DRT across the full observation period, confirming its status as a structurally contentious issue rather than an episodic or event-driven controversy (Nasuto and Rowe, 2024). The synchronized temporal dynamics across both channel groups - increasing from 2021, sustained through 2022–2023, declining in 2024 - suggest that rhetorical intensity tracks broader cycles of political salience, underscoring the effectiveness of such communicative devices as indicators of topical controversiality. The consistently higher levels observed in Right-leaning channels align with political science findings linking immigration salience to active elite contestation (Jin, 2025), reflecting the central role of immigration in right-wing political discourse where it is more frequently framed as a contested and problematized issue.

RQ2 confirms that most techniques are deployed at comparable rates across channel groups, consistent with prior characterizations of divisive rhetoric as a shared toolkit (Bassi et al., 2025a; Zompetti, 2015). The technique-specific asymmetries that do emerge are theoretically interpretable: the persistent right-leaning excess in *Slogans* and *Appeal to Time* is consistent with compressed, mobilizing formulations associated with right-wing populist communication style (De Vreese et al., 2018), while the left-leaning excess in *Appeal to Hypocrisy* aligns with the investment in consistency and liberal international order characteristic of left-wing discourse (Millar, 2024). The shifting alignment of *Smears/Doubt* - predominantly right-leaning during Democratic administrations and reversing in 2024 - further illustrates how oppositional discourse dynamics modulate technique prevalence independently of stable ideological preferences.

Drawing on Bassi et al. (2024)'s distinction between the argumentative *how* and the content *what* of rhetorical strategies, RQ3 and RQ4 extend the analysis to content variation in technique instantiation. RQ3 shows that most techniques display limited content displacement over time - consistent with techniques such as *Slogans*, *Name Calling*, and *Loaded Language* deriving their pragmatic force from stable, sedimented lexicon. A smaller subset - particularly techniques whose efficacy relies on argumentative configuration (*Appeal to Values*, *Appeal to Hypocrisy*, *Causal Oversimplification*) - exhibits episodic content shifts aligned



(a) Appeal to Values 2022-S1 (b) Smears/Doubt 2022-S1 (c) Causal Oversimplification

Figure 8: Contrastive TF–IDF top terms (Right minus Left) illustrating semantic reconfiguration across political leanings.

with salient geopolitical events. These techniques function through recontextualization mechanisms: users renew their perlocutionary effect by adapting referential content while preserving argumentative structure, enabling responsiveness to changing political environments without expanding the underlying rhetorical repertoire.

RQ4 shows that cross-leaning content divergence is similarly concentrated in this same subset of argumentatively configured techniques, and in specific time periods. Where divergence occurs, it reflects asymmetric patterns of semantic integration: Left-leaning discourse more readily incorporates transnational, humanitarian, and geopolitical reference frames, while Right-leaning discourse anchors the same techniques to domestically bounded narratives centered on national governance and internal political accountability. Ideological positioning is thus sustained not through divergent rhetorical repertoires, but through selective framing strategies that regulate which external events are deemed relevant for immigration debates.

6 Conclusions

This paper investigated how DRT are deployed and adapted in YouTube immigration-related discussions over a full U.S. presidential cycle. By combining span-level annotation, supervised detection models and longitudinal semantic analysis, we provide a fine-grained account of how issue positioning is communicatively constructed and transformed over time.

Our findings show that DRT constitutes a largely shared and stable repertoire: techniques tend to be used at comparable rates across political leanings and to persist over time. Differences emerge primarily at the level of intensity, with higher overall usage in right-leaning channels and fluctuations tracking broader cycles of political salience, confirming immigration as a structurally contentious issue. While most techniques display limited content change, a subset relying on argumentative configuration exhibits episodic shifts in response to geopolitical events. Specifically, Left-leaning channels more readily integrate transnational frames while Right-leaning ones remaining more anchored to domestic issues.

