Harvard Kennedy School Misinformation Review[1]
August 2025, Volume 6, Issue 4

DOI: https://doi.org/10.37016/mr-2020-181
Website: https://misinforeview.hks.harvard.edu/

Research Article

Toxic politics and TikTok engagement in the 2024 U.S. election

What kinds of political content thrive on TikTok during an election year? Our analysis of 51,680 political videos from the 2024 U.S. presidential cycle reveals that toxic and partisan content consistently attracts more user engagement—despite ongoing moderation efforts. Posts about immigration and election fraud, in particular, draw high levels of toxicity and attention. While Republican-leaning videos tend to reach more viewers, Democratic-leaning ones generate more active interactions like comments and shares. As TikTok becomes an important news source for many young voters, these patterns raise questions about how algorithmic curation might amplify divisive narratives and reshape political discourse.

Authors: Ahana Biswas (1), Alireza Javadian Sabet (1), Yu-Ru Lin (1)
Affiliations: (1) Department of Informatics and Networked Systems, University of Pittsburgh, USA
How to cite: Biswas, A., Javadian Sabet, A., & Lin, Y.-R. (2025). Toxic politics and TikTok engagement in the 2024 U.S. election. *Harvard Kennedy School (HKS) Misinformation Review*, *6*(4).
Received: March 12th, 2025. Accepted: July 9th, 2025. Published: August 20th, 2025.

Research questions

- How does the political content of TikTok videos affect engagement levels? More specifically, how do the number of videos posted, viewed, and interacted with vary across the leanings of political content?
- How does user interaction with political TikTok videos differ across various content characteristics, such as the presence of toxic content and other features?
- How common is toxic content across different political topics, and how does toxic content in these topics relate to user engagement?
- What patterns emerge in user engagement with toxic content during major political events?

Essay summary

- This study offers one of the first empirical examinations of how partisanship, political toxicity, and topical focus—such as immigration, racism, and election fraud—shaped user engagement with TikTok videos during the 2024 U.S. presidential election.
- The analysis was drawn from 51,680 political videos on TikTok, using models adjusted for feed ranking, user behavior (author, music, posting time), and platform engagement metrics.

[1] *A publication of the Shorenstein Center on Media, Politics and Public Policy at Harvard University, John F. Kennedy School of Government.*

- The majority of videos analyzed (77%) were explicitly partisan and were associated with approximately twice the engagement of nonpartisan content. Republican-leaning videos received more views, while Democratic-leaning ones showed more interactions—measured by total likes, comments, and shares.
- Toxic videos were associated with 2.3% more interactions. Partisan content also tended to show higher engagement, with Democratic-leaning toxic videos linked to even higher interactions.
- Racism, antisemitism, and election fraud were among the most toxic topics, with toxicity defined as rude or disrespectful language. Toxic videos on elections (+1.3%) and immigration (+3.5%) received higher engagement.
- Toxicity and engagement levels changed after major political events. Following Trump's conviction, videos with severe toxicity and sexual attacks saw an approximate 2% surge in interactions.
- Captions alone were weak predictors of partisanship and toxicity, but transcripts of the audio from the videos improved alignment with manual labels (68.2%) and had significantly more (56.2%) toxic content. These results highlight the limitations of surface-level text features and the need for multimodal analysis in political content detection.

Implications

This study examines how political toxicity—such as insults, threats, and harassment—manifests and gains engagement on TikTok, a recommendation-driven platform especially popular among users under 30 (Leppert, & Matsa, 2024; TikTok, 2025e). While toxicity is often treated separately from misinformation and disinformation, recent literature shows that these forms of harmful content frequently overlap, particularly in polarized or conflict settings (Mosleh et al., 2024; Wardle, 2024). False or misleading narratives often co-occur with toxic or identity-targeted language (Ferrara et al., 2020; Marwick & Lewis, 2017). Many topics in our analysis—for example, racism, immigration, and election fraud—are frequently linked to misinformation campaigns, particularly during elections or crises (Donovan & Boyd, 2021; Guess et al., 2018; Neidhardt & Butcher, 2022). While misinformation and disinformation can intensify hostility, rising toxicity may signal deeper engagement with conspiratorial or extremist narratives (Bennett & Livingston, 2018; Meleagrou-Hitchens & Kaderbhai, 2017; Wardle & Derakhshan, 2017). Our findings reveal that engagement with political toxicity increases during politically sensitive events, raising concerns about how algorithmic systems may amplify such content, echoing similar concerns raised in recent work on platform visibility dynamics (Biswas, Lin, et al., 2025). This study contributes to ongoing discussions on platform governance and informs strategies for responsibly addressing harmful political content.

We focused on nuanced political content—ranging from broadly political discourse to explicitly partisan posts—and how user interaction varies based on both toxicity and partisan alignment of the content. TikTok's algorithmically curated feed presents a unique context in which content visibility is shaped not just by popularity but by engagement-driven ranking systems. Thus, what users encounter is filtered through algorithmic choices, not solely their preferences or followings. We also focused on user engagement with this political content, not merely passive exposure (views) but also active interactions such as likes, comments, and shares—forms of interaction that help us understand what types of political content resonate with users. Our analysis shows that toxic content draws higher engagement when tied to high-profile issues like labor rights, socio-cultural controversies, or major political events. Moreover, politically aligned content with toxic framing tends to receive more interaction than non-partisan toxic content, indicating patterns of differential responsiveness likely influenced by algorithmic sorting.

This raises important questions about the platform's role in amplifying divisive content: moderation practices and content delivery algorithms are inseparable, and accountability for one cannot be meaningfully separated from the other. *Platform accountability*—the responsibility to explain, justify, and refine how their systems govern content exposure—must address not only enforcement (what gets removed) but also amplification (what gets promoted), particularly when these systems mediate access to political expression.

While TikTok has stated its commitment to balancing freedom of expression and community safety, its moderation policies and enforcement processes remain opaque (TikTok, 2024; TikTok, 2025c; TikTok, 2025d). Although the platform uses a combination of human review and machine learning for moderation, there is limited public documentation on how these systems function, particularly in the political domain. The observed association between toxicity and partisan content engagement suggests that moderation processes may not be uniformly sensitive to context, though we caution against overgeneralizing these effects. Rather than implying moderation failure, these patterns point to the complexity of content governance in politically charged environments.

Moreover, TikTok's current data-sharing infrastructure, particularly its Research API (TikTok, 2025a), presents significant barriers to external auditing. The platform restricts access to historical content and disallows targeted user queries, limiting researchers' ability to trace moderation decisions over time or across demographics and making it difficult to assess whether moderation practices are equitable or effective. Addressing this would require not only greater data access but also disaggregated transparency reporting, including enforcement actions by topic, timeframe, and moderation rationale. Procedural clarity and the opportunity to appeal can help ensure that moderation does not feel arbitrary or politicized.

We found that signals from full video transcripts were more predictive of toxicity and political alignment than simply using captions or other content descriptors (e.g., hashtags). This is particularly relevant for political TikToks, which often rely on verbal arguments delivered in visually minimal formats. While TikTok has stated that it employs multimodal moderation techniques—including audio and visual processing (TikTok, 2025b)—public documentation is limited. Given the rhetorical characteristics of political videos, we argue that moderation strategies should be tailored to the dominant communicative modality. For political content, transcript-based analysis may yield more reliable signals than visual classifiers or keyword filters.

These accountability mechanisms should be embedded not just in daily moderation operations but also in the platform's crisis response protocols. Our study shows that engagement with toxic content spikes around political flashpoints, such as Trump's conviction, mirrors well-documented patterns in crisis-driven disinformation surges (Pierri et al., 2023; Starbird, 2017). These moments of heightened uncertainty and collective attention are precisely when platforms are most vulnerable to coordinated influence efforts and viral toxicity. We, therefore, echo calls for event-triggered moderation protocols (Goldstein et al., 2023), enabling platforms to promptly curb the spread of harmful political content during sensitive periods such as elections, protests, or geopolitical conflict. Platforms have implemented such measures in the past—for example, TikTok's removal of Capitol riot-related hashtags, or Facebook's pause on political ads during election periods (Paul, 2020; Perez, 2021)—but these interventions are often ad hoc. These mechanisms should be designed to prevent overreach while enabling human oversight and third-party auditability (Koshiyama et al., 2024).

If platforms like TikTok continue to supply toxic and partisan content—particularly around charged sociopolitical issues—there is a risk of normalizing incivility and entrenching polarization. This is especially concerning given the platform's influence over users under 30 (Karimi & Fox, 2023), many of whom rely on it as a source of political news (McClain, 2023; Siegel-Stechler et al., 2025). These patterns also create challenges for content creators who produce nuanced commentary, as they may struggle to compete with

more provocative content optimized for engagement. Greater platform transparency (Balasubramaniam et al., 2022; Diab, 2024; Felzmann et al., 2020) could empower educators, researchers, and civil organizations to use legal yet norm-challenging content as tools for media literacy and develop critical awareness of how algorithmic systems shape political understanding.

Findings

Finding 1: Partisan videos garner higher engagement.

Figure 1A shows the dominance of partisan over non-partisan videos on TikTok in the overall share of posts, views, and interactions. A Mann-Whitney U[2] test showed that partisan videos receive significantly more views (U = 144089417.0, r = 0.08, 95% CI = 0.07, 0.09, p < .001) and interactions (U = 140706446.5, r = 0.10, 95% CI = 0.08, 0.11, p < .001) than non-partisan ones.

Figure 1B presents total posts, median views, and interactions per post across political leanings. Partisan videos receive nearly 2.2 times more median views and interactions compared to non-partisan videos. Republican-leaning videos receive slightly more views than Democratic-leaning content on average (Med_{Repub} = 4,428, Med_{Dem} = 4,359); however, Democratic-leaning videos have higher interactions per post (Med_{Repub} = 664, Med_{Dem} = 739). This suggests that while Republican content reaches a wider audience, Democratic content may foster more interaction and discussion.

