Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-06277-7 OPEN

Drawing digital lines: pattern analysis of divisive rhetoric in social network discussions

Davide Bassi[1]✉, Giovanni Da San Martino[2], Renata Vieira[3] & Martín Pereira-Fariña[4]

Social dialogue is a cornerstone for political decision-making and maintaining mutual understanding between diverse societal groups when addressing collective challenges. However, this dialogue is increasingly strained in digital environments where users regularly encounter opposing viewpoints. While research has examined how political actors strategically leverage divisive rhetoric, less attention has been paid to how ordinary users utilize these devices in everyday online interactions. This study investigates how users employ divisive rhetorical strategies across social networks, examining the relationships between topic controversiality, user stance, and interactive patterns. Through a large-scale analysis of 146K YouTube comments on immigration and climate change discussions—two highly polarizing topics in contemporary discourse. The research combines computational methods for rhetoric mining with network analysis to track patterns of user interaction and manifestation of divisive rhetoric. Our analysis reveals three key findings: (1) Controversial topics elicit significantly higher frequencies of divisive rhetorical strategies compared to non-controversial ones, with distinct patterns across topics; (2) Users demonstrating strong stance commitment (Pro and Contra) use significantly more divisive rhetoric with parallel patterns, regardless of ideological position, distinguishing them from neutral users; (3) Users strategically adapt their rhetorical behavior to their interlocutor's stance, suggesting that stance intensity rather than specific ideological content drives rhetorical similarity. Framed through Social Identity Theory, we conceptualize these wedge rhetorical devices as an interactive toolkit that users deploy to navigate social positioning in deindividualized discussions, either reinforcing solidarity among users sharing similar positions or creating distinctions from those holding opposing views. This study shows how computational methods can effectively track and analyze the ways citizens strategically navigate social positioning on sensitive issues, contributing to our understanding of online political discourse dynamics.

[1] Universidade de Santiago de Compostela, Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Santiago de Compostela, Spain. [2] University of Padua, Padova, Italy. [3] University of Évora, Evora, Portugal. [4] Universidade de Santiago de Compostela, Instituto de Investigación en Humanidades (iHUS), Santiago de Compostela, Spain. ✉email: davide.bassi@usc.es

Introduction

Social dialogue is a cornerstone of political decision-making and an essential mechanism for maintaining social cohesion and mutual understanding between diverse societal groups, especially when addressing collective challenges (Turchi et al., 2023). Social networks (SNs) have made communication easier than ever, connecting users from vastly different backgrounds and creating unprecedented opportunities for engagement with different viewpoints (Friedland and Kunelius, 2023; Zhang, 2023). Yet, despite these enhanced communication capabilities, online polarization and societal fragmentation continue to worsen (Habermas, 2022).

This deterioration is characterized by the increased deployment of divisive rhetoric—a set of communicative devices ranging from argumentative fallacies to rhetorical manipulations and hostile language (Zompetti, 2015). These communicative means aim at circumventing interlocutors' intellectual autonomy[1](Bassi et al., 2024; Godber and Origgi, 2023), transforming public discourse into adversarial exchanges where participants treat discussion as battles, reconfiguring other interlocutors as means to their ends rather than partners in understanding (Deutsch, 1973; Niemi, 2005).

On one hand, research focusing on political actors explains the rise of such interactions as a fundamental characteristic of populism. This rhetoric operates through in-group/out-group dynamics (Zhang, 2023), where 'the elite' is portrayed as serving 'the Others' at the expense of 'the people' (Engesser et al., 2017; Mudde and Kaltwasser 2017). Zeitzoff (2023) broadens this framework through the concept of "nasty politics," encompassing a spectrum of defamatory, dehumanizing, and threatening forms of language. Such rhetoric serves strategic functions: demarcating support bases, signaling ingroup loyalty, and capturing attention in noisy information environments (Kosmidis and Theocharis, 2020; McDermott, 2020; Zeitzoff, 2023).

On the other hand, users exposed to opposing viewpoints exhibit heightened emotional arousal and cognitive engagement (Storbeck and Clore, 2008), reactions amplified when discussions feature derogatory and emotionally charged language, ultimately correlating with greater engagement (Brady et al., 2017; de León and Trilling, 2021; Mercadante et al., 2023). The recommendation systems of social networks create destructive synergies with these controversy-specific reactions (Narayanan, 2023), as content that elicits strong emotional responses both generates higher engagement and receives algorithmic validation (Berger and Milkman, 2012; Guadagno et al., 2013; Rieder et al., 2018). This dangerous coupling between algorithmic optimization and user psychology is exacerbated by additional systemic features of social network interactions: globally audience diversity (Chen and Berger, 2013), rapid content spread (Emamgholizadeh et al., 2020), immediate visibility metrics (Rega and Marchetti, 2021), and coordinated manipulation campaigns (Cinelli et al., 2022). Taken together, these elements create self-reinforcing feedback loops that systematically reward controversial interactions, leading to the further amplification of divisive rhetoric (Heltzel and Laurin, 2024).

However, the prevalence of divisive rhetoric presents an intriguing paradox. Contrary to its widespread deployment, such antagonistic communication has proven ineffective in gaining public favor (Drake and Kiley, 2019; Frimer and Skitka, 2018, 2020; Zeitzoff, 2023). Moreover, audiences show clear aversion to divisive and nasty rhetoric (Frimer and Skitka, 2018; Kim, 2024; Wolter et al., 2023), contrasting with the observed proliferation of such content online. In light of this paradox, the main goal of this paper is to address this oddity, focusing on how users deploy specific divisive rhetorical techniques in their daily interactions across social networks.

Focusing solely on "top-down" processes provides an incomplete picture of how divisive rhetoric proliferates through digital networks. Ordinary users have become active agents—rather than passive recipients—of these communicative strategies. The disagreement space generated by SNs' contentious discussions constitutes, in fact, a social arena in which users navigate issues' positioning to articulate their own values and create shared understanding, ultimately negotiating social belonging and identity (Blitvich et al., 2013; Brown, 2022; Tomasello, 2010). This dynamic is especially evident on deindividualized platforms such as YouTube[2](Andersson, 2021), where flexible engagement patterns enable users to enter and exit conversations at will (Bou-Franch and Blitvich, 2014), diminishing personal identity salience while elevating social category membership as the dominant communicative framework (Blitvich, 2010).

To achieve these goals, users perform (dis)affiliation strategies, aimed at reinforcing connections with like-minded perspectives while distancing themselves from opposing viewpoints (Andersson, 2021; Blitvich et al., 2013). At the communicative level, these social positioning efforts are often enacted through specific practices ranging from impolite language to humor and emotional support (Andersson, 2021; Blitvich et al., 2013; Haugh and Chang, 2015; Locher and Bolander, 2015). While each of these communicative strategies serves distinct social positioning functions, divisive rhetorical techniques are particularly suited for studying (dis)affiliation in controversial debates. These vicious argumentative and rhetorical mechanisms (Bassi et al., 2024) circumvent interlocutors' intellectual autonomy and counter their epistemic interest—that is their genuine desire to acquire quality information (Godber and Origgi, 2023). By conflating (dis) affiliation processes with these manipulative dynamics, they undermine the conditions necessary for a constructive debate and transform discussion into antagonistic interaction (Deutsch, 1973; Martin, 2013). Thus, tracking these techniques provides an ideal systematic framework for examining the grassroots mechanisms through which public discourse transforms from collaborative knowledge-building into adversarial boundary-making, providing insights into how social fragmentation unfolds and spreads.

In light of this, our study empirically examines these dynamics by analyzing user interactions in YouTube comments sections across controversial and non-controversial topics. Our methodology leverages Large Language Models (LLMs) to (1) automatically detect users' stances on topics, deriving their social positioning, and to (2) track the deployment of thirteen specific divisive rhetorical techniques in their comments. Through this computational framework, we systematically investigate how different contextual factors shape the deployment of divisive rhetorical techniques in digital interactions. Specifically we examine how topic characteristics, users' stances, and their interlocutors' positions influence communicative choices in practice.

The remainder of this paper is structured as follows: Section "Related works" situates our work within the existing literature, synthesizing prior research on divisive rhetoric in digital spaces and identifying the key research gaps that our study addresses. Section "Methods" presents our methodological framework, detailing our approach to data collection across controversial and non-controversial topics, our analytical procedures for characterizing user stance and rhetorical techniques, and our methodology for mapping interactive patterns between users. Section "Analysis and results" presents our empirical findings and tests both their statistical and practical significance. Section "Discussions" provides a theoretical analysis of our findings using Social

Identity Theory and examines the connection between divisive rhetorical techniques and social positioning mechanisms in digital environments. Finally, section "Conclusion" synthesizes our theoretical and methodological contributions, discusses research impact and intervention strategies, and suggests future research directions.

Related works. Our study is positioned within the broader field of interpersonal pragmatics, that is the examination of the relational aspects of interpersonal interaction between people, which affects and is affected by their understandings of culture, society, and their own as well as others' interpretations (Locher and Graham, 2010). Research in this field investigates numerous linguistic phenomena, each serving as mechanism through which speakers navigate the complex terrain of social relationships. Within this broader domain, research on online (dis)affiliation processes has generated substantial scholarly attention, highlighting the multi-faceted ways in which users negotiate social belonging and construct identities in digital environments.

Digital platforms provide unique contexts where users employ diverse communicative strategies to signal social membership and ideological alignment. Haugh and Chang (2015) demonstrated that even supportive interactions entail complex (dis)affiliation work, with users navigating emotional support through both affiliative and disaffiliative responses in parenting discussion boards. Beyond emotional contexts, users strategically deploy valued social practices such as humor to manage identity construction, as demonstrated by Locher and Bolander (2015) in their analysis of Facebook status updates. The power of discourse to shape collective identity becomes particularly evident in politically charged environments, where McGlashan (2020) demonstrated how Twitter followers of the Football Lads Alliance —a group explicitly opposing "all extremism"—used biographical descriptions and tweets to construct collective identities aligned with radical right-wing and anti-Islamic discourse. Their findings revealed a sophisticated followership that recognized the group's calculatedly ambivalent public stance while deploying systematic discourse practices for explicit (dis)identification.

These studies collectively demonstrate that digital (dis)affiliation extends beyond simple agreement or disagreement, encompassing sophisticated communicative practices that serve identity construction and boundary-making. However, when such practices involve techniques that systematically undermine constructive dialogue, they warrant closer analytical scrutiny. In this regard, the study of impoliteness has become particularly significant, representing one of the most potent mechanisms through which users perform (dis)affiliation, while fundamentally altering the nature of digital discourse. Rather than viewing impoliteness as merely uncooperative behavior, recent scholarship reveals its strategic deployment as a vehicle for identity construction and group affiliation. Andersson (2024), for instance, demonstrates that face-threatening acts function as manifestations of digital social capital, granting users distinction through recognition by some groups while distancing others. Moreover, the structural features of digital environments amplify these dynamics, creating reinforcing cycles that intensify the deployment of impolite language. Specifically, deindividuation in anonymous, asynchronous contexts (e.g., YouTube comments section) heightens social group salience, leading users to employ impolite communication strategies to construct (dis)affiliation interactions (Andersson, 2021, 2024; Blitvich et al., 2013). This phenomenon becomes especially pronounced in politically and ideologically charged contexts (Blitvich, 2010). Andersson (2021) analyzed YouTube comments criticizing climate activism, demonstrating that impoliteness serves as a mechanism for reinforcing ideological boundaries and consolidating group identity through value homophily. This pattern also extends to more extreme forms of discourse, with Morales et al. (2025) examining how toxic and hateful speech structures in- and out-group formation processes, marking dynamics of belonging and othering that define community boundaries in digital spaces.

Our study builds closely on this research. However, while impoliteness provides valuable insights into face-threatening-based (dis)affiliative practices, the analysis of divisive rhetorical techniques offers a novel framework within relational pragmatics. This perspective examines how social identity and boundary negotiation become inherently destructive to the epistemic goals of public discourse. Specifically, divisive rhetoric captures a broad spectrum of vicious persuasion techniques (Bassi et al., 2024; Godber and Origgi, 2023)—including rational manipulative strategies, nationalistic appeals, and authority-based argumentation—that function as rhetorical strategies to mark ideological affiliation or disaffiliation (Mercier, 2020) while fundamentally violating the conditions for a constructive and reasoned discourse (Grice, 1975). This framework intersects three critical domains: the study of identity work, the analysis of social fragmentation around controversies, and the examination of misinformation dynamics.

Methodologically, our study further distinguishes itself from previous research in this domain. While existing work has employed varying analytical approaches—from purely qualitative discourse analysis to computational methods with varying degrees of sophistication (ranging from word embedding clustering (Andersson, 2021) to automatic toxic language detection (Morales et al., 2025))—most studies have been constrained in their capacity to systematically analyze large-scale interaction patterns.