Methodologically, we release a dataset of YouTube comments span-annotated with DRT and the trained detection model, designed to support computational research on contentious discourse across platforms, topics, and political contexts. Taken together, our work provides tools for studying how lay users navigate disagreement online, offering empirical leverage to examine how differences are articulated, stabilized, and transformed in large-scale public discourse - and supporting more informed inquiry into the conditions under which online debates become persistently polarized or remain open to contestation.

Limitations

Data Our analysis focuses on a single issue, i.e. immigration, and on U.S.-based YouTube channels with enabled comments. While immigration is a paradigmatic controversial topic, this scope limits its generalizability across issues, countries, and platforms. The dataset may also reflect platform-specific participation and moderation biases. Future work should extend this approach to other topics and socio-political contexts and compare persistent versus episodic commenters to assess how rhetorical practices stabilize or diffuse over time.

Methods We do not model users' stance toward immigration, preventing direct links between rhetorical technique usage and issue positioning; future work should integrate stance detection to address this limitation. In addition, our analysis does

not account for the content of the videos to which comments respond. Incorporating video narratives and rhetoric would enable investigation of which content features elicit DRT and facilitate the study of rhetorical contagion.

Interpretation Our findings are descriptive and do not support causal claims. Future work should combine longitudinal observational data with causal or experimental designs to assess the relations between divisive rhetorical techniques on issue positioning and deliberative quality.

Ethical Considerations

Data Collection and Privacy All data used in this study consists of publicly available YouTube comments. No private or protected user information was collected. All comments were anonymized prior to annotation and analysis by removing usernames and any other identifying information, ensuring that individual users cannot be identified from the data or examples reported in this paper. The dataset will be released in anonymized form only.

Annotation and Annotator Welfare The annotation task required expert annotators to engage with large volumes of potentially offensive, inflammatory, and politically charged content. To mitigate the risk of annotator fatigue and distress, annotation sessions were structured with regular breaks and consensus meetings, which also served as a space for annotators to flag distressing content and discuss difficult cases collaboratively. Annotators were fully informed of the nature of the content prior to participation and were compensated fairly for their work.

Verbatim Examples This paper reproduces a small number of YouTube comments verbatim for analytical purposes. These examples were selected solely to illustrate specific rhetorical patterns and may contain offensive, inflammatory, or discriminatory language. We include them because their precise wording is analytically relevant - paraphrasing would obscure the rhetorical features under study - and we flag their presence explicitly in the main text. We do not endorse any of the views expressed in these examples.

Dual-Use Risks The models and dataset introduced in this paper are designed to support the study of DRT in online discourse, with potential applications in computational social science, platform governance, and media literacy research. We acknowledge that, as with any detection system for political speech, these resources carry dual-use risks: they could in principle be repurposed to identify and suppress legitimate political expression, or to optimize the production of divisive rhetoric rather than to detect it. We mitigate this risk by releasing the dataset and models for research purposes, with documentation that contextualizes their intended use and known limitations. We encourage users to critically assess the suitability of these tools for any downstream application, particularly in high-stakes or politically sensitive contexts.

Acknowledgments

ChatGPT-5.2-Codex was used for debugging assistance during the development of data processing and modeling pipelines. Claude Sonnet-4.6 was used to assist with proofreading and language polishing of the manuscript. In both cases, all outputs were carefully reviewed, verified, and revised by the authors, who take full responsibility for the correctness of the writings presented in this paper.

References

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, and Preslav Nakov. 2022. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marta Andersson. 2024. E-mpoliteness–creative impoliteness as an expression of digital social capital. *Journal of Politeness Research*, 20(2):227–248.

Chris Bail. 2022. Breaking the social media prism: How to make our platforms less polarizing. In *Breaking the Social Media Prism*. Princeton University Press.

Davide Bassi, Giovanni Da San Martino, Renata Vieira, and Martín Pereira-Fariña. 2025a. Drawing digital lines: pattern analysis of divisive rhetoric in social network discussions. *Nature - Humanities and Social Sciences Communications*, 12(1).

Davide Bassi, Søren Fomsgaard, and Martín Pereira-Fariña. 2024. Decoding persuasion: a survey on ml and nlp methods for the study of online persuasion. *Frontiers in Communication*, 9:1457433.

