Triangulating Political Toxicity on Twitter: Evidence from a Cross-Model LLM Measurement Approach

Neil Fasching[1], **Sam Wolken**[1], **and Tim Dörr**[1]
[1]University of Pennsylvania
{neil.fasching,sam.wolken,timothy.dorr}@asc.upenn.edu

Abstract

Toxic political speech can inhibit democratic processes. However, despite these concerns, past attempts to quantify political toxicity have been limited by insufficient measurement techniques and narrow timeframes. Here, we deploy two separate state-of-the-art toxicity detection systems (OpenAI's OMNI-MODERATION endpoint and Mistral's MODERATION model) to quantify harassment, hate, and violent political discourse in a large, representative sample of tweets ($n = 46.7$M) covering 2012 through 2022. We compare political tweets to a benchmark of randomly selected tweets to disaggregate change in political toxicity from broader, platform-wide shifts in language. Our results show a sharp increase in the prevalence of harassment in political discourse from 2016 to 2020, and our regression analyses reveal that different forms of toxicity have distinct relationships with account reach with hate speech associated with higher follower counts, while harassment and violent content generally correlate with reduced audience size.

1 Introduction

Commentators have long argued that social media networks have the potential to function as a "public square" in which citizens can engage in current events and politics (Fuchs, 2015). As a growing number of citizens experience politics through social media networks (Mitchell et al., 2020), the importance of social media to democratic processes may be growing. Thus, scholars and pundits have argued that the presence of toxicity in online political discourse poses a substantial threat to social welfare (Settle, 2018; Suhay et al., 2018) because toxic behavior in political discussions shuts down deliberation (Juncosa et al., 2024) and contributes to polarization by exposing citizens to extreme behavior, leading them to perceive partisans as more extreme (Kim et al., 2021). Ultimately, if citizens perceive politics as hostile due to toxicity in political discourse on social media, they will be less likely to engage with and participate in democracy (Klar and Krupnikov, 2016).

Past work has shed light on the engines of toxicity on social media. A small share of users generate most toxic content (Kim et al., 2021), and these users engage in toxic speech in both political and nonpolitical contexts (Mamakos and Finkel, 2023). However, toxicity in online political conversation varies in degree over time. Discourse on social media reflects the broader news environment (King et al., 2017), and rhetoric from political elites can instigate toxic behavior in online communities (Guldemond et al., 2022). As a result, past work on political toxicity that focuses on specific points in time (e.g., Juncosa et al., 2024; Xu, 2024) may not generalize to the future or capture long-term trends (Munger, 2023). In addition, the composition and behavior of users on social networks is in constant flux. Furthermore, lexicon-based content analysis strategies are unable to reliably identify toxic political speech because toxicity can be subtle and contextual (Sheth et al., 2022; Kiritchenko et al., 2023).

The societal impact of toxicity in online political discourse depends on its prevalence, but no research to date offers a consistent measurement of toxicity over time. We address this gap by conducting a comprehensive longitudinal analysis of political discourse on Twitter,[1] spanning a decade that provides unprecedented temporal scope. To process this extensive social media corpus, we employ both OpenAI's OMNI-MODERATION endpoint and Mistral's MODERATION model, ensuring measurement consistency across the dataset. As recent research has shown substantial variation in toxicity

[1]For clarity and consistency, we refer to the company as Twitter and the posts as tweets throughout this paper, as these were their designated names during the period when this data was first collected.

classifications across moderation models (Fasching and Lelkes, 2025), we employ a dual-model approach to achieve several important goals: enhancing reliability through multi-model validation of toxicity classifications; reducing potential biases inherent in single-model systems; and providing detailed categorization of different toxicity types (e.g., hate speech, harassing speech, and violent speech) with quantifiable confidence scores.

2 Methodology

2.1 Data

To analyze trends in toxic content on Twitter over time, we leveraged the Dataset of Historical Tweets (DHT), a representative sample of tweets compiled by The Wharton School and The Annenberg School for Communication at the University of Pennsylvania. This dataset, which spans from April 2012 through November 2022, captures approximately 1% of Twitter's total volume and includes roughly 4.6 trillion tweets in total.

From the DHT, we extracted two datasets for our analysis. The first consists of roughly 200,000 randomly sampled English-language tweets per month. This random sample of tweets (n = 21,877,547) provides a baseline of toxicity on Twitter over time to compare to the political tweets. The second subset of tweets captures political discourse. To identify political tweets, we filtered the DHT using a curated list of political keywords created using a multi-step approach.