However, recent advances in Natural Language Processing (NLP) and Machine Learning tools have expanded the analytical scope of online discourse: ranging from the automatic detection of users' stance on a topic (Alturayeif et al., 2023), to the extraction of pragmatic features from their messages (Mao et al., 2025).

Studies in sociolinguistic and interpersonal pragmatics have employed similar methods to examine social boundary negotiation in email communications (Bhatt et al., 2022), lexical shifts in discussions of sensitive topics (Fariello and Jemielniak, 2025), and the influence of discursive strategies on opinion change in structured debates (Monti et al., 2022). In the context of controversy analysis, computational controversy detection methods combined network and language analysis to track and measure societal division around specific topics (Hessel and Lee, 2019; Zhong et al., 2020).

Despite these methodological advances, a significant gap remains in the fine-grained understanding of how users deploy divisive rhetorical strategies in their daily social network interactions. The recent development of sophisticated methods for detecting and analyzing specific rhetorical patterns (Bassi et al., 2024; Da San Martino et al., 2019) offers an opportunity to address this gap. Such tools have been successfully applied to various contexts, including news media (Piskorski et al., 2023) and multimodal content (Dimitrov et al., 2024).

By combining these capabilities with established methods for analyzing user interactions, stance detection, and controversy dynamics, we aim at developing a more nuanced understanding of how users deploy divisive rhetorical strategies in daily online interactions. Building on this foundation, our study examines three key aspects of rhetorical behavior in social network discussions. First, we investigate the relationship between topic controversiality and the use of specific rhetorical patterns. Second, we examine how users' stance influences their rhetorical

choices. Finally, we analyze the connection between user interactive patterns and divisive rhetorical strategies uses. Specifically, we aim at answering at the following research questions:

- **Discursive Patterns and Topic Controversy**
  - RQ1: Is the controversiality of a topic connected with the frequency and the patterns of use of divisive rhetorical strategies?
- **Discursive Patterns and User's Stance**
  - RQ2: Does the user's stance toward a topic influence its use of divisive rhetorical strategies?
- **User Interactive Patterns**
  - RQ3: Do users modify their rhetorical behavior based on interlocutor's stance?

Table 1 Dataset composition showing video count and comment volume across topics.

Controversial Topics			
Topic	**Position**	**Videos**	**Comments**
Immigration			
	Pro	6	27.122
	Contra	6	28.510
	Other/ Neutral	3	27.416
	Subtotal	15	83.048
Climate Change			
	Pro	8	18.300
	Contra	6	21.582
	Other/ Neutral	3	7.517
	Subtotal	17	47.399
Non-Controversial Topics			
Archeology and Space Exploration		10	16.249

Methods

Data collection. We chose to focus on YouTube due to its distinctive characteristics as a deindividualized social network platform (Blitvich et al., 2013). Unlike more community-based platforms, YouTube features allow flexible participation, letting users freely enter and exit conversations without established community membership requirements. This creates an environment where participants remain largely anonymous and construct their identities primarily through social category membership rather than individual persona (Blitvich, 2010). The resulting deindividualized context, combined with loose community bonds, makes YouTube particularly suited for examining how users deploy divisive rhetorical techniques for ideological positioning (Andersson, 2021; Blitvich et al., 2013).

The data collection procedure followed a systematic approach to ensure a representative and balanced dataset. First, we selected immigration and climate change as controversial topics because of their strong polarizing role in contemporary political discourse. We focused on English-language content from the United States (2013-2024), where both topics exhibit well-documented patterns of political polarization and where research supports their controversial status (Falkenberg et al., 2022; Ollerenshaw and Jardina, 2023). We then crawled YouTube to identify the 100 most viewed videos for each topic, using query sets designed to capture diverse viewpoints (see Appendix 1 for complete query and video lists). We restricted our sample to videos with at least 1000 comments. We then ranked these videos by comment volume to identify those generating the highest volume of discussion. For control topics we chose archaeology and space exploration, and selected the 10 most-commented videos.

Previous research on social networks highlights how source partisanship influences user engagement and comment civility (Rho et al., 2018; Törnberg, 2022), promoting higher participation (Labarre, 2024; Yu et al., 2024) and affecting both the civility and targets comments (Su et al., 2018).

We mitigated this possible "source partisanship bias" by manually analyzing each video's content to determine its stance on the topic (see Appendix 2 for guidelines). This produced a balanced dataset of 15 videos per topic, with a deliberate emphasis on clear stances (6 supportive and 6 opposing) and a smaller neutral set (3). This distribution was chosen to better observe how users employ divisive rhetorical strategies in polarized environments, where content creators take clear positions on the issue. When necessary, we added videos from underrepresented stances to balance comment volumes across positions (see also section "Network creation"). Finally, we scraped comments from the selected videos using YouTube APIs[3], Table 1 details the composition of our dataset.

Stance detection. We classified comments into three stance categories: Contra (0), Neutral/Other (1), and Pro (2). For immigration-related content, Contra comments expressed opposition or concerns to immigration policies, while Pro comments defended immigrant rights or highlighted positive contributions. In climate change discussions, Contra referred to dismissal of climate activism or expressions of skepticism, while Pro comments defended climate action or expressed environmental concerns. The Neutral/Other category combines genuinely neutral comments (e.g., factual statements, clarifying questions) with off-topic comments addressing unrelated subjects. This consolidation primarily serves computational efficiency, though we recognize that these subcategories may reflect different rhetorical motivations while sharing the core trait of ambiguous stance positioning toward the focal topic. Specifically, since our research focuses on topical stance-driven dynamics, comments whose topical-stance cannot explicitly be determined—whether due to genuine neutrality or topic irrelevance—are grouped into one category for comparison with clear Pro/Contra positions. Importantly, "neutral" here refers exclusively to stance ambiguity regarding the focal topic (immigration/climate change), not to pragmatic neutrality, since comments may still deploy divisive rhetorical devices when addressing tangential issues in the same thread. The implications of this methodological choice for our findings are discussed in detail in sections "Discussions" and "Conclusion". Detailed information on immigration and climate stances are provided in Appendix 2.

Stance labeling on YouTube presents unique challenges due to the conversational nature of the platform and the often-implicit expression of opinions. Many comments (e.g., "Yes, I agree") can only be accurately labeled in relation to their parent comments. For context reconstruction, we implemented the methodology of Bassi et al. (2024), using their scripts to rebuild YouTube's hierarchical comment structure. Our gold standard dataset was created by two annotators with university-level training in social sciences. They manually labeled two sets of comments: 1,300 immigration-related comments (achieving an inter-annotator agreement of Cohen's $\kappa = 0.61$), and 1000 comments with their respective parent comments for context (Cohen's $\kappa = 0.68$). All

disagreements were resolved through discussion and consensus to ensure a robust gold standard.

To scale our annotation beyond manual capacity, we leveraged GPT-4o via its API, which enables cross-domain application without fine-tuning on our specific context[4]. Following the effectiveness demonstrated in previous research (Bassi et al., 2024), we provided the model with contextual information for each comment.

We validated our automated annotation approach against the established gold standard. Table 2 reports performance across test sets, both overall and by stance categories. The results show substantial agreement and consistent performance across all stance classes, validating our approach for large-scale annotation.

Divisive rhetorical techniques detection. Scaling the identification of divisive communication in online interactions requires a compromise between theoretical completeness and the capabilities of available computational tools.

To address this challenge, we operationalize the construct of divisive rhetorical strategies based on Zompetti (2015), which provides a comprehensive taxonomy of divisive rhetorical and argumentative devices. Since existing NLP tools for detecting divisive rhetorical devices are largely confined to formal texts (Bassi et al., 2024), we leveraged GPT-4o-mini[5], whose broad base knowledge enables adaptation across diverse textual genres and tasks (Hasanain et al., 2024; Sprenkamp et al., 2023). This approach allowed us to capture the nuanced use of rhetorical techniques in informal social media discourse without the costly, domain-specific annotation campaigns otherwise needed for fine-tuning on YouTube comments.

We adopted the taxonomy proposed by Sprenkamp et al. (2023) (Table 3)[6], which aligns with Zompetti (2015) and has been validated with GPT-4 on news articles. The only modification was merging the "Slogans" and "Thought-terminating Cliches" labels, as their overlap would have added unnecessary complexity without yielding further analytical insight.

Research shows that LLMs exhibit systematic cultural biases (Hu et al., 2025), so our analysis reflects both user rhetorical patterns and the model's interpretive lens rather than purely objective pattern detection. To evaluate whether this interpretive framework aligns with our theoretical construct, we validated it against expert human judgment (Bassi et al., 2025).

An expert in psychology and argumentation manually labeled 1100 YouTube comments (500 from the Immigration dataset, 500 from the Climate dataset and 100 from the Non-Controversial dataset) using this taxonomy. We then tested our model on this gold dataset to evaluate the rigorousness of its scalability to the entire corpus (details on gold-set composition and the classification prompt can be found in Appendix 3).

As shown in Table 4, the results indicate consistent overall performance ($F1\text{-}micro = 0.825$; $F1\text{-}macro = 0.696$), comparable to state-of-the-art (Sprenkamp et al., 2023), supporting reliable predictions across the dataset. In particular, the model performed especially well on techniques such as Loaded Language (F1 = 0.940), Name Calling/Labeling (F1 = 0.883), and Exaggeration/Minimization (F1 = 0.871). The strong alignment between GPT-4o-mini's classifications and expert judgment suggests that, although the model introduces interpretive elements, these do not compromise the theoretical coherence of our framework.

However, detection performance varied notably for some techniques, with Bandwagon/Reductio ad Hitlerum (F1 = 0.308) and Slogans/Thought-terminating Cliches (F1 = 0.350) proving particularly challenging. These results align with Sprenkamp et al. (2023), suggesting inherent detection challenges for certain rhetorical devices rather than methodological limitations. Additionally, lower performance on these categories reflects their limited representation in our dataset, where individual misclassifications significantly affect metrics.

Nevertheless, high precision values ($Prec. > 0.80$) across most categories indicate reliable positive predictions. This is particularly relevant for the aim of our research since we can be confident in the positive predictions. Complete data about technique distribution in the gold-set can be found in Appendix 3.

Table 2 Performance metrics of the stance classification model for climate change and immigration comments.

Stance Classification Performance

Topic	Class	Precision	Recall	F1-score	Support
Climate Change					
	Against	0.727	0.770	0.748	135
	Neutral/Other	0.766	0.725	0.745	149
	Support	0.739	0.739	0.739	115
	Macro	*0.744*	*0.745*	*0.744*	*399*
	Weight. Avg.	-	-	*0.744*	*399*
Immigration					
	Against	0.833	0.743	0.785	502
	Neutral/Other	0.602	0.730	0.660	400
	Support	0.823	0.759	0.790	403
	Macro	*0.752*	*0.744*	*0.745*	*1305*
	Weight. Avg.	-	-	*0.748*	*1305*

Table 3 Divisive rhetorical techniques used and their definitions.

Technique	Definition
Loaded Language	Emotional words and phrases intended to influence audience feelings and reactions
Name Calling, Labeling	Attaching labels or names to discredit or praise without substantive argument
Repetition	Multiple restatements of the same message (or word) to reinforce acceptance
Exaggeration, Minimization	Presenting issues as either much worse or much less significant than reality
Appeal to Fear-Prejudice	Creating anxiety or panic about potential consequences to gain support
Flag-Waving	Exploiting group identity (national, racial, gender, political or religious) to promote a position
Causal Oversimplification	Reducing complex issues to a single cause when multiple factors exist
Appeal to Authority	Using expert or authority claims to support an argument without additional evidence
Slogans/Thought-terminating Cliches	Striking ready-made phrases that use simplification and common-sense stereotypes to discourage critical thinking
Whataboutism, Straw Men	Deflecting criticism by pointing to opponent's alleged hypocrisy
Black-and-White Fallacy	Presenting complex issues as having only two possible outcomes, or one solution as the only possible one
Bandwagon, Reductio ad hitlerum	Promoting ideas based on popularity or rejecting them by negative association
Doubt	Undermining credibility through questioning motives or expertise

Network creation. To reconstruct the network of user interaction, we first aggregated comments by author, mapping each stance to a numerical value: contra (−1), neutral (0), and pro (+1). We then computed the average stance as $\frac{\sum_{i=1}^{n} s_i}{n}$, where s_i represents the stance value of the i-th comment.

The results showed a trimodal distribution with peaks at −1, 0, and 1, as shown in Fig. 1. Sharp clustering at extreme and neutral positions, coupled with low density in intermediate regions, suggests that users tend to maintain consistent stances, with only a small minority expressing mixed or shifting ideological positions.