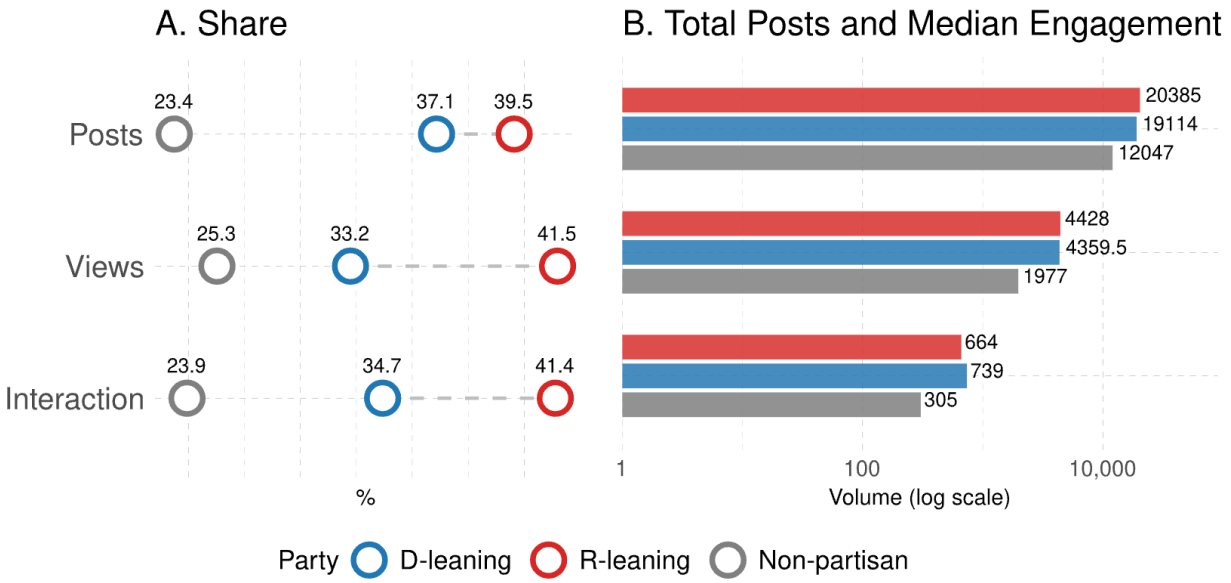

Figure 1. Distribution of engagement in political TikTok posts. *Panel A shows the share of posts, views, and interactions for non-partisan, Republican-leaning, and Democratic-leaning content. Partisan content has higher engagement shares compared to non-partisan, with Republican-leaning videos having the highest engagement share on the platform. Panel B shows the total volume of posts and median views and interactions.*

[2] Mann–Whitney U test: A non-parametric statistical test used to determine whether two groups differ in their distribution. It is an alternative to the *t*-test when data do not meet the assumptions of normality.

Finding 2: Toxicity is linked to higher interaction; partisan toxicity has a stronger association than non-partisan toxicity.

To examine whether toxic language drives user engagement, we used linear mixed-effects regression models[3] that account for platform-driven exposure biases. Adjusting for these sources of bias helps reduce confounding from factors that influence content visibility and engagement. Specifically, we included random effects for post author, featured music, and posting time to control for algorithmic amplification and habitual engagement patterns (e.g., highly followed users or trending sounds).

We found that videos containing toxic language received 2.3% (or $b = 0.023$)[4] more interaction than non-toxic ones (95% CI = 0.017, 0.028, $p < .001$). The effect is stronger when toxicity appears in partisan content: toxic partisan videos received significantly more interaction than nonpartisan ones ($b = -0.014$, 95% CI = -0.022, -0.007, $p < .001$), with Democratic-leaning toxic posts slightly outperforming Republican-leaning ones ($b = -0.006$, 95% CI = -0.013, 0.000, $p < .10$).

Beyond toxicity, we found that affective and demographic features also influence engagement. Videos with stronger red hues[5] received 0.7% more interaction ($b = 0.007$, 95% CI = 0.004, 0.010, $p < .001$), and longer videos gained 1.0% more interaction ($b = 0.010$, 95% CI = 0.006, 0.014, $p < .001$). Videos featuring older speakers were also associated with slightly higher engagement ($b = 0.005$, 95% CI = 0.003, 0.008, $p < .001$).

In contrast, hedging language[6] was associated with lower interaction ($b = -0.016$, 95% CI = -0.025, -0.008, $p < .001$). Compared to Democratic-leaning videos, nonpartisan and Republican-leaning content received 4.0% (95% CI = -0.049, -0.031, $p < .001$) and 0.7% (95% CI = -0.014, 0.000, $p < .05$) less interaction, respectively. These findings suggest that rhetorical certainty, visual salience, speaker characteristics, and political framing all shape user engagement on TikTok.

[3] A type of regression model that accounts for both fixed effects (variables of interest like toxicity or topic) and random effects (unobserved variations across clusters such as users or videos). This allows for more accurate estimates when data is grouped or hierarchical.

[4] The regression estimate, or effect size (b), is a quantitative measure of the strength or magnitude of a relationship between two variables. For instance, videos containing toxic language receive 2.3% (or $b = 0.023$) more interaction. This means that, on average, a toxic video receives 2.3% more interactions than a similar non-toxic video, after accounting for topic, political leaning, user-level effects, and views.

[5] Red hue was measured as the average saturation of red pixels across video frames—a visual cue associated with emotional intensity or urgency.

[6] Hedging language refers to linguistic expressions that convey uncertainty or soften claims, such as "it seems," "perhaps," or "I believe."

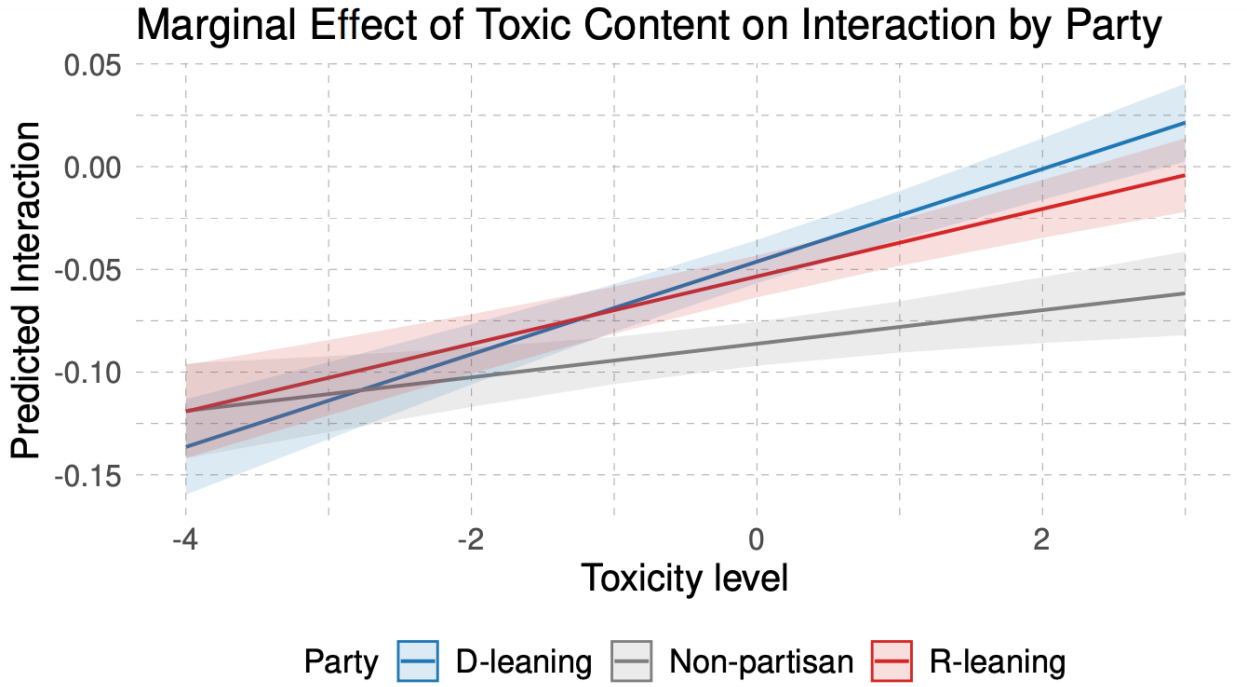

Figure 2. Marginal effects of toxicity on predicted interaction levels across political leanings, based on mixed-effects regression models controlling for views. *Toxicity has a positive effect on interaction for all political leanings, with partisan-aligned posts showing stronger associations than non-partisan.*

Finding 3: Associations between toxicity and engagement vary by topic and partisanship.

To understand how toxicity and engagement interact across political themes, we categorized each video into one or more of 22 manually validated political topics (see Appendix C), including salient issues like immigration, racism, antisemitism, and election fraud. We found that toxicity levels were highest in topics such as racism, antisemitism, Nazi references, election fraud, and Trump's assassination attempt (see Figure D1, Appendix D). Toxic content was associated with higher user interaction in videos about elections (b = 0.013, 95% CI = 0.000, 0.026, p < .05) and immigration (b = 0.035, 95% CI = -0.004, 0.074, p < .10) (Figure 4). In Republican-leaning videos, toxic geopolitical content (e.g., discussions of international conflict) showed a stronger engagement effect than nonpartisan equivalents (b = 0.050, 95% CI = 0.005, 0.095, p < .05) (see Figure 5).

We also examined how topics themselves—not just their toxicity—shaped engagement. Topics related to social and cultural issues (b = 0.034, 95% CI = 0.020, 0.047, p < .001), political figures and events (b = 0.012, 95% CI = -0.001, 0.025, p < 0.1), and labor (b = 0.032, 95% CI = 0.002, 0.062, p < .05) were all associated with significantly higher interaction (Figure 3). By contrast, immigration was associated with lower interaction overall (b = -0.029, 95% CI = -0.058, 0.000, p < .05), except for Republican-leaning immigration posts, which saw higher engagement (b = 0.047, 95% CI = 0.012, 0.083, p < .01). Republican-leaning videos on labor issues also saw significantly lower engagement (b = -0.085, 95% CI = -0.134, -0.036, p < .001) (see Figure 4).

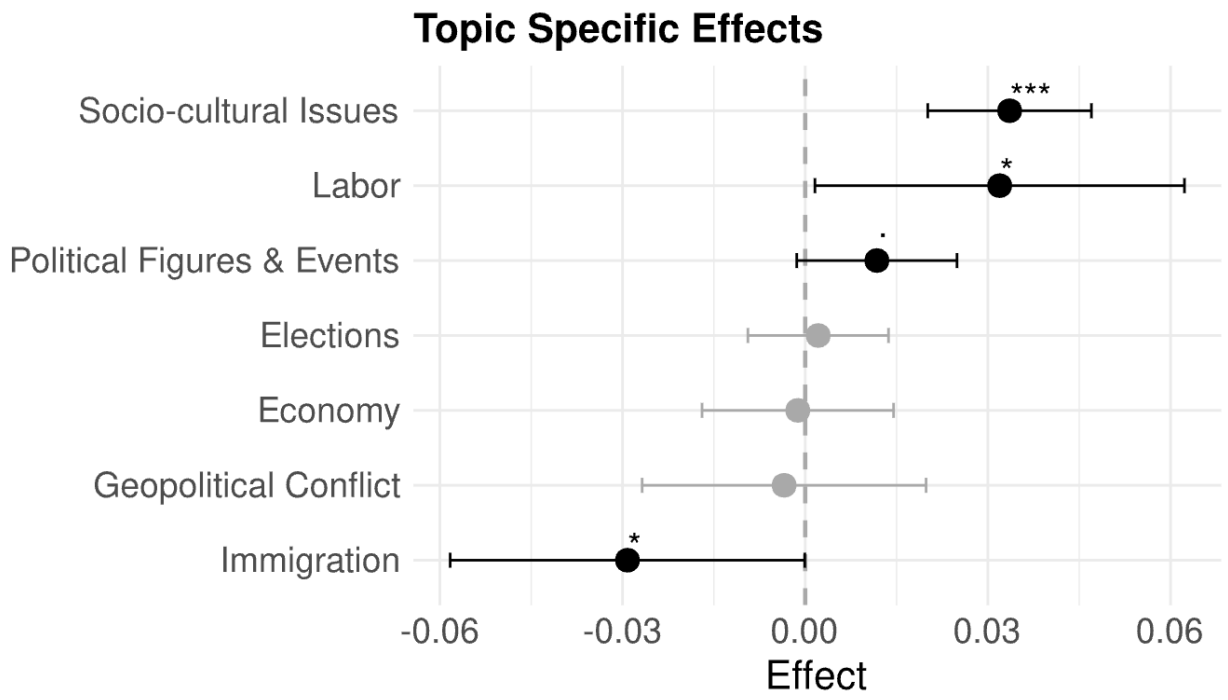

Figure 3. Effects of political topics on user engagement. *User interaction varies by the primary political topic of each video. Topics such as labor rights, political figures, and socio-cultural issues are associated with higher engagement, while immigration shows negative associations.*