First, we extracted all article headlines published between 2012 and 2023 from the New York Times Archive API, generated n-grams from these headlines, and trained a logistic regression model to predict whether a headline belonged to the "Politics" section. We identified the top 5% of n-grams most strongly associated with political content and manually reviewed each for relevance. Next, we employed GPT-4o and DeepSeek V3 to classify a random sample of 2 million tweets as political or non-political. Using tweets classified as political by both models, we trained another logistic regression model to identify words most predictive of political discourse. To ensure comprehensive coverage, we selected tweets classified as political by either model but missed by our keyword approach, applied further logistic regression analysis to these tweets, and conducted manual reviews of tweet subsets to identify additional relevant terms. We also specifically added names and Twitter handles of all

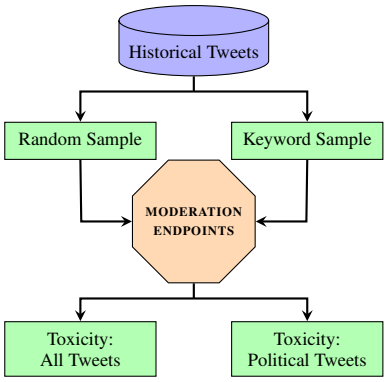

Figure 1: Overview of toxicity labeling pipeline.

U.S. presidents and vice presidents who held office since 2012 (see Appendix A for a comprehensive description of the keyword creation process and a complete list of the keywords).

In total, we use 268 political keywords that can be categorized into six different categories: institutions (n = 23), issues (n = 30), locations (n = 29), political terms (n = 28), politicians (n = 86), and processes (n = 72). Using these keywords, coupled with regular expressions (regex) to capture variations of these terms, we sampled roughly 200,000 political tweets per month (n = 24,913,379) from the DHT.[2] In total, our dataset, which comprised both the baseline and the political keywords tweets, consisted of 46,790,926 tweets spanning 128 months.

2.2 Measuring Toxicity

As toxicity is a multifaceted concept (Hanscom et al., 2024), we operationalize toxic speech in this analysis using three key dimensions: *hate* speech, *harassing* speech, and *violent* speech. To quantitatively measure these dimensions, we employed OpenAI's moderation endpoint, which has been upgraded to utilize a new, more advanced model, which is based on their GPT-4o model. For this analysis, we specifically utilized the OMNI-MODERATION-2024-09-26 model[3], which offers enhanced accuracy over previous versions.[4] While the moderation endpoint provides both binary clas-

[2]Due to data sparsity, some months have fewer than 200,000 containing at least one keyword in the DHT, especially in the earlier years of the dataset.

[3]https://platform.openai.com/docs/guides/moderation

[4]While toxicity classification is well known to a challenging and subjective task for human coders (Kumar et al., 2021), we find reasonably high agreement with human labels and GPT Omni-MODERATION—with accuracy ranging from .94 to .99 and F1 ranging from .4 to .62—across the three measured dimensions (see Table 7).

sifications and "confidence scores," we use the "confidence scores" provided by the endpoint (varying between 0 and 1) to measure our three dimensions of interest, as the continuous confidence scores retain more information than the binary indicators and OpenAI recommends using the scores rather than the binary classifications.[5] However, our results are robust to this decision, with the main findings remaining consistent when utilizing the binary classifications provided by the endpoint (see Appendix C).

To address potential model-specific variations in moderation scores highlighted by recent research (Fasching and Lelkes, 2025), we additionally employed Mistral's MODERATION Model as a robustness check to validate the consistency of our primary findings. This model, based on Ministral 8B 24.10, offers comprehensive classification across nine policy dimensions. Notably, Mistral's model provides a score for *hate and discrimination* and another for *violence and threats*. For the purposes of our analysis, we utilize Mistral's *hate and discrimination* score as a combined proxy for OpenAI's separate hate and harassment values, and Mistral's *violence and threats* score as analogous to OpenAI's violence value. Similar to OpenAI's OMNI-MODERATION model, Mistral's moderation endpoint provides category scores ranging from 0 to 1, with reported precision between 0.8-0.9 and recall between 0.7-0.99 on their internal test sets. Figure 1 provides a high-level overview of the data sampling and annotation process.

3 Results

3.1 Overall Trends in Toxicity

Toxic speech was consistently more prevalent in political tweets than Twitter conversation as a whole across the entire time period. As shown in Appendix B, this pattern holds across all three measured categories—*hate* speech, *harassment*, and *violent* content—compared to random tweets, though with varying degrees of disparity. Figure 2 shows the average difference in toxicity scores between political tweets and randomly selected tweets over time.[6] As shown in Appendix E (Table 8 and Table 9), these differences are statistically significant.