Davide Bassi, Erik Bran Marino, Renata Vieira, and Martin Pereira. 2025b. Old but gold: LLM-based features and shallow learning methods for fine-grained controversy analysis in YouTube comments. In *Proceedings of the 12th Argument mining Workshop*, pages 46–57, Vienna, Austria. Association for Computational Linguistics.

Pilar Garcés-Conejos Blitvich. 2010. The youtubification of politics, impoliteness and polarization. In *Handbook of research on discourse behavior and digital communication: Language structures and social interaction*, pages 540–563. IGI Global.

Pilar Garcés-Conejos Blitvich, Nuria Lorenzo-Dus, Patricia Bou-Franch, Istvan Kecskes, and Jesús Romero-Trillo. 2013. Relational work in anonymous, asynchronous communication: A study of (dis) affiliation in youtube. *Research trends in intercultural pragmatics*, pages 343–366.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Claes H De Vreese, Frank Esser, Toril Aalberg, Carsten Reinemann, and James Stanyer. 2018. Populism as an expression of political communication content and style: A new perspective. *The international journal of press/politics*, 23(4):423–438.

Morton Deutsch. 1973. *The Resolution of Conflict: Constructive and Destructive Processes*. Yale University Press, New Haven.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, Mexico City, Mexico. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Alexandros Efstratiou. 2025. On youtube search api use in research. In *Proceedings of the 2025 ACM Internet Measurement Conference*, IMC '25, page 919–927, New York, NY, USA. Association for Computing Machinery.

Amelia Godber and Gloria Origgi. 2023. Telling propaganda from legitimate political persuasion. *Episteme*, 20(3):778–797.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.

Soyeon Jin. 2025. Why is immigration important to you? a revisit to public issue salience and elite cues. *European Journal of Political Research*, 64(2):671–696.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Kateryna Krykoniuk, Cleo Hopkin-King, and Seán G Roberts. 2025. Developing a discourse space for analysing online discourse. *Discourse, Context & Media*, 67:100929.

Yann Mathet. 2017. The agreement measure γ cat a complement to γ focused on categorization of a continuum. *Computational Linguistics*, 43(3):661–681.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Elyas Meguellati, Stefano Civelli, Pietro Bernardelle, Shazia Sadiq, Irwin King, and Gianluca Demartini. 2026. Towards detecting persuasion on social media: From model development to insights on persuasion strategies. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Katharine M Millar. 2024. Limitations of hypocrisy as a strategy of critique in international politics. *International Theory*, 16(3):295–320.

Fathali M Moghaddam, Rom Harré, and Naomi Lee. 2007. *Global conflict resolution through positioning analysis*. Springer Science & Business Media.

Pablo Moral, Jesús M Fraile, Guillermo Marco, Anselmo Peñas, and Julio Gonzalo. 2024. Overview of dipromats 2024: Detection, characterization and tracking of propaganda in messages from diplomats and authorities of world powers. *Procesamiento del lenguaje natural*, 73:347–358.

Andrea Nasuto and Francisco Rowe. 2024. Understanding anti-immigration sentiment spreading on twitter. *PloS one*, 19(9):e0307917.

Jakub Piskorski, Dimitar Dimitrov, Filip Dobranić, Marina Ernst, Jacek Haneczok, Ivan Koychev, Nikola Ljubešić, Michal Marcinczuk, Arkadiusz Modzelewski, Ivo Moravski, and Roman Yangarber. 2025. SlavicNLP 2025 shared task: Detection and classification of persuasion techniques in parliamentary debates and social media. In *Proceedings of*

the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 254–275, Vienna, Austria. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023a. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023b. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Alan Ramponi, Agnese Daffara, and Sara Tonelli. 2025. Fine-grained fallacy detection with human label variation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 762–784, Albuquerque, New Mexico. Association for Computational Linguistics.

Bernhard Rieder, Adrián Padilla, and Òscar Coromina. 2025. Forgetful by design? a critical audit of youtube's search api for academic research. *Information, Communication & Society*, pages 1–20.