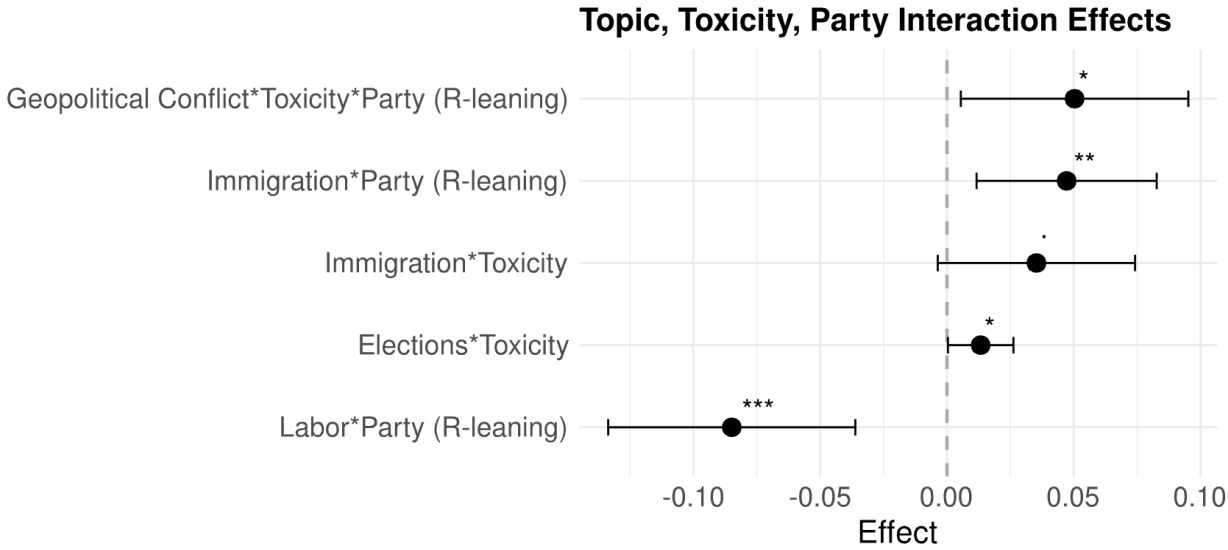

Figure 4. Interaction effects between toxicity, political topics, and party. *The relationship between toxicity, political topics and engagement differs, influenced by partisan leaning. For example, the coefficient for "Toxicity*Election" captures the added effect of toxicity in election-related posts, beyond the average effect of toxicity alone.*

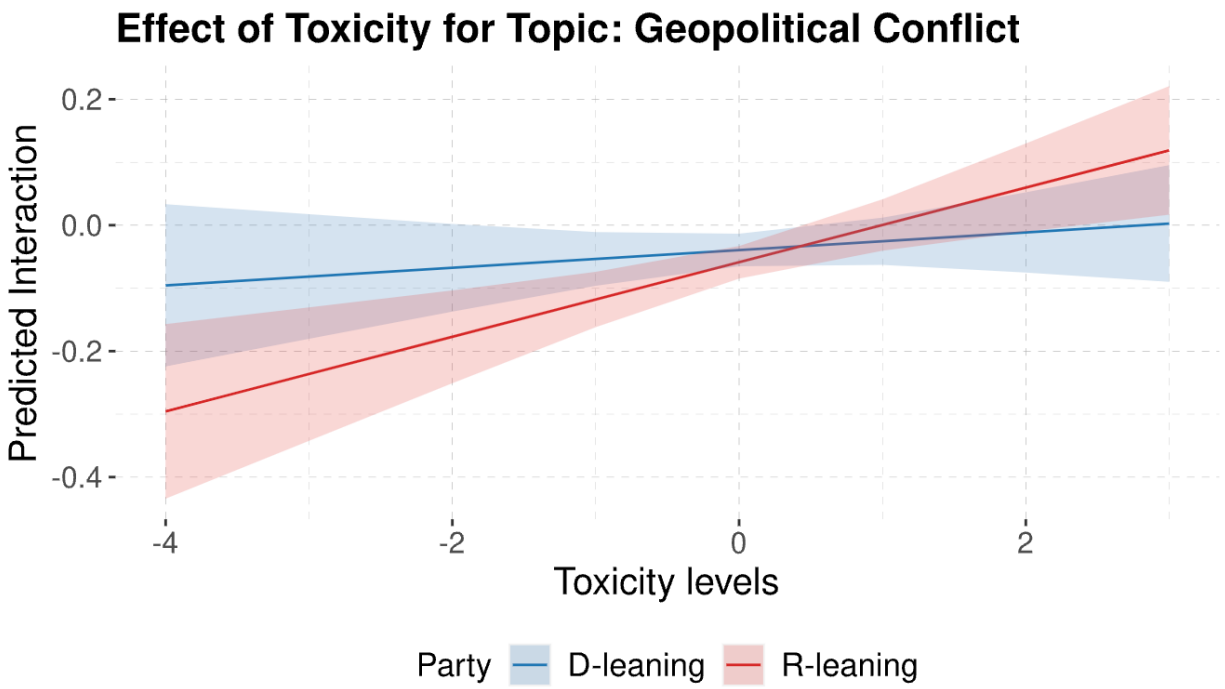

Figure 5. Interaction effects of toxicity and partisan alignment in geopolitical content. *Toxicity interacts with partisan alignment in shaping engagement on geopolitical topics. Specifically, Republican-leaning videos that contain toxic language about international conflicts are associated with significantly higher levels of user interaction than comparable Democratic content.*

Finding 4: Toxic content dynamics shift following political events.

Toxic content notably changed after major political events—such as Trump's conviction, the first presidential debate between Biden and Trump, the Republican National Convention (RNC), the Harris campaign announcement, and Democratic National Convention (DNC)—as shown in Figure 6. We found that overall toxicity (U = 68401.0, r = -0.22, 95% CI = -0.31, -0.14, p < .001), sexual toxicity (U = 62777.0, r = -0.11, 95% CI = -0.21, -0.02, p < .05), and severe toxicity (U = 66374.0, r = -0.21, 95% CI = -0.29, -0.12, p < .001) decreased in Democrat-leaning videos, while identity attacks (U = 69685.0, r = 0.18, 95% CI = 0.10, 0.26, p < .001) increased in Republican-leaning videos after the Harris campaign announcement. Political videos with sexual attacks and severe toxic language saw increased user interactions by 2.0% (95% CI = 0.004, 0.035, p < .05) and 1.6% (95% CI = 0.001, 0.032, p < .05), respectively, following the Trump conviction controversy, as shown in Figure 7.

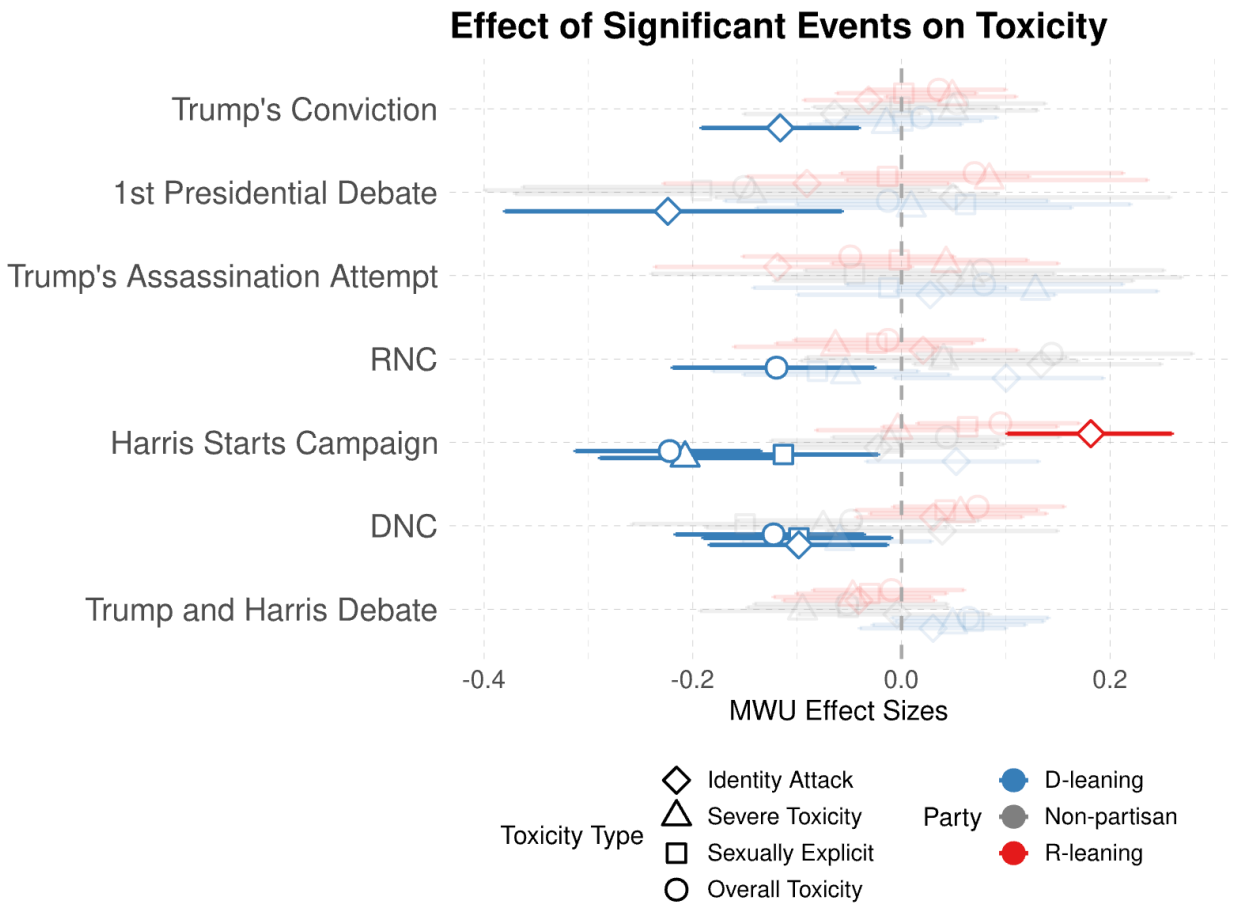

Figure 6. Changes in toxicity types following major political events. *Toxic language—including sexual toxicity, identity attacks, and severe toxicity—shifted following key political events such as Trump's conviction and Harris's campaign launch, suggesting that political flashpoints coincide with changes in hostile rhetoric.*