To validate this pattern, we compared stance distributions for single-comment users (72.3% immigration, 76.6% climate) and multi-comment users (27.7% immigration, 23.4% climate). For both immigration and climate topics, multi-comment users exhibited nearly identical distributions, confirming that the trimodal pattern reflects genuine ideological positioning rather than a methodological artifact. Critically, users falling exactly at classification boundaries (±0.5) constitute only 5.6% (immigration) and 4.2% (climate) of all users. Of these, 84–86% arise from 2-comment (i.e., 1 Contra or 1 Pro + 1 Neutral) rather than sustained stance ambiguity. Additionally, among multi-comment users, 43.9% (immigration) and 47.0% (climate) maintain perfectly consistent stances across all comments and ~20% fall into the boundary cases, mainly due primarily to 2-comment mathematical positioning. The remainder—exhibiting stance variation—still demonstrated stance positioning toward ideological extremes, justifying their stance-label classification within our analytical framework.

In light of this, we defined stance classification thresholds at ±0.5, aligning with the minima in the density distribution to minimize potential misclassification between groups. Users were categorized as contra ($s \le -0.5$), neutral ($-0.5 < s < 0.5$), or pro ($s \ge 0.5$) based on these thresholds.

Network edges were constructed using the comment-reply structure (see section "Stance detection" and Bassi et al. (2024)): directed edges connected replying users to parent comment authors, while top-level comments linked users to video nodes.

As shown in Fig. 2, contra users dominate the discussion across all video types for both topics. Video stance has a more pronounced influence in climate change discussions, where user stance distributions vary based on video position. In contrast, immigration videos show a more consistent pattern regardless of video stance, with contra users as the majority (52–60%), followed by neutral (32%) and pro users (8–16%).

Table 4 Performance metrics of the divisive rhetorical techniques detection.

Divisive Techniques Detection Performance

	Precision	Recall	F1-score
Overall Performance			
Micro Average	0.840	0.797	0.818
Macro Average	0.791	0.659	0.696
Individual Techniques			
Appeal to Authority	0.652	0.577	0.612
Appeal to Fear/Prejudice	0.840	0.748	0.791
Bandwagon/Reductio ad Hitlerum	0.667	0.200	0.308
Black-and-White Fallacy	0.828	0.485	0.611
Causal Oversimplification	0.676	0.881	0.765
Doubt	0.852	0.762	0.805
Exaggeration/Minimization	0.862	0.880	0.871
Flag-Waving	0.882	0.833	0.857
Loaded Language	0.915	0.966	0.940
Name Calling/Labeling	0.869	0.896	0.883
Repetition	0.571	0.462	0.511
Slogans/Thought-terminating	0.821	0.222	0.350
Whataboutism/Straw Men	0.848	0.659	0.742

Analysis and results

RQ1: Topic controversiality and discursive patterns

General comparison. As shown in Table 5, the analysis of rhetorical technique distribution across topics reveals a significant relationship between topic controversiality and divisive rhetorical devices prevalence. Immigration discussions exhibited the highest frequency (64.49% of comments), followed by climate change (53.36%), while non-controversial topics showed substantially lower rates (34.85%). Complete data on technique-specific distribution across topics are provided in Appendix 4. Specifically, immigration discussions were nearly twice as likely to contain divisive rhetorical devices compared to non-controversial topics (Mean Diff. = 29.64; $FR^7 = 1.85\times$), while climate change discussions showed a smaller but substantial increase (Mean Diff. = 18.51; FR = 1.53×). This pattern is reinforced by the intensity of divisive rhetorical strategies usage, with controversial topics averaging more devices per comment (Immigration: 1.84; Climate: 1.38) compared to non-controversial discussions (0.76).

While these findings support the hypothesis regarding divisive rhetorical strategies' prevalence in polarized topics, the data reveal a continuous gradient rather than a binary controversial/non-controversial distinction. Immigration discussions exhibited higher divisive rhetorical devices usage compared to climate change, suggesting rhetorical technique deployment varies along a spectrum of conversational controversy.

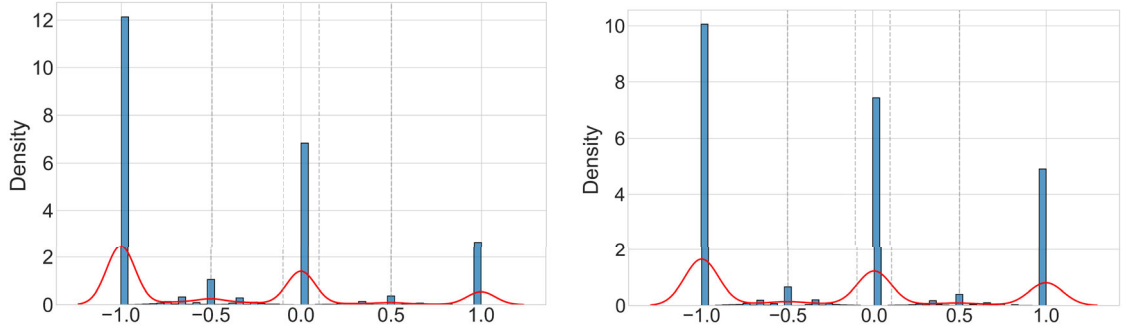


Fig. 1 Users's Stance Distribution per Topic. Panel **a**, **b** shows the Immigration and Climate distributions, respectively. Blue bars represent histogram bins. Red lines show Kernel density estimation.

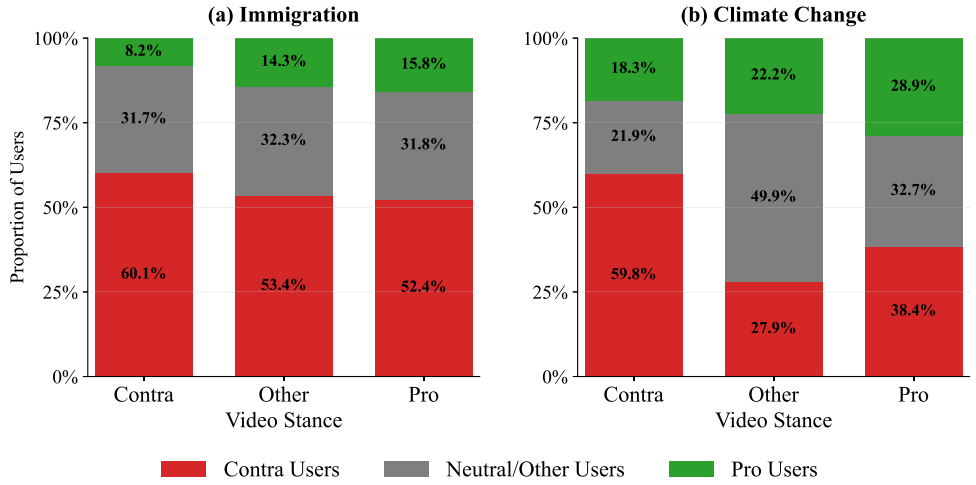


Fig. 2 Distribution of Users Stances by Video Stance Position. Panels **a**, **b** show the distribution for Immigration and Climate Change, respectively.

Table 5 Dataset characteristics (top) and statistical comparisons of rhetorical technique prevalence between topics (bottom).

Dataset Summary

Measure	Immigration	Climate	Control
Total techniques	152,780	65,238	12,304
Total comments	83,048	47,399	16,249
Avg. tech. per comment	1.84	1.38	0.76
Comments with tech. (%)	64.49	53.36	34.85

Topic Comparison Metrics

Comparison	χ^2	Mean Diff. [95% CI]	FR
Immigration vs Control	5409.24***	29.64 [28.84, 30.44]	1.85×
Climate vs Control	2438.76***	18.51 [17.65, 19.37]	1.53×
Immigration vs Climate	1897.33***	11.13 [10.71, 11.55]	1.21×

All tests were conducted with df = 1; *** = $p < 0.001$, Bonferroni-adjusted for multiple comparisons.

Notably, divisive rhetorical devices presence in ostensibly non-controversial topics demonstrates their broader social function. Consider the following exchange[8] on an archeology video:

> **Parent Comment**: [The mammut] It's not 30,000 years old.
> **Comment**: So many brain dead religious people in here [Name-Calling/labeling + Loaded Language]

This example shows how even seemingly technical discussions about archeological findings can trigger divisive rhetorical strategies deployment. Rather than engaging with archeological evidence, the commenter attacks the interlocutor employing Name-Calling/Labeling, thereby reframing the discussion as a conflict between scientific and religious worldviews. This illustrates how divisive rhetorical strategie scan be used to (1) shift the discussion from evidence-based to identity-based debate, (2) position the speaker within implicitly evoked group boundaries, and (3) othering opposing participants, extending beyond the topic's inherent controversy level (Blitvich et al., 2013).

Fine-grained comparison. To examine rhetorical differences between topics more granularly, we conducted χ^2 tests comparing technique frequencies across the different topics and measured effect sizes using Log Ratio[9] and Frequency Ratio (FR). As shown in Fig. 3, the analysis revealed two key patterns in the deployment of divisive rhetorical devices across controversial topics (see Appendix 5 for detailed results).

First, some divisive rhetorical devices exhibit content-agnostic adaptability, operating at an argumentative level that transcends specific subjects. This allows the same technique to be effectively deployed across diverse contexts, with only the semantic content adjusted to the topic. This flexibility likely explains why these divisive rhetorical devices are the most frequent across all groups. Loaded Language (Immigration: 23.91%, Climate: 21.58%, Control: 23.85%), Name Calling (Immigration: 22.19%, Climate: 17.91%, Control: 20.29%), Causal Oversimplification (Immigration: 14.01%, Climate: 15.76%, Control: 13.15%).

These divisive rhetorical devices consistently accounted for over 60% of all identified techniques across the topics, highlighting their central role in controversial discourse regardless of the specific subject matter. For example:

> **Immigration**
> **Parent Comment**: You know how it is... If Trump does something: ITS BAD!!! If Obama does it: GOOD.
> **Comment**: Why??? is the question. The answer. because Obama is payed for and Trump cant be bought. [Causal Oversimplification]

> **Climate**
> **Parent Comment**: This guy is spot on. All alarmists want is your money.
> **Comment**: All climate change deniers want oil money or have oil money, or envy oil money. [Causal Oversimplification]

At the same time, certain rhetorical devices exhibit strong topic-specificity. Flag-Waving is strongly associated with Immigration discussions (Log Ratio = 2.225, FR = 4.38×), while Doubt is substantially more prevalent in Climate Change ones (Log Ratio = −1.111, FR = 2.00×). Appeal to Authority and Exaggeration/Minimization also emerged as distinctively associated with Climate Change content (Log Ratio = −1.566, FR = 2.93× and Log Ratio = −0.626, FR = 1.47×, respectively), albeit to a lesser extent. For the control topics, we observed technique-specific

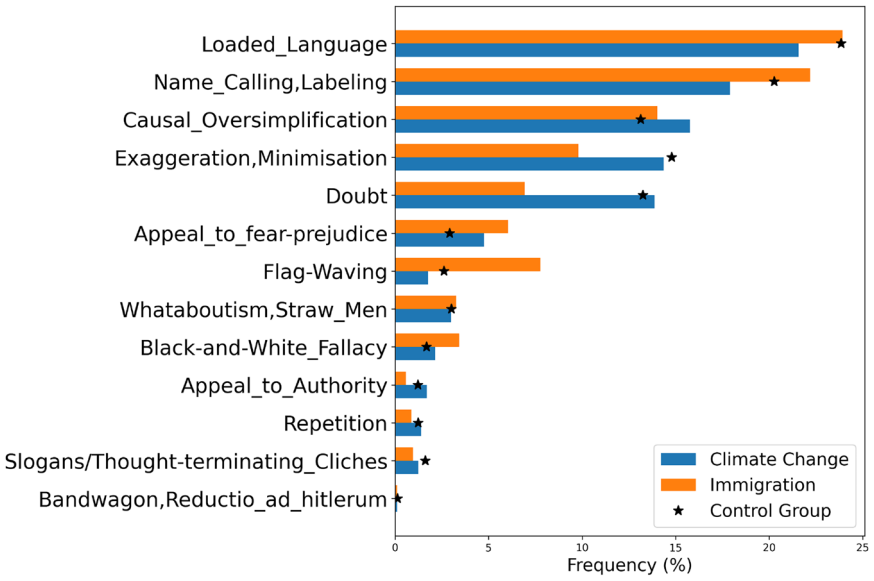


Fig. 3 Frequency Distribution of Divisive Rhetorical Techniques Across Topics.

alignments between topic pairs, with Doubt showing similar frequencies in Climate and Control discussions (13.87% and 13.25%). In contrast, Flag-Waving demonstrated significantly higher prevalence in Immigration discussions compared to both Climate and Control topics (FR =4.38× and 2.95× respectively) (see Appendix 5). In general, across the topics, the remaining techniques showed modest effect sizes ($|LogRatio| < 0.5$, FR < 1.3×), despite high overall frequency and statistical significance.

These topic-specific patterns reflect alignments between the specific techniques and the underlying semantic and epistemological characteristics of each debate. The heightened presence of "Doubt" in climate change discourse (13.87%) illustrates this: challenging the credibility and expertise of opposing voices, this rhetorical device aligns with climate change debates' focus on scientific authority, models' validity and skepticism. This scientific foundation also explains similar rhetorical patterns between Climate Change and Control topics (archeology and space exploration), as all three center on empirical evidence and expert credibility, creating comparable epistemic frameworks for rhetorical work.