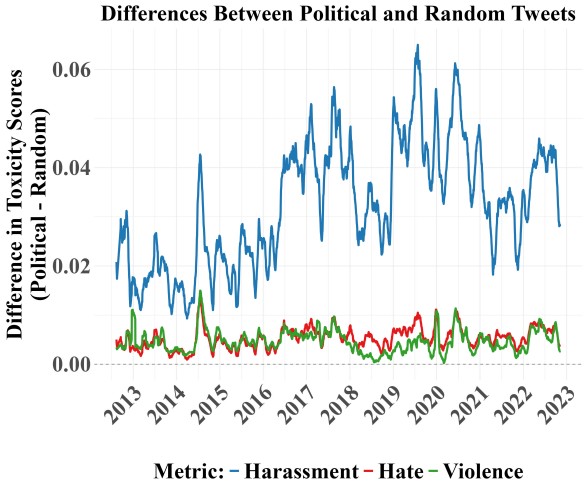

Figure 2: Difference in proportion of toxic content across three toxicity dimensions–*hate*, *harassment*, and *violent speech*–between random and political tweets, as measured by OpenAI's Moderation endpoint. Plotted values are 30-day rolling averages to smooth daily noise. Positive values indicate that political tweets had higher toxicity scores on average at a given time.

The prevalence of toxicity in political tweets was volatile over the course of our sample. There are significant spikes across all three toxicity dimensions, notably in late 2014 and in mid-2020.[7] In contrast, the prevalence of toxicity in the broader Twitter discourse remained comparatively stable over this timeframe.[8] The average *hate* speech score for randomly selected tweets, for instance, hovered just below .01 throughout the observation period. Across all years, political tweets had an average *hate* score of 0.014 (sd = 0.06), while the random sample of tweets had an average *hate* score of 0.009 (sd = 0.05). While the overall prevalence of *hate* was low, the difference between political and random tweets was statistically significant ($p < .001$)

As shown in Figure 2, *harassment* was the dimension with the most pronounced disparity between political and random tweets. Overall, political tweets exhibited higher *harassment* scores (m = 0.064, sd = 0.17) compared to the baseline random sample (m = 0.035, sd = 0.13), the differ-

[5] OpenAI uses the following thresholds to binarize these scores: 0.44 for harassment, 0.4 for hate, and 0.58 for violence. Mistral uses 0.63 for hate and harassment and 0.82 for violence.

[6] In general, the average political tweet had approximately a 1 to 2 percent probability of toxicity compared to 0.7 to 1 percent for random tweets.

[7] The 2014 spike approximately coincides with widespread Black Lives Matter protests, and the spike in early 2020 coincides with the outbreak of the COVID-19 pandemic.

[8] These trends are robust to alternative political classification techniques. In Figure 5, we show time trends in *hate*, *harassment*, and *violent* speech within tweets identified as political by GPT-4O-MINI rather than our lexicon-based strategy. The general patterns are the same across both approaches.

ence of which is statistically significant ($p < .001$). Political tweets exhibited substantial fluctuations in *harassment* scores: As shown in Figure 2, the average harassment score varied from .04 to .10. In contrast, the average *harassment* score was more stable in the random tweets, ranging between .025 and .04. The gap between political tweets and the random tweets was largest during the 2020 presidential election period. The peak in average *harassment* scores for political tweets occurred during the 2020 presidential election period, at which point the average *harassment* score for political tweets was three times that of randomly selected tweets.

The disparity between political tweets and randomly selected tweets is smallest when it comes to violent content, as depicted in Appendix B. *Violent* scores for political tweets range from 0.01 to over 0.02, compared to 0.005 to 0.01 in random tweets, a difference that is statistically significant ($p < .001$). While this represents a narrower gap than observed in other categories, political tweets again displayed greater volatility, with notable spikes reaching 0.022 in early 2015 and mid-2020. Interestingly, *violent* content in random tweets increased between 2018 and 2020, with a subtle but noticeable elevation from 0.006 to 0.01, suggesting a broader societal shift in online discourse during this time. As with the two previous metrics, political tweets showed slightly higher *violent* scores ($m = 0.012$, $sd = 0.06$) compared to the baseline random sample ($m = 0.008$, $sd = 0.05$).

These findings collectively indicate that political discourse on Twitter consistently generates higher toxicity scores across all three categories, with *harassment* showing the largest disparity, followed by *hate* speech, and then *violent* content. The data also reveals that major political events and societal crises tend to amplify these differences, particularly in political tweets, while Twitter conversation at large generally maintained more stable toxicity scores.