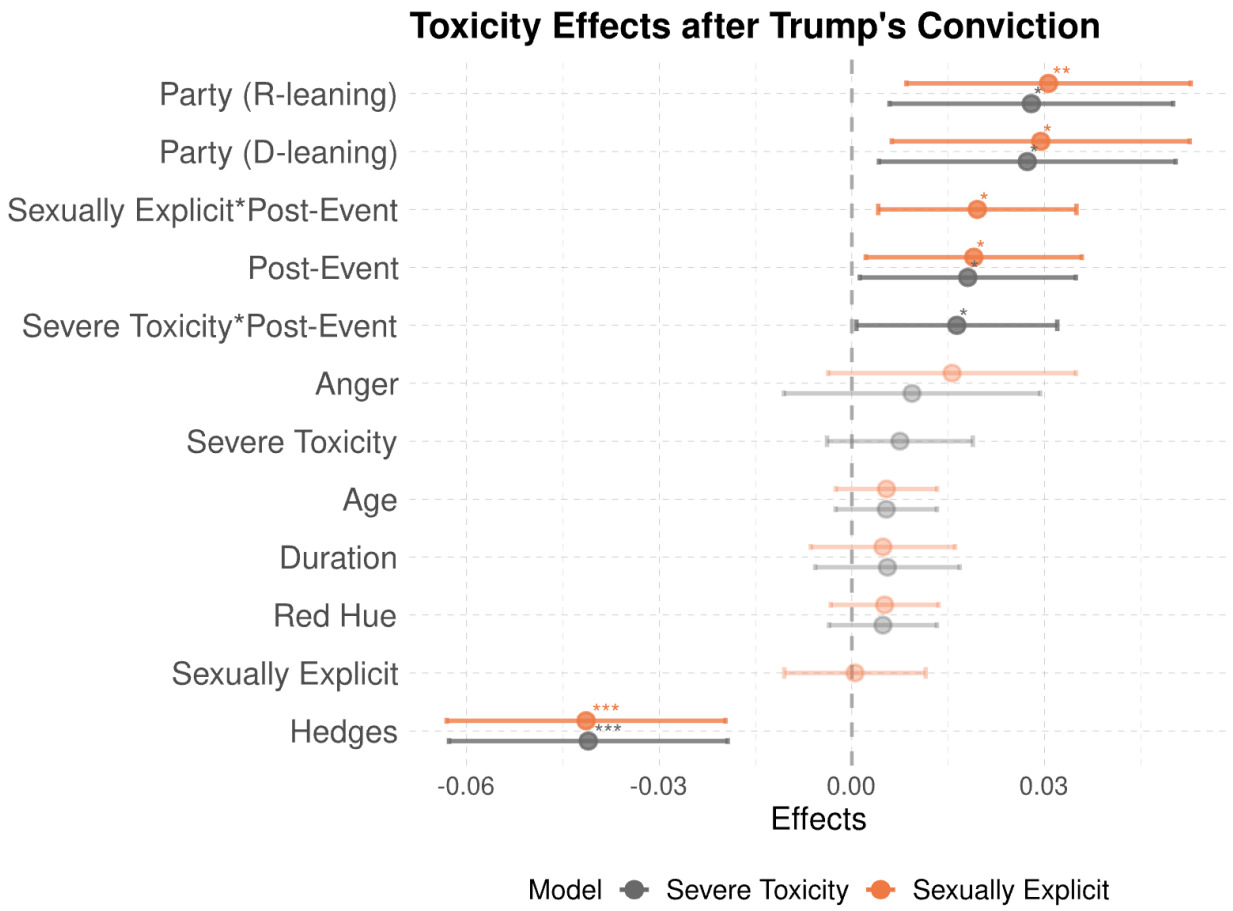

Figure 7. Effects of toxicity types on engagement after major political events. *This figure shows the effect of different toxicity types on predicted user interaction with political videos before and after Trump's conviction. Sexually explicit language and severe toxicity are associated with significantly higher engagement in the post-event period, controlling for views and other covariates.*

Finding 5: Shallow textual features alone fail to detect partisanship and toxicity.

To evaluate how well different inputs capture political content signals, we compared the performance of caption-based inputs (i.e., short text metadata) with transcript-based inputs (full spoken content extracted from videos). Specifically, we assessed whether models using full transcripts were better at identifying toxicity and political alignment than those using captions alone.

For partisan alignment detection[7], using captions, we achieved a Cohen's Kappa of 0.44 with human-coded labels, while using transcripts improved agreement to 0.74—indicating substantially better performance. For toxicity detection, using transcript, we identified[8] higher toxicity levels in 56.2% more cases compared to caption-based models (Cliff's δ[9] = 0.12, 95% CI = 0.04, 0.21). This improvement was especially pronounced for toxicity subtypes like sexual toxicity, identity attacks, and severe toxicity, increasing by 60.4% (Cliff's δ = 0.21, 95% CI = 0.12, 0.29), 62.9% (Cliff's δ = 0.26, 95% CI = 0.17, 0.34) and 72.6% (Cliff's δ = 0.45, 95% CI = 0.37, 0.52) respectively (see Figure E1, Appendix E).

[7] Partisan leaning was identified using Mistral-7B-Instruct-v0.3 model as described in Appendix B.
[8] Perspective API was used to detect toxicity, please see Appendix H for details on manual validation.
[9] Cliff's δ is a non-parametric effect size measure that quantifies the degree of difference between two groups. It ranges from -1 to 1, where 0 indicates no difference, and values closer to -1 or 1 indicate stronger group differences.

These findings suggest that models relying only on shallow metadata, such as captions, risk systematically missing toxic or partisan cues that are more clearly expressed in spoken language. Our results highlight the importance of incorporating multimodal content signals—especially full transcripts—for accurate political content analysis on platforms like TikTok.

Methods

Data collection and feature extraction

We used the TikTok Research API to collect video posts in three waves. Our process began with identifying politically active users (or seed users) who posted at least three videos between January and March 2024 using U.S. election-related hashtags (e.g., #election2024, #biden2024, #trump2024, #maga; see full list in Appendix A). We then applied a snowball sampling strategy, retrieving up to 1,000 liked videos per user, and iteratively expanding the sample of seed users by including authors of these liked videos. Our data collection period ranged from January 01, 2024, until the 2024 U.S. presidential election on November 7, 2024 (see Appendix A for details). This yielded a dataset of 51,680 downloadable political video posts created by 15,344 unique users.

We extracted video transcripts using Whisper (Radford et al., 2023), enabling the analysis of full spoken content. We classified the partisan alignment of videos based on the transcripts, using the Mistral-7B-Instruct-v0.3 model (Mistral AI, 2024). This yielded 20,385 Republican (R) leaning, 19,114 Democratic (D) leaning, and 12,047 non-partisan posts, which we further validated by manual annotation of 150 samples (see Appendix B).

To examine how language and visuals influence engagement, we extracted a combination of linguistic, visual, and structural features from each video (Appendix A). Visual features included speaker demographics—age, gender—identified using DeepFace (Serengil & Özpınar, 2024) and verified through manual annotation (see Appendix G). We also extracted facial expressions (e.g., joy, anger), average red hue as a measure of color saturation, and technical video properties such as duration and frames per second (FPS). Linguistic features derived from transcripts included markers of generalization, causation, hedging (e.g., "I think," "maybe"), subjectivity, and emotion words, based on lexicons and prior NLP work (Appendix A). Toxicity was assessed using Perspective API (Jigsaw, 2025), which detects general incivility as well as specific forms such as identity attacks, sexually explicit language, and severe toxicity. We adopted Perspective's definition of toxicity as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion," and validated this output through manual annotation, achieving Cohen's Kappa = 0.79 (see Appendix H). Finally, we used a hybrid method combining keyword filtering and semantic clustering to identify politically salient video topics such as immigration, racism, and labor rights. Topic labels were finalized through manual review (see Appendix C). By integrating these diverse modalities—spoken text, visual signals, and structural attributes—our approach provides a robust framework for analyzing how TikTok videos shape political discourse and drive user engagement.

Significant events

We focused our analysis on several significant events during the U.S. presidential election cycle, as these moments had a substantial impact on political discourse and public opinion. These events include Trump's conviction on 34 felony counts, the first presidential debate between Biden and Trump, the assassination attempt on Trump, the date Kamala Harris officially launched her presidential campaign after Joe Biden's resignation, and the debate between Kamala Harris and Donald Trump following her entry into the race.

Additionally, we examined the Democratic National Convention (DNC) and the Republican National Convention (RNC), which spanned multiple days and served as critical moments for each party to energize their base and present their platforms.

Statistical analyses

We conducted Mann-Whitney U (MWU) tests with Benjamini false discovery rate[10] (FDR) correction to compare differences in views and interactions (total number of likes, comments, and shares on a video) for partisan and non-partisan videos (RQ1). For research questions 2–4 (see Appendix F for model specifications and full regression results), we used linear mixed effect models with random effects on the user, post timing (i.e., day posted), and music effects in the video—to account for the hierarchical data structure and unobserved heterogeneity, such as variations in user behavior, content trends, or time-specific factors influencing engagement.

We assessed the impact of toxicity and how it varied by partisan leaning while accounting for other significant predictors of interactions (RQ2). To understand the effect of political topics, the model was extended by introducing interaction effects between toxicity, party affiliation, and topic groups (RQ3). For RQ4, we evaluate the impact of significant events by comparing changes in overall toxicity as well as its subtypes in a week before versus after each event (i.e., all videos in a window of seven days pre- vs. post-event) using MWU tests with FDR corrections. Furthermore, we evaluate how toxicity is associated with interactions after three of the most significant political events: Trump's conviction, Trump's assassination attempt, and Harris' campaign announcement.

Bibliography

Balasubramaniam, N., Kauppinen, M., Hiekkanen, K., & Kujala, S. (2022). Transparency and explainability of AI systems: Ethical guidelines in practice. In V. Gervasi & A. Vogelsang (Eds.), *28th International Working Conference, REFSQ 2022, Birmingham, UK, March 21–24, 2022, Proceedings* (pp. 3–18). Springer. https://doi.org/10.1007/978-3-030-98464-9_1

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European journal of communication*, *33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Biswas, A., Lin, Y.-R., Tai, Y. C., & Desmarais, B. A. (2025). Political elites in the attention economy: Visibility over civility and credibility? *Proceedings of the International AAAI Conference on Web and Social Media*, *19*(1), 241–258. https://doi.org/10.1609/icwsm.v19i1.35814

Biswas, A., Javadian Sabet, A., & Lin, Y.-R. (2025). TikTok political engagement dataset (Version V1) dataset. *Harvard Dataverse*. https://doi.org/doi:10.7910/DVN/CHYOPR

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22* [Technical report]. University of Texas at Austin. https://www.liwc.app

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 77–91). Proceedings of Machine Learning Research (PMLR). https://proceedings.mlr.press/v81/buolamwini18a.html

[10] A statistical method used to adjust *p*-values when performing multiple comparisons. It controls the false discovery rate—the expected proportion of incorrect rejections of the null hypothesis—helping reduce the likelihood of false positives.