> **Video/Parent Comment**: Based on models that are based on real science.
> **Comment**: Are you taking about the scientists who get grant money if they find a crisis to scare and control us with? Or the thousands of scientists who know it's all a scam? [Doubt]

Similarly, the prominent use of Flag-Waving in immigration discourse (FR = 4.38×) reflects the inherent connection between immigration policy and national sovereignty. Immigration debates center on questions of national identity and citizenship rights, making patriotic appeals particularly effective.

> **Video/Parent Comment**: It took me years, lots of paperwork, medical tests, background checks, most of my lifes savings and many visits to the INS offices before being granted the privilege of taking an oath of allegiance and becoming a US Citizen. Honor, integrity and the utmost respect for the rules and laws should be required by anyone who desires to be part of our nation.
> **Comment**: [Parent Username] made it a better place. Nobody is saying they can't come in—it just can't come in illegally and unregulated. We cannot turn our back on fellow Americans and veterans. We MUST help our own FIRST! [Flag-Waving]

These patterns suggest that, while the fundamental structure of rhetorical technique usage remains similar across topics, their intensity and specific combinations vary according to the nature of each topic.

RQ2: Discursive patterns and user's stance

General comparison. In this section, we analyze whether, and to what extent, a user's stance on a topic influences both the amount and types of rhetorical devices they employ in discourse. We examined this relationship across the two datasets focusing on controversial topics (Immigration and Climate Change), as we assumed users do not take a specific stance on non-controversial topics.

Given unequal sample sizes across stance groups and Levene's test showing significant variance heterogeneity for both the Immigration ($F = 1.02 \times 10^8$, $p < 0.001$) and the Climate Change ($F = 8.08 \times 10^7$, $p < 0.001$) datasets, we employed Welch's ANOVA and Games-Howell post-hoc tests, which accommodate both unequal sample sizes and variances.

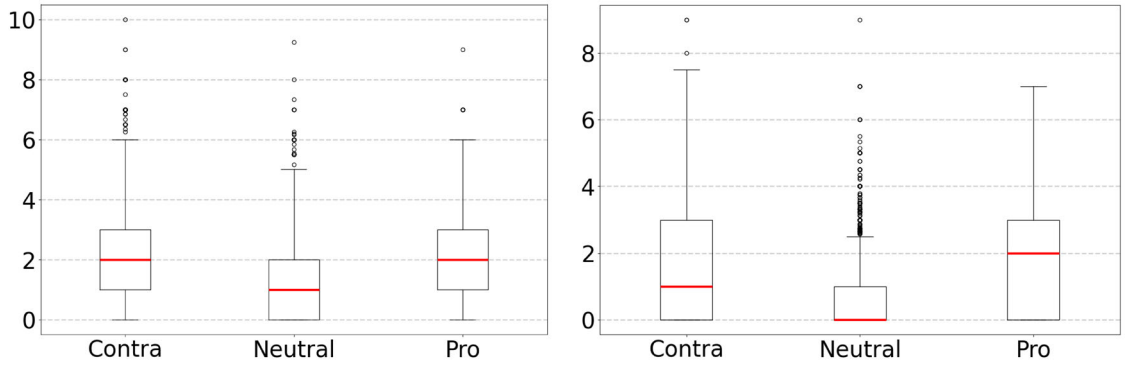
Table 6 and Fig. 4 present the average number of divisive rhetorical strategies used per comment across different stances for both topics:

Our analysis reveals that expressing a particular stance influences the use of divisive rhetorical strategies. As shown in Table 7, the Welch's ANOVA F-statistics ($F = 1.02 \times 10^8$ for Immigration and $F = 8.08 \times 10^7$ for Climate Change, both $p < 0.001$) indicate significant differences between stance groups in each dataset. However, given our large sample size, we focused on Cohen's d effect sizes[10] to better understand the practical significance of these differences.

For Immigration discussions, we observe a clear pattern of divisive rhetorical strategies usage across stances. Users with a Contra stance show the highest usage ($M = 2.12$, SD = 1.64), followed closely by Pro-stance users ($M = 2.03$, SD = 1.53), while Neutral users show markedly lower usage ($M = 1.19$, SD = 1.29). While all differences are statistically significant, the Cohen's effect sizes reveal a more nuanced picture. The difference between users holding strong positions (Contra vs Pro) is minimal (Mean Difference = 0.09,

Table 6 Usage patterns of divisive rhetorical devices by topic and stance.

Topic	Stance	Users (*n*)	Comments	Median	Mean	SD
Immigration	Contra	24,595	39,003	2.00	2.12	1.64
	Neutral	17,612	35,834	1.00	1.19	1.29
	Pro	5379	8211	2.00	2.03	1.53
Climate	Contra	12,009	18,492	1.00	1.46	1.45
	Neutral	10,578	19,329	0.00	0.68	1.05
	Pro	5873	9578	2.00	1.74	1.43



(a) Immigration Related Videos (b) Climate Change Related Videos

Fig. 4 Average Divisive Rhetorical Techniques per Comment per Stance. Panels **a**, **b** show the distributions for Immigration and Climate Change, respectively.

Table 7 Statistical analysis of stance differences in technique usage.

Topic	Test	Mean Difference	Test Statistic	Cohen's *d*
Immigration	Welch's ANOVA	—	$F = 1.02 \times 10^{8***}$	—
	Contra vs. Neutral	0.93	$t = 64.75^{***}$	0.61
	Contra vs. Pro	0.09	$t = 3.85^{***}$	0.06
	Neutral vs. Pro	−0.84	$t = -36.33^{***}$	−0.62
Climate Change	Welch's ANOVA	—	$F = 8.08 \times 10^{7***}$	—
	Contra vs. Neutral	0.79	$t = 47.13^{***}$	0.62
	Contra vs. Pro	−0.28	$t = -12.22^{***}$	−0.19
	Neutral vs. Pro	−1.07	$t = -50.05^{***}$	−0.89

Note: ***Indicates $p < 0.001$.

$d = 0.06$), whereas both groups differ substantially from Neutral users (Contra vs Neutral: Mean Difference = 0.93, $d = 0.61$; Pro vs Neutral: Mean Difference = −0.84, $d = -0.62$).

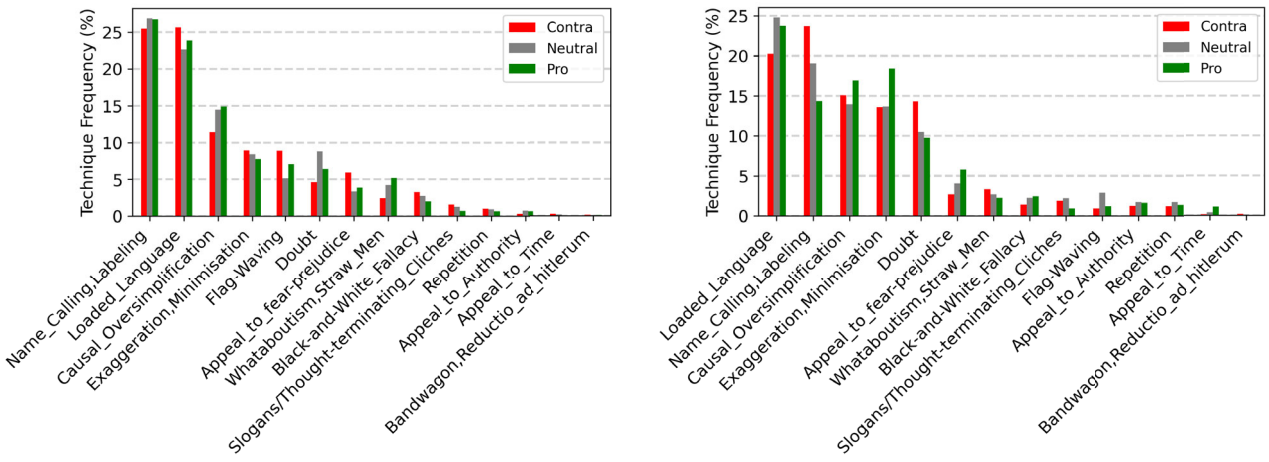
Climate Change discussions mirror this pattern. Here, Pro-stance users employ slightly more rhetorical techniques ($M = 1.74$, $SD = 1.43$) than Contra users ($M = 1.46$, SD = 1.45), yet both groups show substantially higher usage than Neutral users ($M = 0.68$, SD = 1.05). Effect sizes again demonstrate that the key distinction is between users with strong positions and those with neutral stances (Contra vs Neutral: $d = 0.62$; Pro vs Neutral: $d = -0.89$), rather than between opposing viewpoints (Contra vs Pro: $d = -0.19$).

These findings suggest that the tendency to employ divisive rhetorical strategies is more strongly associated with holding a decisive stance on an issue, irrespective of the position taken. The remarkably symmetric pattern in rhetorical technique usage between Pro and Contra users (with minimal effect sizes of $d = 0.06$ for Immigration and $d = -0.19$ for Climate Change) indicates that stance commitment, rather than stance direction, is the primary driver of divisive rhetorical strategies deployment.

Notably, users with neutral positions consistently demonstrate lower propensity to use such rhetorical devices, further supporting this pattern. However, it is important to note that the Neutral category encompasses both "truly non-committal comments" and off-topic discussions, which may blur the patterns observed in this group (discussed further in sections "Discussions" and "Conclusion").

Fine-grained comparison. Figure 5 presents the relative frequencies of divisive rhetorical strategies across user stances. We analyzed these distributions using Chi-squared tests with Bonferroni corrections for multiple comparisons (see Appendix 6). While almost all comparisons were statistically significant, the large sample size (see Table 6) required focusing on effect sizes, particularly Log Ratios, to evaluate practical significance.

A key finding is the overall similarity in divisive rhetorical devices usage across different stance groups, with most techniques exhibiting minimal practical differences despite statistical significance. This consistency is particularly evident in fundamental techniques like Loaded Language, which showed minimal variation in relative frequency across stance groups in Immigration (C-N Lg.Rt = 0.12, C-P Lg.Rt = 0.07, N-P Lg.Rt = −0.05) and Climate Change discussions (C-N Lg.Rt = −0.20, C-P Lg.Rt = −0.16, N-P Lg.Rt = 0.04).



(a) Immigration Related Video (b) Climate Change Related Video

Fig. 5 Techniques Relative Frequencies for Users' Stance. Panels **a**, **b** show results for Immigration and Climate Change related videos, respectively, reporting how often each technique is used by Contra (red), Neutral (grey) and Pro (green) users within the topic.

Other techniques followed topic-dependent patterns. For instance, Causal Oversimplification was more prevalent among Pro and Neutral users than Contra users (C-P Lg.Rt = −0.27, C-N Lg.Rt = −0.23), while Flag Waving exhibited a polarized usage pattern, with both Pro and Contra users employing it more frequently than Neutral users (C-P Lg.Rt = 0.23, N-P Lg.Rt = −0.32).

In climate change discussions more pronounced stance-based variations emerged. Contra users demonstrated significantly higher usage of Name Calling (Lg.Rt = 0.50) and Doubt (Lg.Rt = 0.38) than Pro users. Conversely, Pro-stance users showed greater propensity for using Appeal to Fear/Prejudice compared to contra users (Lg.Rt = −0.76) (see Fig. 5).

Although climate change discussions showed more pronounced variations between stances—suggesting that the relationship between social positioning and divisive rhetorical devices usage may be moderated by topic context—the overall pattern across both topics indicates that users tend to employ similar divisive rhetorical strategies regardless of their ideological position, with most differences remaining modest in practical terms. This observation aligns with and expands our earlier findings (see section "RQ2: Discursive patterns and user's stance"), configuring divisive rhetorical devices as a shared rhetorical toolkit that can be effectively deployed to support divergent stances.

To illustrate this versatility, we present two comparative examples showing how the same rhetorical technique is employed by users on opposing sides: Whataboutism in Immigration discussions and Appeal to fear/prejudice in Climate Change debates.