Daniel Röchert, German Neubaum, Björn Ross, Florian Brachten, and Stefan Stieglitz. 2020. Opinion-based homogeneity on youtube: Combining sentiment and social network analysis. *Computational Communication Research*, 2(1):81–108.

Jennifer Schumann and Steve Oswald. 2024. Pragmatic perspectives on disagreement. *Journal of Language Aggression and Conflict*, 12(1):1–16.

Ivan Srba, Olesya Razuvayevskaya, João A. Leite, Robert Moro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, Carolina Scarton, Kalina Bontcheva, and Maria Bielikova. 2026. A survey on automatic credibility assessment using textual credibility signals in the era of large language models. *ACM Trans. Intell. Syst. Technol.*, 17(2).

Thomas Zeitzoff. 2023. *Nasty Politics: The Logic of Insults, Threats, and Incitement*. Oxford University Press.

Joseph Zompetti. 2015. *Divisive discourse*. Cognella Academic Publishing, Illinois State University.

Appendix

YouTube Data Collection: API Limitations

Prior work documents substantial recency bias and poor reproducibility in the API's search endpoint (Rieder et al., 2025; Efstratiou, 2025). Accordingly, we avoided keyword-based search for video retrieval. Furthermore, although the YouTube API provides access to channel uploads via the Upload-Playlist method, this endpoint is capped at 20,000 videos per channel. Because several channels exceeded 100,000 uploads, this approach would have resulted in incomplete coverage and systematic data loss. To overcome these structural limitations, we relied on the yt-dlp library[7], which enables comprehensive and deterministic crawling of all publicly available videos from a given channel.

Inter-Annotator Agreement over Time

Figure 9 shows that agreement remained stable across annotation sessions, indicating the effectiveness of our quality-control procedure.

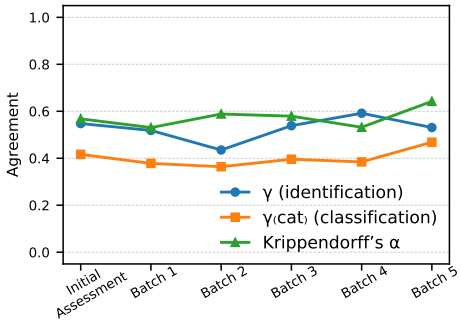


Figure 9: Inter-annotator agreement over time across annotation sessions.

Dataset Statistics

Detailed channel-level statistics, per-technique instance counts, and span length distributions are reported in Table 3 and Table 4 below.

Model Architecture Details

Two-Heads Comment-Level Architecture: Gating Mechanism The span-aware representation is combined with the global [CLS] embedding through a learnable gating mechanism:

$$g = \sigma(W_g h_{\text{CLS}}), \qquad h = g \cdot h_{\text{span}} + (1-g) \cdot h_{\text{CLS}},$$

where g controls the relative contribution of span-level and global contextual evidence. The fused

[7] https://github.com/yt-dlp/yt-dlp

Channel	Videos	Comments	Avg. len.
Left-Leaning Channels			
Legacy Media			
ABC News	1.078	627.235	25.5
CBS News	1.405	477.240	23.5
VICE News	116	210.447	27.9
Inside Edition	94	96.437	15.4
Content Creator			
The Young Turks	538	660.721	27.1
David Pakman Show	148	149.604	26.5
Brian Tyler Cohen	19	77.496	23.8
Left Sub-Total	**3.398**	**2.300.910**	**24.6**
Right-Leaning Channels			
Legacy Media			
Fox Business	2.044	1.422.103	21.4
LiveNOW from FOX	316	345.190	22.4
Content Creator			
Ben Shapiro	161	237.806	26.8
Megyn Kelly	172	142.709	27.4
Matt Walsh	51	86.761	29.8
Charlie Kirk	10	13.092	27.3
DailyWire+	20	20.011	25.6
Right Sub-Total	**2.774**	**2.268.765**	**22.9**
Total	**6.172**	**4.569.675**	**23.8**

Table 3: Final dataset composition with average comment length (number of words) per channel.

representation h is passed to a comment-level classification head that predicts the presence of rhetorical techniques in a multi-label setting.