Siegel-Stechler, K., Hilton, K., & Medina, A. (2025, May 12). *Youth rely on digital platforms, need media literacy to access political information*. CIRCLE, Tisch College, Tufts University. https://circle.tufts.edu/latest-research/youth-rely-digital-platforms-need-media-literacy-access-political-information

Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., & Webson, A. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*(70), 1–53. https://www.jmlr.org/papers/v25/23-0870.html

Diab, R. (2024, July 11). *The case for mandating finer-grained control over social media algorithms.* Tech Policy Press. https://www.techpolicy.press/the-case-for-mandating-finergrained-control-over-social-media-algorithms/

Donovan, J., & Boyd, D. (2021). Stop the presses? Moving from strategic silence to strategic amplification in a networked media ecosystem. *American Behavioral Scientist*, *65*(2), 333–350. https://doi.org/10.1177/0002764219878229

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, *26*(6), 3333–3361. https://doi.org/10.1007/s11948-020-00276-4

Ferrara, E., Cresci, S., & Luceri, L. (2020). Misinformation, manipulation, and abuse on social media in the era of COVID-19. *Journal of Computational Social Science*, *3*, 271–277. https://doi.org/10.1007/s42001-020-00094-5

Goldstein, I., Edelson, L., Nguyen, M. K., Goga, O., McCoy, D., & Lauinger, T. (2023). *Understanding the (in) effectiveness of content moderation: A case study of Facebook in the context of the U.S. Capitol riot.* arXiv. https://doi.org/10.48550/arXiv.2301.02737

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv https://doi.org/10.48550/arXiv.2203.05794

Gilda, S., Giovanini, L., Silva, M., & Oliveira, D. (2022). Predicting different types of subtle toxicity in unhealthy online conversations. *Procedia Computer Science*, *198*, 360–366. https://doi.org/10.1016/j.procs.2021.12.254

Guess, A., Nyhan, B., Lyons, B., & Reifler, J. (2018). *Avoiding the echo chamber about echo chambers: Why selective exposure to like-minded political news is less prevalent than you think.* Knight Foundation. https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/133/original/Topos_KF_White-Paper_Nyhan_V1.pdf

Islam, J., Xiao, L., & Mercer, R. E. (2020, May). A lexicon-based approach for detecting hedges in informal text. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3109-3113). European Languages Resources Association. https://aclanthology.org/2020.lrec-1.380/

Karimi, K., & Fox, R. (2023). Scrolling, simping, and mobilizing: TikTok's influence over Generation Z's political behavior. *The Journal of Social Media in Society*, *12*(1), 181–208. https://www.thejsms.org/index.php/JSMS/article/view/1251

Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., & Adams, J. (2024). Towards algorithm auditing: Managing legal, ethical and technological risks of AI, ML and associated algorithms. *Royal Society Open Science*, *11*(5), 230859. https://doi.org/10.1098/rsos.230859

Leppert, R., & Matsa, K. E. (2024). *More Americans–especially young adults–are regularly getting news on TikTok.* Pew Research Center. https://www.pewresearch.org/short-reads/2024/09/17/more-americans-regularly-get-news-on-tiktok-especially-young-adults/

Llama-3.2-3B-Instruct [Computer software]. (2024). Meta Llama. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

Marwick, A., & Lewis, R. (2017). *Media manipulation and disinformation online.* Data & Society Research Institute. https://datasociety.net/library/media-manipulation-and-disinfo-online/

McClain, C. (2023, August 20). *About half of TikTok users under 30 say they use it to keep up with politics, news*. Pew Research Center. https://www.pewresearch.org/short-reads/2024/08/20/about-half-of-tiktok-users-under-30-say-they-use-it-to-keep-up-with-politics-news/

Meleagrou-Hitchens, A., & Kaderbhai, N. (2017). *Research perspectives on online radicalisation: A literature review, 2006–2016*. International Centre for the Study of Radicalisation, King's College London. https://icsr.info/2017/05/03/icsr-vox-pol-paper-research-perspectives-online-radicalisation-literature-review-2006-2016-2/

Mistral-7B-Instruct-v0.3 [Computer software]. (2024). Mistral AI. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

Mosleh, M., Cole, R., & Rand, D. G. (2024). Misinformation and harmful language are interconnected, rather than distinct, challenges. *PNAS Nexus*, *3*(3), pgae111. https://doi.org/10.1093/pnasnexus/pgae111

Neidhardt, A. H., & Butcher, P. (2022). *Disinformation on migration: How lies, half-truths, and mischaracterizations spread.* Migration Policy Institute. https://www.migrationpolicy.org/article/how-disinformation-fake-news-migration-spreads

Perspective API [Computer software]. (2025). Jigsaw. https://www.perspectiveapi.com/

Pierri, F., Luceri, L., Chen, E., & Ferrara, E. (2023). How does Twitter account moderation work? Dynamics of account creation and suspension on Twitter during major geopolitical events. *EPJ Data Science*, *12*(1), Article 43. https://doi.org/10.1140/epjds/s13688-023-00420-7

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *ICML '23: Proceedings of the International Conference on Machine Learning* (pp. 28492–28518). Association for Computing Machinery. https://dl.acm.org/doi/10.5555/3618408.3619590

Resende, G. H., Nery, L. F., Benevenuto, F., Zannettou, S., & Figueiredo, F. (2024*). A comprehensive view of the biases of toxicity and sentiment analysis methods towards utterances with African American English expressions.* arXiv. https://doi.org/10.48550/arXiv.2401.12720

Serengil, S., & Özpınar, A. (2024). A benchmark of facial recognition pipelines and co-usability performances of modules. *Journal of Information Technologies*, *17*(2), 95–107. https://doi.org/10.17671/gazibtd.1399077

Somasundaran, S., Ruppenhofer, J., & Wiebe, J. (2007). Detecting arguing and sentiment in meetings. In H. Bunt, S. Keizer, & T. Paek (Eds.), *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue* (pp. 26–34). https://doi.org/10.18653/v1/2007.sigdial-1.5

Starbird, K. (2017). Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 230–239. https://doi.org/10.1609/icwsm.v11i1.14878

Perez S. (2021, January 7). *TikTok bans videos of Trump inciting mob, blocks #stormthecapital and other hashtags.* TechCrunch. https://techcrunch.com/2021/01/07/tiktok-bans-videos-of-trump-inciting-mob-blocks-stormthecapital-and-other-hashtags/

TikTok. (2024). *Community principles.* https://www.tiktok.com/community-guidelines/en/community-principles

TikTok. (2025a). *Research API.* TikTok for Developers. https://developers.tiktok.com/products/research-api/

TikTok. (2025b). *Our approach to content moderation.* TikTok Transparency Center. https://www.tiktok.com/transparency/en/content-moderation/

TikTok. (2025c). *Content moderation*. https://www.tiktok.com/euonlinesafety/en/content-moderation/

TikTok. (2025d). *Teen safety*. https://www.tiktok.com/euonlinesafety/en/teen-safety/

TikTok. (2025e). *How TikTok recommends content*. https://support.tiktok.com/en/using-tiktok/exploring-videos/how-tiktok-recommends-content

Paul K. (2020, October 7). Facebook announces plan to stop political ads after 3 November. *The Guardian.* https://www.theguardian.com/technology/2020/oct/07/facebook-stop-political-ads-policy-3-november

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Council of Europe. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

Wardle, C. (2024). *A conceptual analysis of the overlaps and differences between hate speech, misinformation and disinformation*. United Nations Peacekeeping. https://peacekeeping.un.org/en/conceptual-analysis-of-overlaps-and-differences-between-hate-speech-misinformation-and

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). OpinionFinder: A system for subjectivity analysis. In D. Byron, A. Venkataraman, & D. Zhang (Eds), *Proceedings of HLT/EMNLP interactive demonstrations* (pp. 34–35). Association for Computational Linguistics. https://aclanthology.org/H05-2018/

Acknowledgments
We thank NSF, AFOSR, ONR, the Pitt Cyber Institute, and the Pitt CRCD for providing partial support and resources that enabled this research. Any opinions, findings, conclusions, or recommendations expressed here do not necessarily reflect the views of the funding agencies. We also would like to acknowledge and thank Julie Lawler, Sodi Kroehler, and the wider team of PICSO Lab for their early input for this research. We are also grateful to the reviewers for their constructive feedback, which significantly strengthened the final manuscript.

Funding
This project did not receive direct support from any agency or foundation.

Competing interests
The authors declare no competing interests.

Ethics
This study does not involve human subjects, and IRB review was not applicable. A state-of-the-art approach, DeepFace, is employed to infer demographic features (e.g., gender, ethnicity, as defined by DeepFace), which we further manually validated (Appendix G). Recognizing potential minority bias (Buolamwini & Gebru, 2018), we recommend further research and the implementation of transparent mitigation strategies to ensure fairness and accountability in algorithmic detection.

While automated toxicity detection tools like Perspective API offer scalable methods for analyzing harmful language, they are not without limitations. Prior research has shown that such models can overestimate toxicity in certain contexts—particularly when analyzing emotionally expressive content, sarcasm, or speech containing African American Vernacular English (Gilda et al., 2022; Resende et al., 2024). To mitigate this, we use continuous toxicity scores and validated Perspective API outputs against human-annotated labels (see Appendix H), finding strong alignment. However, we acknowledge that residual biases may persist. These biases could potentially skew topic-level or group-level toxicity estimates—for example, inflating toxicity scores for content related to race, protest, or polarizing events—thereby influencing interpretations of which topics or partisan groups appear most toxic. While our statistical models control for topic and party interactions, we still warrant some caution in interpreting the results. These limitations underscore the need for future work on bias-aware, context-sensitive toxicity models.

As with any observational study relying on platform APIs, our dataset is shaped by visibility constraints and discovery mechanisms inherent to TikTok's design. While we cannot claim platform-wide representativeness, our sampling strategy—grounded in election-related hashtags, user engagement behaviors (likes), and multi-step snowball expansion—captures the ecosystem of political content actively circulating during the 2024 U.S. election. This includes videos engaged with from both Democratic- and Republican-aligned creators, as well as a wide range of topics and rhetorical styles. Our findings should therefore be interpreted as reflecting patterns within politically salient and user-engaged content, rather than all political TikTok posts or broader user sentiment. By focusing on videos that users actively engaged with, rather than relying solely on posted content, our study offers insights into how partisanship and toxicity surfaced and amplified within TikTok's recommendation-driven environment. Future work may expand upon this by comparing patterns across less engaging or emergent content spaces.

Data availability

We collected data using the TikTok Research API, following its terms of service. In accordance with TikTok's data-sharing policies, we are unable to share raw video content, transcripts, or user metadata. However, to support reproducibility, we have published the TikTok post IDs corresponding to the analyzed content on Harvard Dataverse: https://doi.org/10.7910/DVN/CHYOPR. All code required for analysis is available on our GitHub repository: https://github.com/picsolab/Toxic-Politics-and-TikTok-Engagement-in-the-2024-U.S.-Election.