Pro Immigration	Contra Immigration
Parent Comment: [your parents] immigrated LEGALLY into the USA! NOT just walking in and expecting a massive handout! **Comment:** If you are part of the UN, then all countries should accept Migrants, JUST LIKE when YOUR family were Migrants and they entered the USA, remember that? [Whataboutism]	**Parent Comment:** How do you deport someone more humanely? Tell em "be careful out there" and hand them a lollypop after releasing them? **Comment:** Meanwhile, it was Obama who separated families and put people in cages. Why take responsibility when you can just blame the next administration? [Whataboutism]

Pro Climate Change	Contra Climate Change
Video/Parent Comment: I honestly think climate is the least of our problems right now **Comment:** When Florida floods remember we warned y'all [Appeal to fear-prejudice]	**Video/Parent Comment:** The intended purpose is to raise taxes and put more parts of our lives under Govt control. **Comment:** YES, SOON YOU WILL BE MOVED INTO CLIMATE CAMPS FOR REEDUCATION LIKE IN CHINA WITH THE WEGERS [Appeal to fear-prejudice]

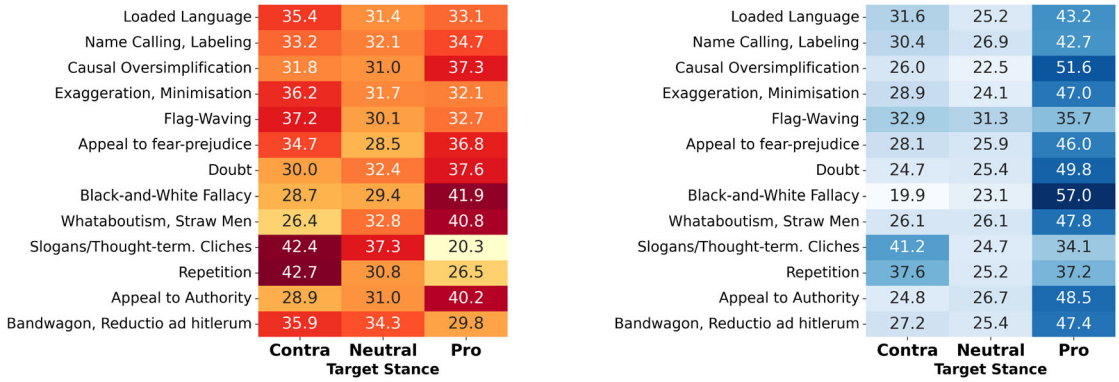
RQ3: User interactive patterns

Divisive language patterns and target user's stance. To investigate how users adapt their rhetorical patterns to interlocutor stance, we analyzed the deployment of divisive rhetorical strategies across stance-based interactions. Following the methodology in section "Network creation", we assigned stances to users and mapped comment-author interactions via network topology analysis.

For each stance group (source), we quantified the proportion of each divisive rhetorical strategy used in both ingroup (interactions within the same stance) and outgroup interactions (interactions with other stances). Specifically, we calculated the percentage of each device used across target stance groups by normalizing raw frequencies against the total amount of techniques employed by the source group.

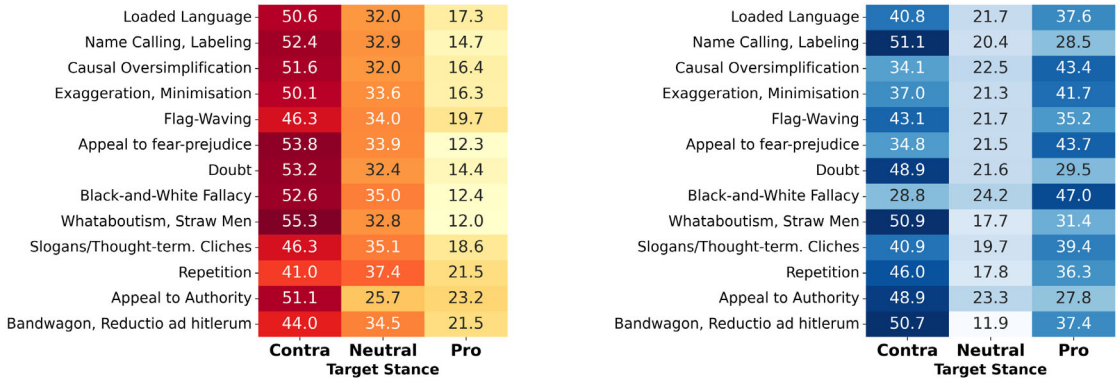
A notable class imbalance characterized our dataset, with Contra users predominating across both topics—most markedly in Immigration discussions—while Pro users were the minority and Neutral users occupied an intermediate position (Fig. 1). This imbalance may introduce bias in our analysis. In fact, initial results indicated techniques were predominantly directed at Contra users (see heatmaps in Appendix 7); however, this pattern likely reflects probabilistic interaction opportunities rather than intentional targeting, an effect particularly pronounced in Immigration discussions where stance asymmetry peaks.

To address this sampling bias, we implemented a bootstrap resampling procedure to create balanced user groups. For each iteration, we randomly selected an equal number of users from each stance group, with the sample size determined by the



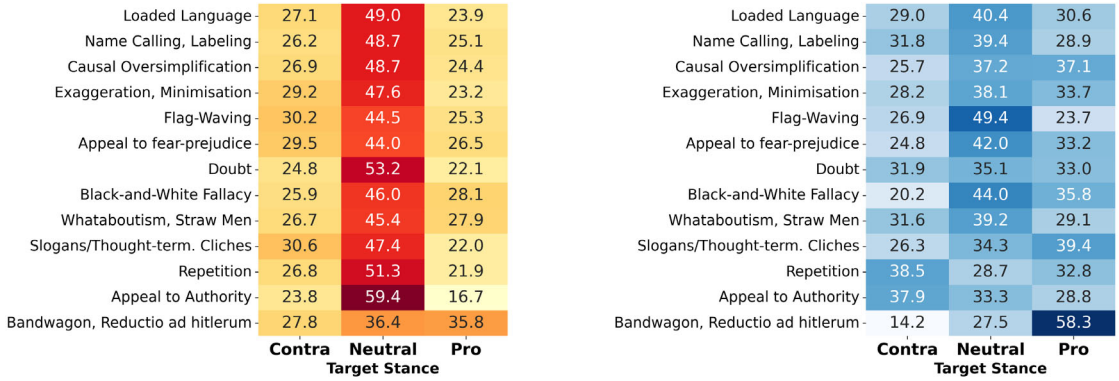
(a) Immigration (b) Climate

Fig. 6 Technique Usage by Target Stance - Contra Users. Heatmaps (**a**) and (**b**) show the results for Immigration and Climate Change, respectively, displaying the average relative frequency of techniques used by Contra users toward Contra, Neutral and Pro targets (after bootstrapping resampling).



(a) Immigration (b) Climate

Fig. 7 Technique Usage by Target Stance - Pro Users. Heatmaps (**a**) and (**b**) show the results for Immigration and Climate Change, respectively, displaying the average relative frequency of techniques used by Pro users toward Contra, Neutral and Pro targets (after bootstrap resampling).



(a) Immigration (b) Climate

Fig. 8 Technique Usage by Target Stance - Neutral Users. Heatmaps (a) and (b) show the results for Immigration and Climate Change, respectively, displaying the average relative frequency of techniques used by Neutral users toward Contra, Neutral and Pro targets (after bootstrap resampling).

minority stance group (Pro users in both cases). Using $N = 1000$ iterations for robust estimation, we analyzed the distribution of divisive rhetorical devices across target stances and computed final distributions by averaging over all iterations. We assessed reliability using standard deviation (Std Dev) and coefficient of variation (CV). Results largely demonstrated consistency across iterations (detailed in Appendix 7), although underrepresented techniques were less reliable due to their low frequency and the resulting risk of omission (see also section "Divisive rhetorical techniques detection").

As shown in Figs. 6, 7, 8, our analysis reveals distinct patterns in the use of divisive rhetorical devices across different topics and user stances.

Contra Users. In *Climate Change* discussions, users holding a contra stance exhibit clear rhetorical behavior, consistently deploying their divisive rhetorical strategies toward pro users (see Fig. 6b). This pattern expresses an adversarial use of divisive rhetorical strategies, systematically targeting opposing viewpoints on the topic. The following exchange exemplifies this dynamic: a contra-climate user attacks a pro-climate comment by deploying a "Black-and-White Fallacy" to dismiss environmental concerns, reducing the complexity of climate-induced species extinction to a simplistic binary choice between accepting polar bear extinction as "natural evolution" or abandoning human existence:

> **Pro-Climate Parent Comment**: When scientists say bears are going extinct, I want people to realize what it looks like. Bears are going to starve to death," said Paul Nicklen (the photograph). "This is what a starving bear looks like.
> **Contra-Climate Reply Comment**: It's called evolution, if you don't like it give up your seat in the race. [Black-and-White Fallacy][11]

On the other hand, in *Immigration*-related discussions, this targeting of users with opposing stances dissipates. Contra users show a more evenly distributed use of rhetorical techniques across all stance groups, with minor variations (see Fig. 6a). The exception appears in techniques like "Slogans/Thought-terminating Cliches" and "Repetition", which show a stronger tendency for interactions within the same stance (directed towards the Contra stance-group in the 42.4% and 42.7% of the cases, respectively). This indicates that in Immigration discussions, Contra users engage in more complex communicative dynamics that extend beyond simple targeting of opposing stance-holders, revealing an affiliative function of divisive rhetorical strategies aimed at reinforcing like-minded solidarity. The following exchange exemplifies this affiliative behavior. A contra-immigration user responds to a comment with a similar stance, using a rallying slogan aimed at consolidating a shared political alignment:

> **Contra Immigration Parent Comment:** Finally people, even New Yorkers, are growing a spine and realizing tolerance without limit is foolish.
> **Contra Immigration Reply:** Vote RED [Slogans/Thought-terminating Cliches]

Pro Users. Pro users demonstrate the most pronounced and consistent divisive rhetorical behavior across topics, predominantly targeting contra users. This pattern is particularly strong in *Immigration* discussions, where nearly all techniques are deployed more than 50% of the time toward contra users, with Whataboutism (55.3%), Appeal to fear-prejudice (53.8%), and Doubt (53.2%) showing the strongest patterns (see Fig. 7a). The following exchange illustrates this dynamic, where a Pro-immigration user responds to a Contra-immigration comment by employing Whataboutism to deflect criticism of immigrant integration by redirecting attention to Western military interventions, shifting the discussion from individual responsibility to geopolitical causation, and challenging the opposing stance through allegations of hypocrisy:

> **Contra Immigration Parent Comment**: Its always someone else's fault, but it's never the fault of "guests" who are given a second chance of living a healthy and safe life, but refuse to adapt and integrate. Respect to Denmark for protecting their country, something countries like France,-Netherlands,and Belgium failed to do. Of course there will be a proper percentage of immigrants behaving themselves, but you can't blame the danish government for putting conclusions out of recent developments.
> **Pro Immigration Reply Comment**: Wasn't it the US and European Union who destroyed Syria, Iraq and the whole middle east by funding ISIS. Isn't it what led them to seek refuge in Europe in the first place? [Whataboutism]

In *Climate Change* discussions, although the deployment of divisive rhetorical strategies toward Contra users remains significant, there's more variation. Some techniques, such as Name Calling (51.1%) and Whataboutism (50.9%), maintain strong contra-directed patterns, while others exhibit a more balanced distribution across stance groups (see Fig. 7b).

Neutral Users. This category encompasses two distinct types of comments: those showing minimal social identity work with minimal or no technique usage due to genuine non-commitment, and those that avoid explicit topical stance-taking while still deploying divisive rhetorical strategies. The first type is illustrated by the following example:

> **Pro Parent Comment**: This is exactly why we stopped using freon.
> **Neutral Climate Reply Comment**: as long as there is air conditioning freon is still being used

In the second case, while not explicitly taking a stance on the topic, Neutral users deploy divisive rhetorical devices for other relational purposes. In *Climate Change* discussions, the divisive rhetorical devicesare deployed evenly across stance groups (see Fig. 8b). By contrast, in *Immigration*-related discussions, Neutral users who deploy techniques predominantly direct them toward other neutral users, with techniques like "Appeal to Authority" and "Doubt" showing particularly strong concentration within the neutral stance group (59.4% and 53.2% respectively; see Fig. 8a). For instance:

> **Neutral Immigration Parent Comment**: Wow, reading comprehension is in short supply on this thread...
> **Neutral Immigration Reply Comment**: I am extremely disappointed by these "news" channels posting meme clips for a living. They are essentially feeding on dividing the nation over stuff. They can't just post a misleading clip from a hearing and call it journalism. [Doubt]

This exchange demonstrates the complexity of Neutral users' relational work: although both Parent and Reply comments avoid explicit reference and positioning to immigration policy, they are not pragmatically neutral. The Parent comment attacks the quality of discourse from all participants (attacking both sides), while the Reply comment expands this statement, employing Doubt to critique media practices and positioning both users as superior to the polarized debate participants. These findings indicate that divisive rhetorical devices function as indicators of topical commitment and social identity construction among Neutral users as well, albeit via more sophisticated rhetorical mechanisms. Specifically, we observed how the strategic deployment of these techniques facilitates meta-discursive positioning, establishing Neutral users' epistemic superiority over overtly committed participants. Although these results reinforce our interpretation of divisive rhetorical strategies as mechanisms for signaling stance and performing social identity, the intricate patterns observed within the Neutral category warrant future research to better distinguish between genuinely non-committal

comments, off-topic ones, and those employing indirect ideological positioning strategies (see sections "Discussions" and "Conclusion").