Token-Level Inference Procedure At inference time, token-level logits for each rhetorical technique are converted into binary predictions using a fixed threshold tuned on the development set. Contiguous sequences of positively predicted tokens for the same technique are merged into predicted spans. To enable comparison with comment-level baselines, comment-level predictions are derived from token-level outputs: a rhetorical technique is predicted as present in a comment if at least one corresponding span is detected within that comment.

Best Training Parameters Below the best parameter for training the token-level single-head model:

Model	DeBERTa-v3-large
Hardware	2× Nvidia Tesla V100S
Optimization	Optuna
Learning Rate	2.6049078843086554e-05
Batch Size	8
Epochs	28 (with early stopping)
Gradient Accumulation Steps	1
Max Length	320
Dropout Probability	0.2
Weight Decay	0.03
Inference Threshold	0.35

Technique	N	Avg.	St.Dev.
Appeal to authority	179	93.77	±121.07
Appeal to pity	155	108.13	±67.01
Appeal to popularity	78	58.29	±40.09
Appeal to hypocrisy	228	140.42	±88.17
Appeal to time	47	50.66	±26.59
Appeal to values/Flag waving	562	113.96	±86.85
Appeal to fear, prejudice	540	129.96	±95.23
Black-and-white fallacy	336	92.45	±59.11
Causal oversimplification	378	147.90	±86.75
Distraction	73	133.47	±122.53
Exaggeration/minimisation	366	72.39	±52.73
Intentional vagueness	34	81.41	±85.12
Loaded language	2150	16.71	±14.67
Name calling	3160	16.84	±14.71
Reductio ad Hitlerum	48	118.83	±81.17
Repetition	1067	10.37	±8.75
Slogans	190	37.97	±31.81
Smears/Doubt	1549	116.42	±75.37
Thought-terminating cliché	466	30.67	±21.97

Table 4: Span statistics by persuasion technique

Per-Technique Performance Technique-specific performance scores are reported in Table 5.

	Comment-level			Span-level		
Technique	Prec.	Recall	F1	Prec.	Recall	F1
Appeal to authority	46.67	63.64	53.85	50.00	54.55	52.17
Appeal to fear, prejudice	52.24	72.92	60.87	29.91	64.00	40.76
Appeal to hypocrisy	53.85	38.89	45.16	40.00	21.05	27.59
Appeal to pity	50.00	41.67	45.45	37.50	50.00	42.86
Appeal to popularity	36.36	57.14	44.44	30.77	57.14	40.00
Appeal to time	40.00	100.00	57.14	22.22	100.00	36.36
Appeal to values / flag waving	50.91	70.00	58.95	30.43	63.64	41.18
Black-and-white fallacy	61.90	46.43	53.06	29.33	73.33	41.90
Causal oversimplification	72.73	48.48	58.18	36.67	64.71	46.81
Distraction	28.57	50.00	36.36	10.00	25.00	14.29
Exaggeration / minimisation	25.00	52.00	33.77	18.37	36.00	24.32
Intentional vagueness	0.00	0.00	0.00	0.00	0.00	0.00
Loaded language	64.46	91.45	75.62	47.19	77.30	58.60
Name calling	83.73	87.42	85.54	62.80	78.33	69.71
Reductio ad hitlerum	40.00	66.67	50.00	25.00	33.33	28.57
Repetition	29.79	77.78	43.08	23.42	56.52	33.12
Slogans	44.00	68.75	53.66	31.58	63.16	42.11
Smears / doubt	76.64	68.91	72.57	50.89	66.67	57.72
Thought-terminating cliché	52.73	70.73	60.42	41.89	65.96	51.24

Table 5: Technique-specific precision (P), recall (R), and F_1 scores (in %) of the best-performing model, evaluated at both comment and span level.

Robustness Checks for Semantic Shift and Cross-Leaning Divergence