Appendix A: Data collection & feature extraction

Data collection

To construct a comprehensive dataset of political TikTok content, we employed a multi-step data collection process focused on identifying politically active users and their engagement patterns during the 2024 U.S. presidential election. We first collected TikTok posts originating from the United States that contained the keyword "election" between January and March 2024, yielding over 361,000 posts by around 110,000 users. To refine our dataset, we identified relevant hashtags associated with U.S. politics, particularly those related to the presidential election, as described next. Then, we constructed a hashtag co-occurrence network (based on co-occurrence in video captions) and selected hashtags that appeared most frequently together using network backbone extraction. Next, to ensure a balanced representation of partisan content, we selected nine general election-related (#usa, #election2024, #election year2024, #electionyear, #america, #election2024prediction, #unitedstates, #election, #2024election), eight Democratic-leaning (#biden, #democrat, #biden2024, #voteblue2024, #joebiden, #joe, #biden2020, #democrats), and ten Republican-leaning (#donaldtrump, #trump2024, #trump, #maga, #republican, #trumptrain, #trump2020, #election2024trump, #electionfraud, #gop) hashtags that were among the most frequent in the co-occurrence network. Note that Biden was still the Democratic presidential candidate during this period. Using these hashtags, we identified politically active users (seed users) who post political content—that is, users who had at least three posts each containing any two of the hashtags.

We detected the partisan leaning of the seed users to monitor and limit partisan imbalance across iterative data collection rounds. This helped ensure that the final dataset was not disproportionately skewed in either direction due to our snowball sampling process. To determine the partisan alignment of these seed users, we employed FLAN-T5 large (Chung et al., 2024) to classify three of their posts that included voice-to-text transcriptions. The prompt was to simply classify a post as Republican/Democrat/Neither based on the stance expressed, and we use majority vote (out of 3 posts) to determine the leaning of the user. We were unable to identify the leaning of users for whom the voice-to-text transcriptions (15% of posts had transcriptions provided by API) were unavailable or who had fewer than three posts with transcriptions. Overall, we labeled over 11,000 posts by 3,826 (3.5%) users. Note that these seed users were responsible for around 27% posts in the dataset. To validate the accuracy of our labeling, we manually labeled 100 accounts, achieving a Cohen's Kappa of 0.53, indicating moderate agreement. Initially, we intended to collect videos posted by these politically active users. However, due to limitations in the TikTok API, we were unable to access their original posts. Instead, we collected videos that these users liked (1000 per user due to API restrictions), allowing us to analyze content that politically active users engaged with. To expand the seed users, we identified additional politically active users from the authors of these liked videos. This process was iteratively repeated three times, with diminishing returns in new user identification, suggesting saturation—3.5% new users in the third round compared to 47.5% after the second round.

Our final dataset comprised 4,624 labeled users, with the majority exhibiting a tendency to lean Republican (69%)—reflecting the initial data trends. Due to API constraints, we retrieved liked videos for 683 users (65% Republican-leaning, similar to the initial user leaning distribution), yielding 1,004,654 videos collected from January 1 to November 7, 2024. Using the previously mentioned hashtags—and additional tags like #kamala, #harris, #kamalaharris, #jdvance, #vance, #timwalz, and #walz—we identified 111,114 political videos containing at least one of these hashtags (11.06% of total liked videos collected), 51,680 (46.5% of political videos) of which were downloadable using a third-party library Pyktok (the remaining videos were likely removed by the user or platform). The mean number of posts during a week (i.e., 7 days) in our dataset was 1,055.8. Note, we were unable to identify the exact reason

for the removal of the remaining videos; nevertheless, our dataset reflects the naturally occurring set of publicly accessible political videos on TikTok during the collection period. Despite initial user imbalances and API limitations, our multi-stage, hashtag-seeded snowball sampling method produced a comprehensive video dataset that is balanced in partisan leanings, with downloadable content reflecting an approximately equal split between Republican- and Democratic-leaning videos (39.5% vs 37.1%) and included posts across a wide range of salient political topics and account demographics.

Video feature extraction

We extracted keyframes[11] from videos to capture the most visually representative moments. This is crucial because TikTok videos rely heavily on visual elements to convey implicit messages, evoke emotions, and engage audiences. We inferred user demographics, including age, gender, and ethnicity, and facial emotions such as happiness, anger, fear, and sadness from keyframes, using DeepFace (see Appendix G for manual validation), and RGB color saturation from each keyframe using OpenCV. We aggregated these features at the video level, such as calculating the mean RGB values or the proportions of various emotions and demographics featured, to provide insights into the overarching visual and emotional tone of each video. Figure A1 shows the demographic and video duration distribution in our dataset.

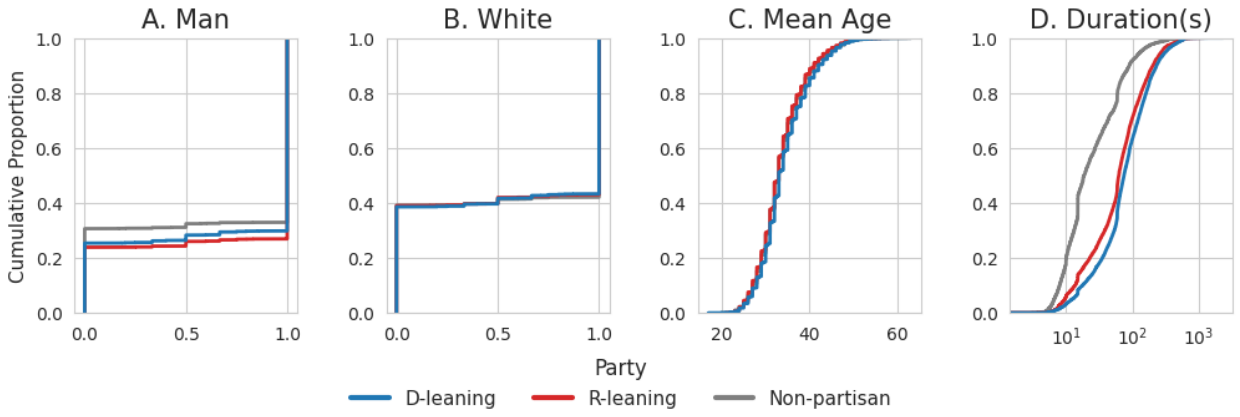

Figure A1. Demographic and video duration distribution for political TikTok videos. *Panel A shows proportion of man, panel B shows proportion of White, panel C shows mean age, as averaged over video keyframes. Our dataset is dominated by videos featuring men and White users, with an average age of 33.9. Panel D shows video duration in seconds, with an average duration of 89.8s.*

Furthermore, we extracted a variety of linguistic features from the video transcripts to analyze how language influences engagement. These features include markers of generalization (Somasundaran et al., 2007) and causation (Boyd et al., 2022) to assess logical structuring, hedges (Islam et al., 2020) and subjectivity (Wilson et al., 2005) to gauge certainty and bias, and emotion-related indicators such as positive emotions, anger, and anxiety (Boyd et al., 2022). These linguistic elements provide insights into the emotional and rhetorical strategies employed by content creators to engage or persuade their audiences.

[11] Keyframes refer to a subset of still frames extracted from each video that represent significant visual transitions or scenes. These are used to summarize visual content and enable analysis of attributes like color saturation, facial expressions, and visual composition without processing every frame of the video.

Appendix B: Partisan alignment detection

We experimented with various prompts and several state-of-the-art large language models (LLMs), including Flan-T5 Large, LLaMA-3.2-3B-Instruct (Meta Llama, 2024), and Mistral-7B-Instruct-v0.3, to classify partisan alignment of post transcripts and captions. We conducted a multi-stage manual validation process to select the final model and prompt. Two annotators iteratively labeled 1,000 samples over four rounds, refining both the prompt and labeling guidelines based on disagreement analysis. Discrepancies were resolved through discussion, and the final inter-annotator agreement reached a Cohen's Kappa of 0.77, indicating substantial agreement. After prompt and model tuning, we selected the Mistral-7B-Instruct-v0.3 model for its high alignment with human-labeled data using the optimal prompt detailed in Table B1. When evaluated against the human labels, this model achieved an accuracy of 82%, precision of 0.82 recall of 0.82, and F1 score of 0.81. Cohen's Kappa is 0.74 between model outputs and human labels, which emphasizes similarity in agreement rates between LLM and human raters. These results demonstrate the robustness of the LLM-based classification pipeline when coupled with prompt engineering and human-in-the-loop validation.

Prompt instruction for detecting partisan alignment

As a highly knowledgeable and objective political analyst, your task is to determine the political leaning of the given text from transcripts of U.S. political videos based solely on the sentiment expressed, regardless of whether it contains offensive, obscene, or derogatory language. You must provide a political leaning even if the text lacks specific policy details or uses harsh language.

Use the following criteria:

(1) "Democrat" if the text supports Democratic figures or Democratic values like climate action, social justice, healthcare reform, and increasing taxes on high-income earners, or criticizes Republican figures or policies.
(2) "Republican" if the text supports Republican figures or Republican values like strong border policies, traditional family values, and gun rights, or it criticizes, opposes, or uses derogatory language toward Democratic figures or Democratic values.
(3) "Neither" if the text does not clearly support, oppose, or criticize any political figures or values of either party, and does not contain any language that indicates clear political affiliation.

It is crucial that you provide an answer based on the sentiment expressed, even if the language is obscene or offensive. Make your choice based strictly on these criteria and respond with only one of the following labels: "Republican," "Democrat," or "Neither."

Provide a rationale for your label and clearly state which of the above criteria you are using. Please first provide the rationale for the answer, followed by the answer in the format: rationale: label. Ensure that the rationale is clearly explained before the label is presented. Text: {text}

Appendix C: Topic analysis

We initially used standard BERTopic (Grootendorst, 2022) to identify topics from the video transcripts. However, after a round of manual validation, we found that these topics tended to be noisy. We selected the top 50 topics based on frequency and manually filtered keywords (provided by BERTopic) to identify the most relevant ones for identifying topics. This was followed by multiple rounds of manual validation and merging certain topics to obtain the most representative topics in our datasets. Finally, we got 22 salient topics using our keyword-based approach (the presence of at least 3 keywords) with around 63% of posts being assigned at least one topic. Two annotators labeled 20 posts per topic to validate the final topic assignment—interrater Cohen's Kappa reliability was 0.92 (near perfect agreement). The precision of keywords ranges between 84–100% for each topic. The topics included "Election 2024," "Religion," "Economic Issues," "Abortion," "MAGA," "January 6 Riots," "Impeachment," "Obama," "Racism," "Project 2025," "Immigration," "Socialism," "Gun," "Israel War," "Ukraine War," "Nazi," "Hunter Biden," "Afghanistan," "Trump's Assassination Attempt," "Labor," "Election Fraud," and "Cyclone Helene." Figure C shows the volume, median views, and median interaction on topical posts by partisan alignment.