Discussions

This study examines how users deploy divisive rhetorical strategies in online interactions—communicative strategies designed to circumvent interlocutors' intellectual autonomy during debates (Bassi et al., 2024; Godber and Origgi, 2023). Overall, our findings contribute to understanding (dis)affiliation processes in digital environments by showing how rhetorical choices function as mechanisms for social positioning across topic characteristics, user stance commitment, and interactive dynamics.

RQ1: Is the controversiality of a topic connected with the frequency and the patterns of use of divisive rhetorical techniques? Addressing RQ1—the link between topic controversiality and rhetorical techniques (section "RQ1: Topic controversiality and discursive patterns")—we found compelling evidence of greater deployment of divisive rhetorical devices in controversial discussions. Immigration discussions exhibit nearly twice the frequency of rhetorical techniques compared to non-controversial topics, while climate change discussions occupies an intermediate position. While algorithmic recommendation systems contribute to the visibility and exacerbation of such controversies (Narayanan, 2023; Stray et al., 2023), our analysis centers on the psychosocial and relational dynamics shaping users engagement with these topics. Unlike non-controversial topics, discussions about immigration or climate change often challenge core aspects of users' worldviews and their perceived social roles, tying them to social identities and issue-based group memberships (Lamont et al., 2015). As underscored by Social Identity Theory (SIT) (Tajfel et al., 1979), this "link" prompts users to defend their positions as proxies for social identity—affirming their own view while rejecting opposing ones. Consequently, when discussions are tied to issue-based group affiliation, expressing a position often transcends the rational evaluation of competing ideas. In this way, the social stakes pervading the debate transform constructive dialogue into destructive conflict, as interlocutors lose their role as collaborative meaning-makers and are cast as either enemies or allies (Deutsch, 1973). The deployment of divisive rhetorical strategies can thus be understood as both a symptom and a mechanism of this deindividualization process, wherein opponents' intellectual autonomy can be systematically undermined because they are no longer viewed as autonomous agents worthy respectful engagement, but as mere means to one's goals (Godber and Origgi, 2023).

Additionally, the analysis of specific rhetorical patterns reveals a duality in users' deployment of these rhetorical strategies. On one hand, we observed striking consistency in the core structure of rhetorical technique usage across topics—with Loaded Language, Name Calling, and Causal Oversimplification accounted for about 60% of identified techniques—aligning with Zompetti (2015)'s "standardized divisive rhetorical toolkit". On the other hand, we found notable topic-specific variations, such as more frequent Flag-Waving in immigration discussions and Doubt in climate change debates, underscoring users' strategic agency in adapting these tools to match the semantic aspects of different topics (Grabill and Pigg, 2012).

RQ2: Does the user's stance toward a topic influence its use of divisive rhetorical techniques? Our second research question (section "RQ2: Discursive patterns and user's stance") explored how users' issue positions shape their rhetorical choices in online discussions. The findings show a consistent symmetry in rhetorical behavior: users with pronounced stances, whether Pro or Contra, use divisive rhetorical strategies far more often than Neutral users, while differing little from one another.

These results can be understood through the concepts of topic salience, stance commitment and their relationship to social identity construction (Flowerday and Shell, 2015; Keusch, 2013; Zarrinkalam et al., 2024). When users choose to engage in controversial discussions, they face a series of commitment decisions: whether to participate, adopt a specific stance, and which position to take. Each choice can thus represent a deeper investment in the topic and, consequently, a different gradient of salience for their identity formation. This choice, in turn, can be not aligned with an individual's broader ideological identity, but instead stem from specific aspects of the topic. In fact, while stance commitment and ideological identity are frequently related and intertwined, they remain distinct (Mason, 2015). Comellas and Torcal (2023) emphasize that ideological identity can persist even when it conflicts with the stance held on a specific topic. This is particularly relevant considering YouTube's volatile interactions, which make the validation of one's own values a compelling incentive to engage in discussions (Andersson, 2021), ultimately promoting this "issue-based" engagement. Additionally, the deindividualized nature of this platform elevates social identity as the predominant facet of communication, leading users to primarily engage through (dis)affiliation with stance-based groups formed around the topic (Blitvich et al., 2013). Consequently, the more frequent use of divisive rhetorical strategies by clear-stance users (either pro or contra) reflects a strategy for signaling commitment to their group. The contrast between users with clear stances and those in our Neutral category —which includes both genuinely neutral and off-topic comments [12]—suggests that stance commitment toward to the focal topic, rather than general engagement or ideological identity, drives the use of such communicative devices. Building on our previous discussion, these rhetorical strategies aim to establish clear social boundaries necessary for identity work through issue-based group differentiation. Finally, our findings show that users employ remarkably similar divisive rhetorical strategies regardless of stance. This observation suggests that divisive rhetorical devices constitute a shared rhetorical toolkit that can be effectively deployed to support divergent positions, further reinforcing the Zompetti (2015) "standardized divisive rhetorical toolkit" discussed in RQ1. Rather than reflecting ideological content, the symmetrical use of these techniques underscores their function as universal instruments for boundary-making and social signaling in controversial discussions.

RQ3: Do users modify their rhetorical behavior based on interlocutor's stance? Our investigation of how users modify their rhetorical patterns based on their interlocutor's stance (section "RQ3: user interactive patterns") offer insights into the dynamics of social identity negotiation in digital spaces.

The analysis reveals a sophisticated pattern of rhetorical deployment serving two functions: offensive (directed towards users holding a different stance, to attack their position) and affiliative (directed towards like-minded users to reinforce bonds around shared opinions). This duality manifests most clearly in users with strong stances (Pro and Contra).

Building on interpersonal pragmatics research, these two "declination" of the divisive rhetorical strategies can be framed as specific mechanisms through which users navigate the complex terrain of digital social relationships (Locher and Graham, 2010). Following Blitvich et al. (2013)'s framework, these strategies serve as vehicles for, respectively, disaffiliative and affiliative responses,

enabling users to negotiate social belonging while constructing clear issue-based boundaries.

This aligns with studies on the relational function of impoliteness—used as a face-threat to distance from opponents—and politeness—used to claim common ground with like-minded users (Andersson, 2021; Blitvich et al., 2013). From an argumentative perspective, (im)politeness parallels divisive rhetorical strategies that to perform ethotic functions—communicative means aimed at manipulating opposing arguments by either attacking the speaker's personality and credibility or exalting their virtues (Budzynska and Reed, 2012).

The following examples illustrate how the "Name-Calling" technique allows users to perform both actions. In the first case, it undermines the opponent's claim by attacking on his/her persona, taking distance from his/her position. In the second case, the attack targets a third (opposing) part to create common ground with the immediate interlocutor. Despite their differences, both cases can be framed as instances of "othering" (Blitvich et al., 2013) — the process by which we reenact our own positive identity through the stigmatization of another (both directly and indirectly) (Canales, 2000).

> **Pro Climate Parent Comment**: So what? The rainforest is huge. That doesn't change the fact that Bolsonaro is destroying it at record levels. Thankfully, he'll be gone soon enough.
> **Contra Climte Change Reply Comment**: you're a disgusting paid DNC troll with many accounts [Name-Calling/Labeling]

> **Contra Immigration Parent Comment**: I swear these people never even learned the words "i am/was wrong"
> **Contra Climte Change Reply Comment**: Just overgrown toddlers who throw hissy fits when you don't let them do whatever they want when they want. [Name-Calling/Labeling]

At the same time, focusing on divisive rhetorical strategies reveals a wider range of communicative devices that transform discussions into manipulative arenas for negotiating social identity. This framework reveals how users intersect topical discussions to broader values, enabling them not only to contest opponents' statements, but also to negotiate their own and other's perceived role in society.

The following example demonstrates this through another instance of othering. In this case the "excluded" entity used to signal distance from the parent comment is not any discussion participants, but illegal immigrants as a broader social category. By invoking (indirectly) "illegal immigrants" as the stigmatized out-group, the user simultaneously performs identity work within the digital conversation while reinforcing real-world social hierarchies and exclusionary narratives that extend beyond the YouTube platform.

> **Pro Immigration Parent Comment**: it is not your land only. It is founded by immigrants and for immigrants.
> **Contra Immigration Change Reply Comment**: news flash it's the land of those who are born here and those who come here legally. America is not for those who want to come illegally. Get that through your head [Flag-Waving].

Using divisive rhetorical strategies as a theoretical lens to study social identity work in online environments shed light on the mechanisms linking social identity, online discourse polarization, and offline societal divisions. This contributes to the study of what McCoy and Somer (2019) characterize as pernicious polarization, which fragments societies along multiple dimensions. Additionally, focusing on divisive rhetorical deviceslets us trace not only the link between YouTube discussions and social identity work, but also their intersection with misinformation dynamics. The following example shows how, starting from a relatively reliable (not-misinformative) comment, identity work through emotive manipulation can expand and reinforce like-minded perspectives. The reply simultaneously signals agreement with the original comment—supporting the user's social identity work—while introducing manipulative elements that may distort the informative value of the overall conversation (Cinelli et al., 2021).

> **Pro Climate Parent Comment:** The melting ice also has methane stored in it so when the ice melts, it's releasing methane. Congrats VOX this is your best video by far, powerful yet simple.
> **Pro Climate Reply Comment:** And dormant ancient pathogens that most humans don't have immunity to are also released. Hooray! More storm surges, extreme weather, fires, AND pandemics! [Appeal to Fear]

The complexity of social identity work revealed by our theoretical-methodological framework is further demonstrated in our findings on Neutral users in RQ3. As shown in the example described in the "Results" section, when users deploy divisive rhetorical devices while avoiding explicit stance-taking, their identity work can be framed as "negative identity work"—"meta-positioning" themselves as disengaged from the debate itself rather than from particular positions within it. This functions by using the discussion as a vehicle to signal distance from the very act of taking sides, rather than allegiance to a particular position. This interpretation aligns with Andersson (2024)'s analysis of creative impoliteness as expression of digital social capital. While some users may avoid taking sides out of fear of conflict or a sense of disempowerment (Strickler et al., 2024), others may engage in meta-positioning that reflects a form of distinction-seeking behavior, presenting themselves as operating from a different vantage point than the polarized participants—effectively claiming superiority through non-participation rather than argumentation. However, it is important to acknowledge that our methodology does not distinguish between genuinely neutral comments and off-topic discussions, which represents a limitation in our ability to fully characterize this form of identity work, making this aspect something that warrants further investigation.

Conclusion

While social networks offer unprecedented opportunities to discuss and create common ground in our fragmented social reality (Fairclough, 1999), users often employ divisive rhetoric that undermines these collaborative potentials (Bail, 2022). Rather than fostering constructive dialogue, these linguistic strategies function to create and maintain social boundaries. As Sciortino (2024) argues, boundaries are not consequences of pre-existing group differences but the very mechanisms through which groups are created and sustained. Through this lens, social networks emerge as arenas where social positions are continuously negotiated and maintained interactively (Bail, 2022). Our research investigated how this boundary-making process manifests in digital spaces by tracking users' demployment of divisive rhetorical devices as interactive tools for shaping and sustaining social divisions. Specifically, we addressed three key questions about these dynamics. First, we examined whether topic controversy influences the frequency and patterns of divisive rhetorical

devices deployment (RQ1). Our findings demonstrate that controversial topics — particularly immigration — exhibit significantly higher frequencies of divisive rhetoric than non-controversial ones, with users adapting their rhetorical strategies to match the semantic demands of each topic. Second, we explored how users' stances on topics influence their rhetorical choices (RQ2). The results reveal that users with clear stances (Pro or Contra) use divisive rhetorical strategies at significantly higher rates than Neutral users, with remarkable symmetry between opposing positions—suggesting that stance commitment, rather than ideology, drives rhetorical deployment. Third, we analyzed whether users adjust their rhetorical behavior to their interlocutors' stances (RQ3). Our findings show a dual function: divisive rhetorical strategiesserve both offensive purposes (attacking opposing stances) and affiliative ones (reinforcing like-minded bonds), acting as strong communicative proof of stance commitment through strategic adaptation to addressees—suggesting that users calibrate their strategies to the identity threats and opportunities of each discursive contexts.

Taken together, these findings extend beyond documenting rhetorical patterns to show how social networks have fundamentally transformed the nature of building social identity in public discourse (Bail, 2022). This transformation shows how online discourse has evolved into a complex arena where dichotomizing rhetorics serve as instruments of social positioning in an increasingly interconnected digital public sphere.

Research impact. Our findings contribute to understanding the societal implications of divisive rhetoric in the following ways:

Communicative Modalities and Digital (Dis)affiliation: on a theoretical level our results illuminate how users enact (dis) affiliation in online environments through divisive rhetorical strategies, extending existing research by revealing how epistemic manipulation serves social boundary construction and how destructive discourse patterns emerge from grassroots social positioning processes. These findings reveal how social identity work can systematically undermine the epistemic conditions necessary for constructive democratic dialogue, raising questions about normative frameworks for governing digital communicative practices.