To validate our primary findings and mitigate possible model-specific biases, we repeated our entire analysis using Mistral's MODERATION model as a robustness check. The patterns observed with the Mistral model largely mirrored our primary findings with OpenAI's OMNI-MODERATION endpoint, confirming that political tweets consistently exhibited higher toxicity scores than random tweets across the study period. As illustrated in Figure 3, the difference between political and random tweets remained pronounced

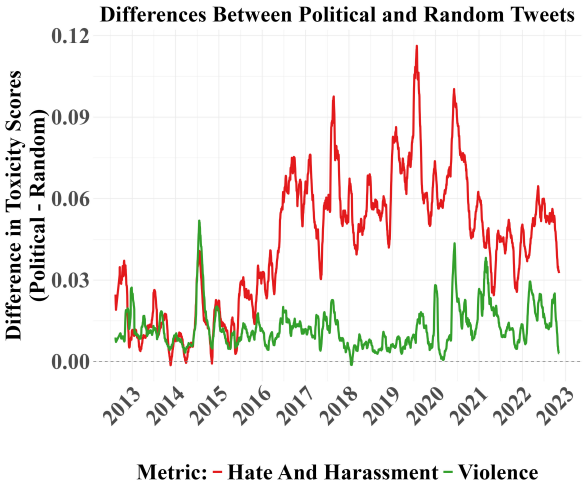

Figure 3: Difference in proportion of toxic content across two toxicity dimensions–*hate and harassment* (red) and *violent* (green) speech–between random and political tweets, as measured by Mistral's Moderation model. Plotted values are 30-day rolling averages to smooth daily noise. Positive values indicate that political tweets had higher toxicity scores on average at a given time.

across both toxicity dimensions measured by Mistral's model, especially for the *hate and harassment* values. Notably, the Mistral model detected the same significant spikes in toxicity that we observed in late 2014 and mid-2020 using the OMNI-MODERATION endpoint, reinforcing our observation that these periods represented particularly volatile moments in online political discourse. The difference between political and baseline tweets is particularly notable for the *hate and harassment* category, where political tweets show significantly elevated scores compared to random tweets throughout the time period. Across all years, political tweets had an average *hate and harassment* score of 0.104 ($sd = 0.249$), while the random sample of tweets had an average *hate and harassment* score of 0.060 ($sd = 0.178$). This difference is statistically significant ($p < .001$). Further aligning with OpenAI's moderation endpoint results, in 2019 and 2020, 14% of all political tweets had *hate and harassment* scores higher than .4, roughly double the rate of 2014 and 2015.

The Mistral analysis confirmed our *hate and harassment* findings, while also revealing similar patterns for *violent* content. Political tweets consistently showed higher *violent* scores (m= 0.038, sd = 0.140) than the random sample ($m = 0.026$, sd = 0.102) across all years. This difference, though less pronounced than for *hate and harassment* met-

rics, aligned with our OpenAI moderation endpoint observations and is statistically significant ($p < .001$). Interestingly, the models diverged regarding the 2014 spike in *violent* political tweets. Both detected the spike, but Mistral indicated *violent* tweets became more prevalent than *hate and harassment* tweets, while OpenAI found the reverse. This discrepancy underscores the importance of using multiple moderation models for cross-validation, as each system may detect different aspects of toxic content. Nevertheless, the overall consistency between Mistral and OpenAI results strengthens our confidence in the observed patterns and reduces concerns about model-specific biases affecting our conclusions.

3.2 Differences by Political Categories

Table 1: Average OpenAI Toxicity Scores by Keyword Category

Category	Hate Speech	Harassment Speech	Violent Speech
Locations	0.023 (0.00004)	0.072 (0.00010)	0.015 (0.00004)
Politicians	0.013 (0.00002)	0.082 (0.00008)	0.013 (0.00003)
Issues	0.016 (0.00004)	0.062 (0.00010)	0.017 (0.00004)
Political Terms	0.012 (0.00003)	0.065 (0.00010)	0.009 (0.00003)
Institutions	0.012 (0.00004)	0.049 (0.00011)	0.012 (0.00004)
Processes	0.008 (0.00002)	0.036 (0.00006)	0.006 (0.00002)

Note: Mean toxicity scores from OpenAI's moderation endpoint are reported with standard errors in parentheses.