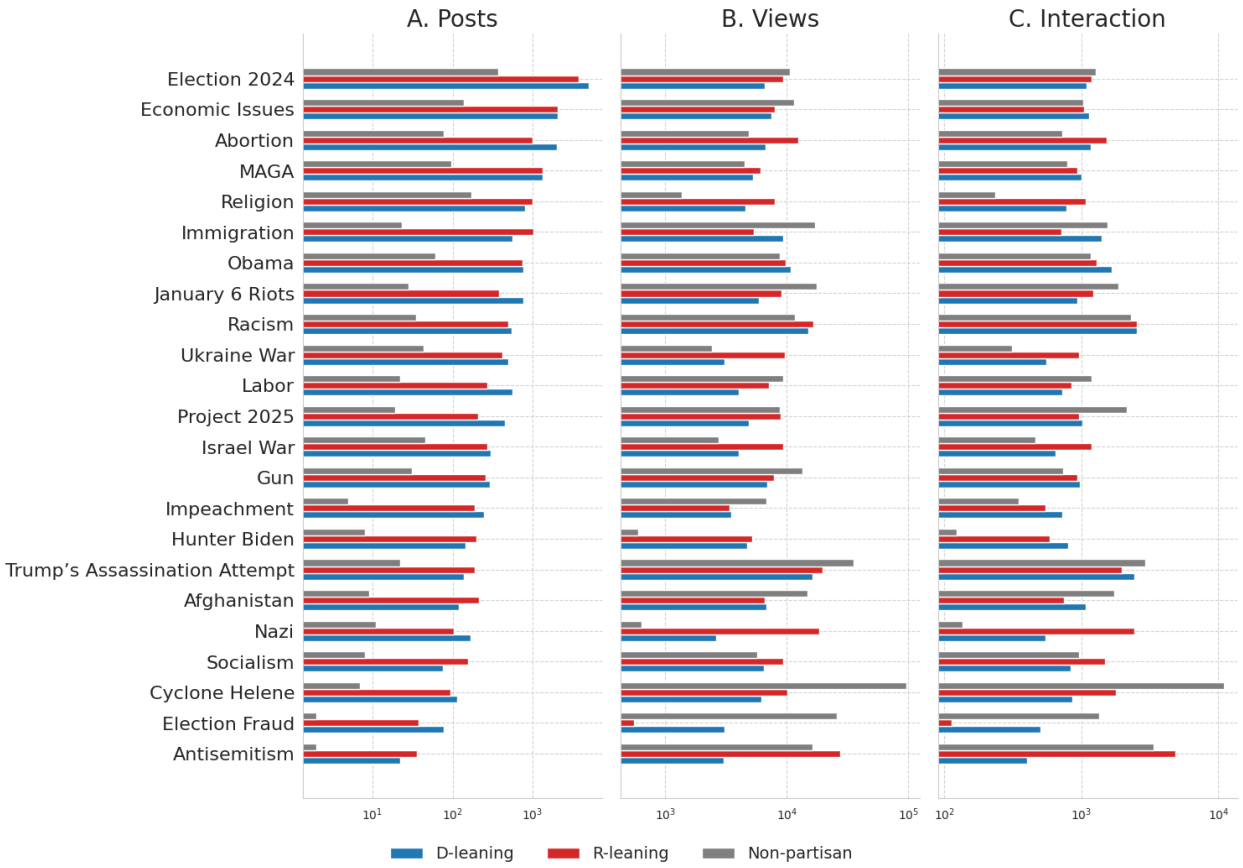

Figure C1. Total posts, views, and interactions for topics. *Posts about the 2024 elections, economic issues, and contentious issues like abortion and immigration are most frequent. Videos talking about issues like racism or antisemitism and Trump's assassination attempt get high views and interactions.*

Table C1 shows the topic grouping used for RQ3. In addition to these topic groups, "Immigration" and "Labor" were analyzed as individual topics, as they were particularly influential in the political discourse surrounding the 2024 U.S. presidential elections.

Table C1. *Topic grouping for RQ3.*

Grouping	**Topics**
Political Figures & Events	*Impeachment, January 6 Riots, Project 2025, MAGA, Obama, Hunter Biden, Trump's Assassination Attempt*
Socio-cultural Issues	*Racism, Abortion, Religion, Socialism, Gun, Nazi*
Elections	*Election 2024, Election Fraud*
Geopolitical Conflicts	*Israel War, Ukraine War, Afghanistan*
Economy	*Economic Issues, Cyclone Helene*

Appendix D: Toxicity distribution by topic

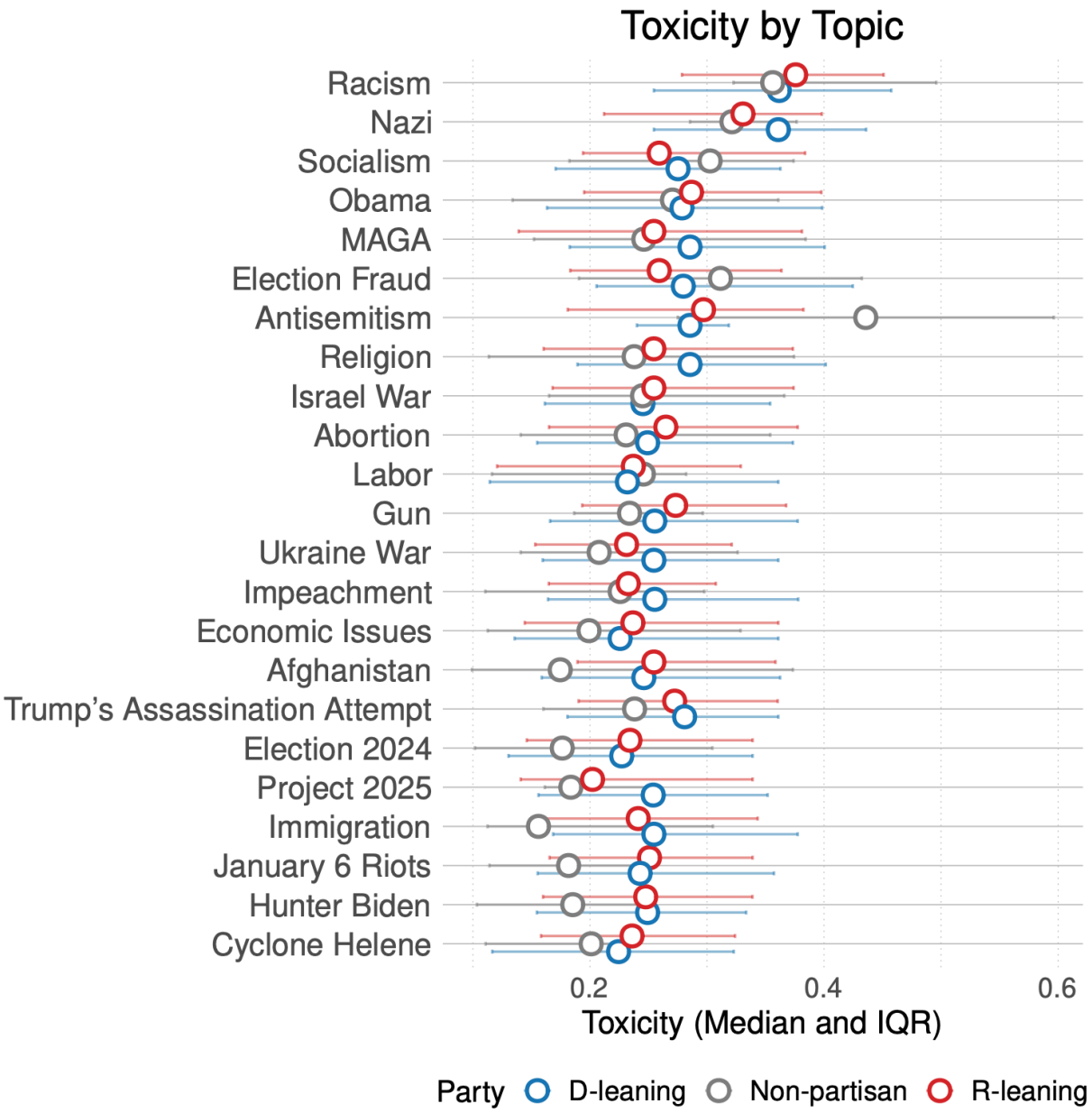

Figure D1. Distribution of toxicity in topical posts. *Topics like racism, antisemitism, and election fraud, and topics about political ideologies like Nazi, MAGA, and socialism showed the highest levels of toxicity.*

Appendix E: Toxicity detection using video captions vs. transcripts

We selected 300 (100 for each partisan alignment) videos randomly and ran the toxicity classification on only video captions. The results are compared to toxicity scores using full video transcripts (see Appendix H for manual validation). As shown in Figure E1, using only captions is not enough to capture the toxicity (including subtypes) levels in the video.

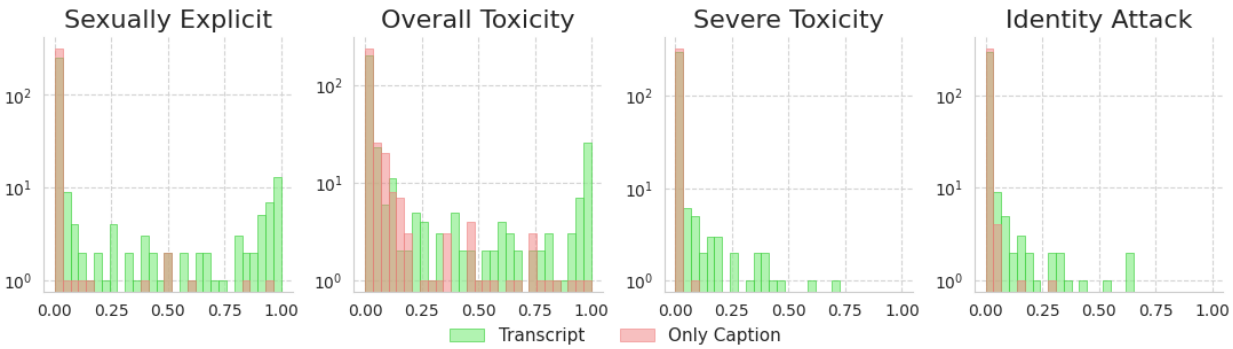

Figure E1. Toxicity distribution using only captions vs. video transcripts. *Using only captions failed to predict toxic content levels present in the video.*

Appendix F: Regression model selection, specification, and full results

We estimated the observed effects (after controlling for views) on interaction for RQ2–4 using linear mixed effects models. We first selected the best-performing baseline model using ANOVA-based model selection to determine the most relevant features influencing user interactions. To capture content-specific idiosyncrasies, we incorporated random effects for post author, featured music, and posting time, which enhances methodological rigor by disentangling algorithmic amplification from organic engagement patterns, offering a more precise analysis of how partisanship, toxicity, and political topics shape audience interactions. We experimented with several variations of models for RQ2–4, the best models in terms of high R^2 and low complexity (using Akaike Information Criterion or AIC) were selected. The baseline model features included partisan alignment, red hue, duration, hedging, anger, and age, after controlling for views. We use the following model for RQ2:

$$y_{ui} \sim \alpha_0 + \alpha_t t_i + \alpha_e t_i p_i + \alpha_p p_i + \overrightarrow{\alpha_K} K_i + m_i + \alpha_v views_i + s_u + w_i$$

Where:

- y_{ui} is the interaction on author u's post i.
- t denotes toxicity score.
- α_t is the effect of toxicity on interaction.
- p denotes partisan leaning of a post.
- α_e captures the joint effect of party and toxicity on interactions.
- K is the vector of other post features (red hue, duration, hedging, anger, and age) from the baseline model.
- s , w , and m denote random effects on the user, post timing, and featured music, respectively.