Methodological Framework for Analyzing Digital Discourse: we demonstrated how LLMs (GPT-4o) can be used to automate stance and divisive rhetorical devices detection in YouTube, enabling fine-grained large-scale analysis of users' interactions. This approach provides a framework for examining how divisive rhetoric intersects with other digital phenomena, such as misinformation (Cinelli et al., 2021) and controversy dynamics (Bassi et al., 2025). As stressed by Nannini et al. (2025), such computationally-based tracking could also allow to evaluate intervention and policy effectiveness in reducing harmful online behavior.

Informing Content Moderation and Intervention Strategies: current moderation approaches face fundamental limitations, as they primarily target content that explicitly violates platform policies while overlooking sophisticated rhetorical manipulations that drive controversy amplification (e.g., fear-based appeals) (Saha et al., 2023, 2021; Stray et al., 2023). While algorithmic suggestion-management solutions face implementation challenges due to social networks' economic dependence on engagement-driven interactions (Narayanan, 2023; Stray et al., 2023), our work addresses detection issues by demonstrating systematic tracking of nuanced vicious techniques in controversial discussions. This can enable bottom-up interventions by reframing destructive interactions in terms of specific argumentative and rhetorical techniques. Unlike conventional content moderation approaches that rely on criticized top-down content removal (Douek, 2021; Puig Larrauri and Morrison, 2022; Ravndal, 2018), bottom-up strategies engaging users and communities (Bjornsgaard 2023)—such as NGO-led counterspeech initiatives (Chung et al., 2021; Gagliardone et al., 2015; Jia and Schumann, 2025)—could leverage our framework's detection mechanisms to devise countermeasures informed by established argumentation research (Baggini, 2016; Lewandowsky et al., 2020). Such training has proven effective for educational and counter-narrative approaches (Hruschka and Appel, 2023), though practical implementation requires further validation.

Overall, considering these contributions, our work advances understanding of how digital discourse shapes social fragmentation while offering pathways for more effective intervention approaches.

Limitations and future research. While this study provides valuable insights into the deployment of rhetorical techniques in social network discussions, several limitations warrant consideration and suggest avenues for future research. (1) Our analysis focused on two controversial topics, which may not capture the full spectrum of online controversies. (2) The research framework prevented us from observing the impact of platform-specific features on users' communication, as well as on the temporal evolution of rhetorical patterns, particularly during periods of heightened social polarization or significant events. (3) Our computational approach, while enabling large-scale analysis, detects only a subset of rhetorical techniques, with varying levels of accuracy across different categories. Techniques that require complex argumentative or contextual understanding show lower detection accuracy. Moreover, relying on GPT-4o-mini introduces interpretive frameworks that actively shape meaning rather than providing objective detection. While our validation shows substantial alignment with expert judgment, the model's cultural and political biases (Hu et al., 2025) may influence classification outcomes in ways that remain difficult to fully explain due to the "black box" nature of LLM decision-making. (4) Our user-level stance aggregation assumes consistent positioning, whereas some users may shift stances within or across discussions—a dynamic our current methodology cannot capture. (5) Our methodology does not adequately distinguish genuinely neutral users (those with no clear position) from users engaging in off-topic discussions, limiting our ability to fully characterize the sophisticated forms of identity work within the Neutral category. This conflation hinders the analysis of how users strategically deploy apparent neutrality as a form of social positioning, distinct from both genuine disengagement and simple off-topic commentary. (6) Our analysis focused exclusively on divisive rhetorical techniques as manifestations of social identity work, without examining how users might perform such identity work through constructive discursive practices, i.e., communicative modalities that preserves epistemic dialogue conditions.

These limitations suggest several directions for future research: (1) analyzing a broader range of controversial topics to validate and expand understanding of rhetorical patterns; (2) deepening the study of how factors such as source characteristics (e.g., YouTuber communication style) and platform features (e.g., algorithmic recommendation patterns) shape users' engagement modalities in time; (3) developing specialized models with greater interpretive transparency, including BERT-based approaches for domain-specific detection; (3a) expanding the analytical framework to encompass the full spectrum of destructive online behaviors, integrating divisive rhetoric with other forms of hostile communication to provide a comprehensive understanding of

digital discourse degradation (Bassi et al., 2025); (4) developing dynamic stance detection methods capable of capturing within-user positional shifts and the contextual factors that trigger them; (5) developing more nuanced computational methods to distinguish types of neutral engagement, including genuine neutrality, off-topic discussions, and strategic meta-positioning—potentially leveraging topic modeling approaches such as BERTopic (Ocal, 2024) to identify off-topic content—to better understand how apparent disengagement functions as a form of social identity work in controversial discussions; (6) investigating how social identity work manifests through constructive discourse in controversial contexts, leveraging the divisive rhetorical strategies framework to distinguish between destructive and constructive comments (Bassi et al., 2025) and develop a more comprehensive understanding of boundary-making modalities in digital controversies.

Data availability

While YouTube policies prevent us from sharing the dataset we created and used, detailed results of all our experiments are available in the supplementary material. We are fully committed to transparency and are happy to provide any additional information readers may require. All the script used to carry out the research are available at the GitHub Repository: https://github.com/BassiDavide/DivisiveRhetoric_YouTube.

Code availability

All the scripts used to carry out the research are available at the GitHub Repository.

Received: 25 January 2025; Accepted: 10 November 2025;
Published online: 30 December 2025

Notes

1 Refers to the capacity of individuals to think critically and independently, forming beliefs based on their own reasoning and evidence, which involves the possibility of appropriately rely on external sources while maintaining intellectual self-direction (Carter, 2020).
2 As opposed to more community-based platforms such as Reddit, where members regularly interact and identify themselves based on pre-established community values (Graham, 2015).
3 In addition to the YouTube API Documentation, in the GitHub Repository we also share the scripts we used for crawling the comments.
4 OpenAI's API enables automated processing of large text corpora by programmatically submitting consistent prompts to the same model underlying the widely accessible ChatGPT interface. Our full implementation code, including prompt engineering and data processing scripts, is available at our GitHub Repository for transparency and reproducibility.
5 For a brief explanation of API functioning see note in section "Stance detection" and OpenAI Website. Also in this case we share the scripts in the GitHub repository.
6 For use examples of the techniques see section "Analysis and results".
7 FR (Frequency Ratio) indicates how many times more frequent the technique is in the more prevalent group. A value of 1 indicates equal frequency between groups, while, for reference, FR = 1.5 indicates the technique is 50% more frequent in one group, FR = 2.0 indicates twice as frequent, and so on.
8 Parent comment text appears for reader context only, but was not included in the LLM prompt.
9 Log Ratio (LR) is calculated as the logarithm base 2 of the proportion between two groups' frequencies. A value of 0 indicates equal frequency in both groups, positive values indicate higher frequency in the first group, and negative values indicate higher frequency in the second group.
10 Cohen's d quantifies the standardized difference between two means, with $|d| \geq 0.2$, 0.5, 0.8 indicating small, medium, and large effects, respectively.
11 While comments may exhibit multiple rhetorical techniques, they are presented as illustrative examples of a specific one.
12 While we acknowledge that this categorical grouping limits our ability to isolate specific mechanisms, our methodology and the consistent patterns support the centrality of issue-based positioning in rhetorical choices. On this, see also sections "Network creation" and "Conclusion".

References

Alturayeif N, Luqman H, Ahmed M (2023) A systematic review of machine learning techniques for stance detection and its applications. Neural Comput Appl 35(7):5113–5144
Andersson M (2021) The climate of climate change: impoliteness as a hallmark of homophily in YouTube comment threads on Greta Thunberg's environmental activism. J Pragmat 178:93–107
Andersson M (2024) E-mpoliteness–creative impoliteness as an expression of digital social capital. J Politeness Res 20(2):227–248
Baggini J (2016) The edge of reason: a rational skeptic in an irrational world. Yale University Press
Bail C (2022) Breaking the social media prism: how to make our platforms less polarizing. In: Breaking the social media prism. Princeton University Press
Bassi D, Dimitrov DI, D'Auria B, Alam F, Hasanain M, Moro C, Orrù L, Turchi GP, Nakov P, Da San Martino G (2025) Annotating the annotators: analysis, insights and modelling from an annotation campaign on persuasion techniques detection. In: Findings of the association for computational linguistics: ACL 2025. Association for Computational Linguistics, Vienna, Austria, pp 17918–17929
Bassi D, Fomsgaard S and Pereira-Fariña M (2024) Decoding persuasion: a survey on ML and NLP methods for the study of online persuasion. Front Commun 9:1457433. https://doi.org/10.3389/fcomm.2024.1457433
Bassi D, Maggini MJ, Vieira R, Pereira-Fariña M (2024) 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) (pp. 146–153). IEEE. https://doi.org/10.1109/SNAMS64316.2024.10883781
Bassi D, Marino EB, Vieira R, Pereira M (2025) 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. Association for Computational Linguistics, pp 46–57
Berger J, Milkman KL (2012) What makes online content viral? J Mark Res 49(2):192–205
Bhatt AM, Goldberg A, Srivastava SB (2022) A language-based method for assessing symbolic boundary maintenance between social groups. Sociol methods Res 51(4):1681–1720
Bjornsgaard K, Dukić S (2023) The media and polarisation in Europe: strategies for local practitioners to address problematic reporting
Blitvich PG-C (2010) The youtubification of politics, impoliteness and polarization. In: Handbook of research on discourse behavior and digital communication: language structures and social interaction. IGI Global, pp 540–563
Blitvich P G-C, Lorenzo-Dus N, Bou-Franch P (2013) Relational work in anonymous, asynchronous communication: A study of (dis)affiliation in YouTube. In Kecskes I, Romero-Trillo J (Ed), Research Trends in Intercultural Pragmatics (pp. 343-366). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9781614513735.343
Bou-Franch P, Blitvich PG-C (2014) Conflict management in massive polylogues: a case study from YouTube. J Pragmat 73:19–36
Brady WJ, Wills JA, Jost JT, Tucker JA, Bavel JJV (2017) Emotion shapes the diffusion of moralized content in social networks. Proc Natl Acad Sci 114(28):7313–7318
Brown AD (2022) Identities in and around organizations: towards an identity work perspective. Hum Relat 75(7):1205–1237
Budzynska K, Reed C (2012) The structure of ad hominem dialogues. In: Computational models of argument. IOS Press, pp 410–421
Canales MK (2000) Othering: toward an understanding of difference. Adv Nurs Sci 22(4):16–31
Carter JA (2020) Intellectual autonomy, epistemic dependence and cognitive enhancement. Synthese 197(7):2937–2961
Chen Z, Berger J (2013) When, why, and how controversy causes conversation. J Consum Res 40(3):580–593
Chung Y-L, Tekiroğlu SS, Tonelli S, Guerini M (2021) Empowering NGOs in countering online hate messages. Online Soc Netw Media 24:100150
Cinelli M, Cresci S, Quattrociocchi W, Tesconi M, Zola P (2022) Coordinated inauthentic behavior and information spreading on Twitter. Decis Support Syst 160:113819
Cinelli M, Pelicon A, Mozetič I, Quattrociocchi W, Novak PK, Zollo F (2021) Dynamics of online hate and misinformation. Sci Rep. 11(1):22083
Comellas JM, Torcal M (2023) Ideological identity, issue-based ideology and bipolar affective polarization in multiparty systems: the cases of Argentina, Chile, Italy, Portugal and Spain. Elect Stud 83:102615
Da San Martino G, Yu S, Barrón-Cedeño A, Petrov R, Nakov P (2019) Fine-grained analysis of propaganda in news articles. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp 5636–5646

de León E, Trilling D (2021) A sadness bias in political news sharing? The role of discrete emotions in the engagement and dissemination of political news on Facebook. Soc Media + Soc 7(4):20563051211059710
Deutsch M (1973) The resolution of conflict: constructive and destructive processes. Yale University Press, New Haven
Dimitrov D, Alam F, Hasanain M, Hasnat A, Silvestri F, Nakov P, Da San Martino G (2024) SemEval-2024 task 4: multilingual detection of persuasion techniques in memes. In: Ojha AK, Doğruöz AS, Tayyar Madabushi H, Da San Martino G, Rosenthal S, Rosá A (eds) Proceedings of the 18th international workshop on semantic evaluation (SemEval-2024). Association for Computational Linguistics, Mexico City, Mexico, pp 2009–2026
Douek E (2021) Governing online speech: from "posts-as-trumps" to proportionality and probability. Colum L Rev 121:759
Drake B, Kiley J (2019) Americans say the nation's political debate has grown more toxic and 'heated' rhetoric could lead to violence. Pew Research Center
Emamgholizadeh H, Nourizade M, Tajbakhsh MS, Hashminezhad M, Esfahani FN (2020) A framework for quantifying controversy of social network debates using attributed networks: biased random walk (BRW). Soc Netw Anal Min 10(1):90
Engesser S, Fawzi N, Larsson AO (2017) Populist online communication: introduction to the special issue. Inform Commun Soc, 20(9):1279–1292. https://doi.org/10.1080/1369118X.2017.1328525
Fairclough N (1999) Global capitalism and critical awareness of language. Lang Aware 8(2):71–83
Falkenberg M, Galeazzi A, Torricelli M, Di Marco N, Larosa F, Sas M, Mekacher A, Pearce W, Zollo F, Quattrociocchi W (2022) Growing polarization around climate change on social media. Nat Clim Chang 12(12):1114–1121
Fariello G, Jemielniak D (2025) The changing language and sentiment of conversations about climate change in Reddit posts over sixteen years. Commun Earth Environ 6(1):3
Flowerday T, Shell DF (2015) Disentangling the effects of interest and choice on learning, engagement, and attitude. Learn Individ Differ. 40:134–140
Friedland LA, Kunelius R (2023) The public sphere and contemporary lifeworld: reconstruction in the context of systemic crises. Commun Theory 33(2-3):153–163
Frimer JA, Skitka LJ (2018) The Montagu principle: incivility decreases politicians' public approval, even with their political base. J Personal Soc Psychol 115(5):845
Frimer JA, Skitka LJ (2020) Americans hold their political leaders to a higher discursive standard than rank-and-file co-partisans. J Exp Soc Psychol 86:103907
Gagliardone I, Gal D, Alves T, Martinez G (2015) Countering online hate speech. Unesco Publishing
Godber A, Origgi G (2023) Telling propaganda from legitimate political persuasion. Episteme 20(3):778–797
Grabill JT, Pigg S (2012) Messy rhetoric: Identity performance as rhetorical agency in online public forums. Rhetor Soc Q 42(2):99–119
Graham SL (2015) Relationality, friendship, and identity in digital communication. In: The Routledge handbook of language and digital communication. Routledge, pp 305–320
Grice HP (1975) Logic and conversation. Syntax Semant 3:43–58
Guadagno RE, Rempala DM, Murphy S, Okdie BM (2013) What makes a video go viral? An analysis of emotional contagion and internet memes. Comput Hum Behav 29(6):2312–2319
Habermas J (2022) Reflections and hypotheses on a further structural transformation of the political public sphere. Theory Cult Soc 39(4):145–171
Hasanain M, Ahmad F, Alam F (2024) Large language models for propaganda span annotation. In: Al-Onaizan Y, Bansal M, Chen Y-N (eds) Findings of the association for computational linguistics: EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, pp 14522–14532
Haugh M, Chang W-LM (2015) Troubles talk, (dis) affiliation and the participation order in Taiwanese-Chinese online discussion boards. In: Participation in public and social media interactions. John Benjamins Publishing Company, pp 99–133
Heltzel G, Laurin K (2024) Why Twitter sometimes rewards what most people disapprove of: the case of cross-party political relations. Psychol Sci 35(9):976–994
Hessel J, Lee L (2019) Something's brewing! Early prediction of controversy-causing posts from discussion features. In Burstein J, Doran C, Solorio T (Eds), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1: Long and Short Papers) (pp. 1648–1659). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1166
Hruschka TM, Appel M (2023) Learning about informal fallacies and the detection of fake news: an experimental intervention. PLoS ONE 18(3):e0283238
Hu T, Kyrychenko Y, Rathje S, Collier N, van der Linden S, Roozenbeek J (2025) Generative language models exhibit social identity biases. Nat Comput Sci 5(1):65–75
Jia Y, Schumann S (2025) Tackling hate speech online: the effect of counter-speech on subsequent bystander behavioral intentions. Cyberpsychol J Psychosoc Res Cybersp 19(1):4
Keusch F (2013) The role of topic interest and topic salience in online panel web surveys. Int J Mark Res 55(1):59–80
Kim Y (2024) Shame on you! How incivility and absence of supporting evidence in likeminded Facebook comments influence evaluations of ingroup members and online political participation. Online Inf Rev 48(3):619–643
Kosmidis S, Theocharis Y (2020) Can social media incivility induce enthusiasm? Evidence from survey experiments. Public Opin Q 84(S1):284–308
Labarre J (2024) French Fox News? Audience-level metrics for the comparative study of news audience hyperpartisanship. J Inf Technol Polit 21(4):510–527
Lamont M, Pendergrass S, Pachucki M (2015) Symbolic boundaries. Int Encycl Soc Behav Sci 2:850–855
Lewandowsky S, Cook J, Ecker U, Albarracin D, Amazeen MA, Kendou P, Lombardi D, Newman E, Pennycook G, Porter E et al. (2020) The debunking handbook 2020. University of Queensland
Locher MA, Bolander B (2015) Humour in microblogging: exploiting linguistic humour strategies for identity construction in two Facebook focus groups. In: Participation in public and social media interactions. John Benjamins Publishing Company, pp 135–155
Locher MA, Graham SL (2010) 1. Introduction to Interpersonal Pragmatics. In: Locher MA, Graham SL (eds) Interpersonal pragmatics. De Gruyter Mouton, Berlin, New York, pp 1–16
Mao R, Ge M, Han S, Li W, He K, Zhu L, Cambria E (2025) A survey on pragmatic processing techniques. Inf Fusion 114:102712
Martin J (2013) Chantal Mouffe: hegemony, radical democracy, and the political. Routledge
Mason L (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. Am j poli sci 59(1):128-145
McCoy J, Somer M (2019) Toward a theory of pernicious polarization and how it harms democracies: comparative evidence and possible remedies. Ann Am Acad Political Soc Sci 681(1):234–271
McDermott R (2020) Leadership and the strategic emotional manipulation of political identity: an evolutionary perspective. Leadersh Q 31(2):101275
McGlashan M (2020) Collective identity and discourse practice in the followership of the Football Lads Alliance on Twitter. Discourse Soc 31(3):307–328
Mercadante EJ, Tracy JL, Götz FM (2023) Greed communication predicts the approval and reach of US senators' tweets. Proc Natl Acad Sci 120(11):e2218680120
Mercier H (2020) Not born yesterday: the science of who we trust and what we believe. Princeton University Press
Monti C, Aiello LM, De Francisci Morales G, Bonchi F (2022) The language of opinion change on social media under the lens of communicative action. Sci Rep. 12(1):17920
Morales E, Hodson J, O'Meara V, Gruzd A, Mai P (2025) Online toxic speech as positioning acts: Hate as discursive mechanisms for othering and belonging. New Media & Society. https://doi.org/10.1177/14614448251338493
Mudde C, Kaltwasser CR (2017) Populism: A very short introduction. Oxford University Press. https://doi.org/10.1093/actrade/9780190234874.001.0001
Nannini L, Bonel E, Bassi D, Maggini MJ (2025) Beyond phase-in: assessing impacts on disinformation of the EU Digital Services Act AI Ethics 5(2):1241–1269
Narayanan A (2023) Understanding social media recommendation algorithms. Knight First Amendment Institute at Columbia University
Niemi JI (2005) Jürgen Habermas's theory of communicative rationality: the foundational distinction between communicative and strategic action. Soc Theory Pr 31(4):513–532
Ocal A (2024) Perceptions of the future of artificial intelligence on social media: a topic modeling and sentiment analysis approach. IEEE Access 12:182386–182409
Ollerenshaw T, Jardina A (2023) The asymmetric polarization of immigration opinion in the United States. Public Opin Q 87(4):1038–1053
Piskorski J, Stefanovitch N, Nikolaidis N, Da San Martino G, Nakov P (2023) Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st annual meeting of the association for computational linguistics, vol 1: long papers. Association for Computational Linguistics, Toronto, Canada, pp 3001–3022
Puig Larrauri H, Morrison M (2022) Understanding digital conflict drivers. In: Fundamental challenges to global peace and security: the future of humanity. Springer, pp 169–200
Ravndal JA (2018) Explaining right-wing terrorism and violence in Western Europe: grievances, opportunities and polarisation. Eur J Polit Res 57(4):845–866

Rega R, Marchetti R (2021) The strategic use of incivility in contemporary politics. The case of the 2018 Italian general election on Facebook. Commun Rev 24(2):107–132
Rho EHR, Mark G, Mazmanian M (2018) Fostering civil discourse online: linguistic behavior in comments of MeToo articles across political perspectives. Proc ACM Hum Comput Interact 2(CSCW):1–28
Rieder B, Matamoros-Fernández A, Coromina Ò (2018) From ranking algorithms to 'ranking cultures' investigating the modulation of visibility in YouTube search results. Convergence 24(1):50–68
Saha P, Garimella K, Kalyan NK, Pandey SK, Meher PM, Mathew B, Mukherjee A (2023) On the rise of fear speech in online social media. Proc Natl Acad Sci 120(11):e2212270120
Saha P, Mathew B, Garimella K, Mukherjee A (2021) "Short is the road that leads from fear to hate": fear speech in Indian WhatsApp groups. In: Proceedings of the web conference 2021, WWW '21. Association for Computing Machinery, New York, NY, USA, pp 1110–1121
Sciortino G (2024) Borders and boundaries. In: Research handbook on the sociology of migration. Edward Elgar Publishing, pp 23–33
Sprenkamp K, Jones DG, Zavolokina L (2023) Large language models for propaganda detection. *arXiv preprint* https://doi.org/10.48550/arXiv.2310.06422
Storbeck J, Clore GL (2008) Affective arousal as information: how affective arousal influences judgments, learning, and memory. Soc Personal Psychol compass 2(5):1824–1843
Stray J, Iyer R, Puig Larrauri H (2023) The algorithmic management of polarization and violence on social media. Knight First Amendment Institute at Columbia University
Strickler Y, Rao V, Appleton M, Limberg P, Fox R, Citarella J, Benkhadda L, Bear AR, Busta C, Wadsworth J (2024) The dark forest: anthology of the internet. Dark Forest Collective, Berlin
Su LY-F, Xenos MA, Rose KM, Wirz C, Scheufele DA, Brossard D (2018) Uncivil and personal? comparing patterns of incivility in comments on the facebook pages of news outlets. N. Media Soc 20(10):3678–3699
Tajfel H, Turner J (2001) An integrative theory of intergroup conflict. In Hogg MA, Abrams D (Eds), Intergroup relations: Essential readings (pp. 94–109). Psychology Press
Tomasello M (2010) Origins of human communication. MIT press
Törnberg P (2022) How digital media drive affective polarization through partisan sorting. Proc Natl Acad Sci 119(42):e2207159119
Turchi GP, Bassi D, Cavarzan M, Camellini T, Moro C, Orrù L (2023) Intervening on global emergencies: the value of human interactions for people's health. Behav Sci 13(9):735
Wolter JS, Bacile TJ, Xu P (2023) How online incivility affects consumer engagement behavior on brands' social media. J Serv Res 26(1):103–119
Yu X, Wojcieszak M, Casas A (2024) Partisanship on social media: In-party love among American politicians, greater engagement with out-party hate among ordinary users. Polit Behav 46(2):799–824
Zarrinkalam F, Noughabi HA, Noorian Z, Fani H, Bagheri E (2024) Predicting users' future interests on social networks: a reference framework. Inf Process Manag 61(5):103765
Zeitzoff T (2023) Nasty politics: the logic of insults, threats, and incitement. Oxford University Press
Zhang M (2023) Digital lifeworld and communicative interaction: conceptualizing the transformative potentials of social networking in the public sphere. J Linguist Commun Stud 2(4):121–131
Zhong L, Cao J, Sheng Q, Guo J, Wang Z (2020). Integrating semantic and structural information with graph convolutional network for controversy detection. In Jurafsky D, Chai J, Schluter N, Tetreault J (Eds), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 515–526). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.49
Zompetti J (2015) Divisive discourse. Cognella Academic Publishing, Illinois State University

Acknowledgements

This work is supported by the EUHORIZON2021 European Union's Horizon Europe research and innovation programme (https://cordis.europa.eu/project/id/101073351/es) the Marie Skłodowska-Curie Grant No.: 101073351. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Author contributions

DB: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Writing—original draft, Writing—review and editing; GDSM: Conceptualization, Investigation, Writing—review and editing, Supervision; RV: Supervision, Writing—review and editing; MPF: Conceptualization, Investigation, Supervision, Writing—review and editing.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

This research was reviewed by the Ethics Committee of the University of Santiago de Compostela (USC), which granted an exemption from full ethics review the 22 October 2025. The exemption was granted because the study analyzes publicly available YouTube comments without collecting personally identifiable data, involves no direct interaction with human participants, examines content only in aggregate form, and ensures individual identification is technically impossible.

Informed consent

Informed consent was not obtained as the study analyzed publicly posted YouTube comments that users voluntarily shared on a public platform accessible without restrictions. All data was collected in aggregated form with no usernames, profile information, or personally identifiable information retained or reported.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-025-06277-7.

Correspondence and requests for materials should be addressed to Davide Bassi.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025