Model fit and sample:

- *N* = 37,929 observations
- Marginal R^2 = .894 (variance explained by fixed effects)
- Conditional R^2 = .930 (variance explained by fixed and random effects)

Table F1. *Regression estimates, standard errors, 95% CI, and p-values for RQ2.*

Predictor	**Estimate**	***SE***	**95% CI (LL, UL)**	***p***
Intercept	0.000	0.006	[-0.011, 0.012]	.943
Toxicity	0.023	0.003	[0.017, 0.028]	< .001 ***
Party: Neither	-0.040	0.004	[-0.049, -0.031]	< .001 ***
Party: R-leaning	-0.007	0.004	[-0.014, 0.000]	.042 *
Red Hue	0.007	0.002	[0.004, 0.010]	< .001 ***
Duration	0.010	0.002	[0.005, 0.014]	< .001 ***
Hedges	-0.016	0.004	[-0.025, -0.008]	< .001 ***
Anger	0.003	0.004	[-0.004, 0.011]	.356
Views	0.917	0.002	[0.913, 0.920]	< .001 ***
Age	0.005	0.002	[0.003, 0.008]	< .001 ***
Toxicity*Party: Neither	-0.014	0.004	[-0.022, -0.007]	< .001 ***
Toxicity*Party: R-leaning	-0.006	0.003	[-0.013, 0.001]	.078 †

Note: p values: † *p* < .10, * *p* < .05, ** *p* < .01, *** *p* < .001. CI = Confidence Interval; *SE* = Standard Error.

For RQ3, individual topics are grouped into broader topic categories to enhance statistical power and interpretability, and nonpartisan content is excluded due to insufficient topical posts (see Figure C1, Appendix C). We use the following model:

$$y_{ui} \sim \alpha_0 + \overrightarrow{\alpha_G} G_i + \overrightarrow{\alpha_L} t_i G_i + \overrightarrow{\alpha_N} p_i G_i + \overrightarrow{\alpha_H} t_i p_i G_i + \alpha_t t_i + \alpha_e t_i p_i + \alpha_p p_i + \overrightarrow{\alpha_K} K_i + m_i + \alpha_v views_i + s_u + w_i$$

Where:

- G is the vector of topic groups.
- α_G gives the effect of topic groups on interaction.
- α_L and α_N capture the interaction effects of toxicity and party with topic groups.
- $\alpha_H$captures the three-way interaction effect between topic group, party and toxicity.

Model fit and sample:

- *N* = 29,425 observations
- Marginal R^2 = .896 (fixed effects)
- Conditional R^2 = .931 (fixed + random effects)

Table F2. *Regression estimates, standard errors, 95% CI and p-values for RQ3.*

Predictor	**Estimate**	***SE***	**95% CI (LL, UL)**	***p***
Intercept	-0.010	0.007	[-0.023, 0.003]	.118
Toxicity	0.020	0.003	[0.013, 0.026]	< .001 ***
Party: R-leaning	-0.004	0.004	[-0.012, 0.004]	.335
Elections	0.002	0.006	[-0.009, 0.014]	.717
Economy	-0.001	0.008	[-0.017, 0.015]	.880
Socio-cultural Issues	0.034	0.007	[0.020, 0.047]	< .001 ***
Political Figures & Events	0.012	0.007	[-0.001, 0.025]	.080 †
Geopolitical Conflict	-0.003	0.012	[-0.027, 0.020]	.771
Immigration	-0.029	0.015	[-0.058, ~0.000]	.049 *
Labor	0.032	0.015	[0.002, 0.062]	.039 *
Red Hue	0.011	0.002	[0.008, 0.015]	< .001 ***
Duration	0.007	0.003	[0.002, 0.012]	.011 *
Hedges	-0.007	0.005	[-0.017, 0.003]	.177
Anger	0.003	0.004	[-0.005, 0.011]	.502
Views	0.910	0.002	[0.906, 0.914]	< .001 ***
Age	0.007	0.002	[0.004, 0.011]	< .001 ***
Toxicity*Party: R-leaning	-0.005	0.004	[-0.013, 0.003]	.207
Toxicity*Elections	0.013	0.007	[0.000, 0.026]	.044 *
Toxicity*Immigration	0.035	0.020	[-0.004, 0.074]	.076 †
Party: R-leaning*Immigration	0.047	0.018	[0.012, 0.083]	.009 **
Party: R-leaning*Labor	-0.085	0.025	[-0.134, -0.036]	< .001 ***
Toxicity*Party: R-leaning* Geopolitical Conflict	0.050	0.023	[0.005, 0.095]	.028 *

Note: p values: † *p* < .10, * *p* < .05, ** *p* < .01, *** *p* < .001. CI = Confidence Interval; *SE* = Standard Error.

For RQ4, we use the following model:

$$y_{ui} \sim \alpha_0 + \alpha_z z + \alpha_l z x_i + \alpha_x x_i + \alpha_p p_i + \overrightarrow{\alpha_K} K_i + m_i + \alpha_v views_i + s_u + w_i$$

Where:

- z is a binary indicator denoting the event (i.e., $z = 1$).
- α_z shows the effect of the event on interaction.
- α_l effect of toxicity subtype x on interaction following the event.

Severe toxicity model fit and sample:

- *N* = 3,008 observations
- Marginal R^2 = .910 (fixed effects)
- Conditional R^2 = .961 (fixed + random effects)

Table F3. *Regression estimates, standard errors, 95% CI and p-values for RQ4 (severe toxicity).*

Predictor	**Estimate**	**SE**	**95% CI (LL, UL)**	***p***
Intercept	-0.037	0.014	[-0.065, -0.010]	.007 **
Severe Toxicity	0.007	0.006	[-0.004, 0.019]	.196
Post-Event	0.018	0.009	[0.001, 0.035]	.035 *
Party: D-leaning	0.027	0.012	[0.004, 0.050]	.020 *
Party: R-leaning	0.028	0.011	[0.006, 0.050]	.013 *
Red Hue	0.005	0.004	[-0.004, 0.013]	.258
Duration	0.006	0.006	[-0.006, 0.017]	.333
Hedges	-0.041	0.011	[-0.063, -0.019]	< .001 ***
Anger	0.009	0.010	[-0.011, 0.029]	.357
Views	0.952	0.006	[0.941, 0.963]	< .001 ***
Age	0.005	0.004	[-0.003, 0.013]	.182
Severe Toxicity*Post-Event	0.016	0.008	[0.001, 0.032]	.040 *

Note: p values: † *p* < .10, * *p* < .05, ** *p* < .01, *** *p* < .001. CI = Confidence Interval; *SE* = Standard Error.

Sexually explicit toxicity model fit and sample:

- N = 3,008 observations
- Marginal R^2 = .910 (variance explained by fixed effects)
- Conditional R^2 = .961 (variance explained by fixed and random effects)

Table F4. *Regression estimates, standard errors, 95% CI and p-values for RQ4 (sexually explicit toxicity).*

Predictor	**Estimate**	**SE**	**95% CI (LL, UL)**	***p***
Intercept	-0.042	0.014	[-0.069, -0.015]	.002 **
Sexually Explicit	0.001	0.006	[-0.011, 0.011]	.929
Post-Event	0.019	0.009	[0.002, 0.036]	.027 *
Party: D-leaning	0.029	0.012	[0.006, 0.053]	.013 *
Party: R-leaning	0.031	0.011	[0.009, 0.053]	.007 **
Red Hue	0.005	0.004	[-0.003, 0.013]	.233
Duration	0.005	0.006	[-0.006, 0.016]	.398
Hedges	-0.041	0.011	[-0.063, -0.020]	< .001 ***
Anger	0.016	0.010	[-0.004, 0.035]	.113
Views	0.952	0.006	[0.940, 0.963]	< .001 ***
Age	0.005	0.004	[-0.002, 0.013]	.180
Sexually Explicit*Post-Event	0.020	0.008	[0.004, 0.035]	.013 *

Note: p values: † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$. CI = Confidence Interval; *SE* = Standard Error.

Appendix G: Manual validation of DeepFace demographic and emotion predictions

To evaluate the accuracy of DeepFace on our dataset, we conducted a manual validation of its predictions for gender, ethnicity, age group, and facial emotion. We randomly selected one keyframe from each of 50 randomly sampled TikTok videos, and a single human annotator labeled the demographic and emotional attributes visible in the selected frame. Given the visual clarity of the tasks and the binary or coarse-grained nature of the labels, a single coder was used for this evaluation. For gender (male/female) and ethnicity (white/non-white), DeepFace achieved an accuracy of 84.4% and 95.5% respectively. For age and dominant emotion, the proportion of correct predictions was 78% and 74%, respectively. When no face was detected, DeepFace returned null values. These results suggest that DeepFace performs reliably on the types of political TikTok videos in our dataset and is suitable for extracting high-level demographic and affective trends at scale.

Appendix H: Manual validation of toxicity detection

To evaluate the validity of toxicity scores used in our study, and to substantiate the claim that full video transcripts better capture toxic content than captions alone, we conducted two manual validation tasks using a sample of TikTok videos from our dataset. Two annotators independently labeled each sample for both tasks, and final labels were decided by mutual discussion.

Validation of perspective API toxicity scores

We randomly sampled 100 TikTok videos and manually labeled each as toxic or non-toxic, based on the primary verbal content in the transcript. Since our study employs the continuous toxicity scores from the Perspective API, we binarized the scores using the median toxicity score in our dataset as the cutoff point to assess agreement with human annotations. The validation yielded an accuracy of 92%, precision 0.92, recall 0.92, and F1 score 0.92. The inter-rater Cohen's Kappa agreement is 0.79 (substantial agreement). These results indicate high agreement between human judgment and Perspective API predictions, justifying its use for toxicity inference in our analysis.

Validation of transcript vs. caption toxicity comparison

To test our claim in Finding 5—that full video transcripts capture more toxicity than video captions—we sampled 100 random video pairs, each consisting of a caption and a transcript for the same video. A pair was labeled as positive by human annotators if the transcript conveyed more toxic language or sentiment than the caption. This was compared against the Perspective API scores, where a difference of ≥ 0.1 between transcript and caption scores (transcript > caption) was treated as a positive case. The validation yielded an accuracy of 90%, precision 0.89, recall 0.89, and F1 score 0.89. The inter-rater Cohen's Kappa agreement is 0.81 (near perfect agreement).

Manual inspection of the captions revealed that many consist solely of hashtags, often generic and non-toxic (e.g., #usa, #vote2024, #biden, #maga). These captions lack the linguistic content needed to detect nuanced or hostile rhetoric. In contrast, the transcripts frequently contain substantive political commentary, including personal attacks, emotionally charged statements, and explicit toxicity, which are not reflected in the caption alone.