We disaggregated the political tweets dataset into six groups of keyword categories: institutions, issues, locations, political terms (including offices and terms related to political processes), politicians, and processes. There is substantial variation in OpenAI's toxicity values across these categories (see Appendix D). As shown in Table 1, tweets that include location keywords exhibit the highest average *hate* speech, tweets with politician names have the highest *harassment* scores, and tweets identified based on issues had the highest mean *violent* content scores. Descriptively, these results suggest that toxicity is prevalent in political tweets on a variety of topics and not unique to a subset of political discourse. Process-related terms showed the lowest amount of toxicity across all categories.

Likewise, using the same six groups of keyword categories, we found similar variation in Mistral's toxicity values across the different categories (see Table 2). Unlike OpenAI's evaluation, tweets that include politician names exhibit the highest average *hate and harassment* scores, while tweets identified based on issues had the highest *violent* scores. The pattern for process-related terms remained consistent, showing the lowest toxicity levels across both measured categories. Institutional keywords demonstrated moderate toxicity scores, falling below political terms but above processes. Aligning with past research, the Mistral endpoint returned higher scores across both *hate and harassment* speech and *violent* speech categories compared to OpenAI's metrics, as well as a larger difference between political tweets and the baseline randomly sampled tweets. These cross-model findings reinforce that toxicity patterns in political discourse vary by topic category, with certain political subjects consistently provoking more toxic language than others, regardless of the evaluation system employed.

Table 2: Average Mistral Toxicity Scores by Keyword Category

Category	Hate & Harassment	Violent Speech
Locations	0.104 (0.00013)	0.042 (0.00008)
Politicians	0.143 (0.00012)	0.034 (0.00006)
Issues	0.102 (0.00014)	0.048 (0.00009)
Political Terms	0.118 (0.00016)	0.029 (0.00007)
Institutions	0.072 (0.00016)	0.042 (0.00011)
Processes	0.056 (0.00008)	0.018 (0.00004)

Note: Mean toxicity scores from Mistral's evaluation are reported with standard errors in parentheses.

3.3 Differences by Top Political Keywords

Among the top 10 most common political keywords (see Table 3), tweets including the term `America` had the highest levels of *hate* speech, followed by `China` and `Obama`, with the latter showing the highest levels of *harassment*. Tweets containing `War` showed elevated levels of *violent* content. Process-related keywords such as `Vote/Voting` showed the

Table 3: Average OpenAI Toxicity Scores by Top 10 Keywords

Keyword *Frequency*	**Hate Score**	**Harassment Score**	**Violence Score**
America *700,320*	0.035 (0.00013)	0.109 (0.00029)	0.017 (0.00010)
China *403,076*	0.023 (0.00015)	0.070 (0.00031)	0.014 (0.00012)
Obama *768,767*	0.022 (0.00009)	0.123 (0.00028)	0.014 (0.00008)
War *837,190*	0.020 (0.00010)	0.062 (0.00023)	0.031 (0.00013)
Government *605,434*	0.019 (0.00009)	0.068 (0.00024)	0.012 (0.00008)
Law *634,739*	0.019 (0.00010)	0.076 (0.00026)	0.016 (0.00010)
Trump *2,641,931*	0.015 (0.00004)	0.088 (0.00013)	0.014 (0.00005)
President *829,427*	0.010 (0.00005)	0.058 (0.00018)	0.009 (0.00006)
National *640,944*	0.007 (0.00005)	0.034 (0.00016)	0.006 (0.00005)
Vote/Voting *2,481,230*	0.005 (0.00002)	0.024 (0.00007)	0.005 (0.00002)

Note: Mean toxicity scores for the top 10 keywords are reported with standard errors in parentheses. Keywords ordered by average hate speech scores.

Table 4: Average Mistral Toxicity Scores by Top 10 Keywords

Keyword *Frequency*	**Hate Score**	**Violence Score**
Obama *768,767*	0.185 (0.00036)	0.039 (0.00015)
America *700,320*	0.166 (0.00037)	0.037 (0.00016)
Trump *2,641,931*	0.152 (0.00019)	0.041 (0.00010)
President *829,427*	0.106 (0.00027)	0.033 (0.00014)
China *403,076*	0.104 (0.00040)	0.030 (0.00020)
Law *634,739*	0.103 (0.00031)	0.041 (0.00018)
Government *605,434*	0.100 (0.00031)	0.039 (0.00018)
War *837,190*	0.083 (0.00025)	0.107 (0.00026)
National *640,944*	0.061 (0.00025)	0.020 (0.00013)
Voting *2,481,230*	0.035 (0.00009)	0.013 (0.00004)

Note: Mean Mistral toxicity scores for the top 10 keywords are reported with standard errors in parentheses. Keywords ordered by average hate speech scores.

lowest toxicity scores across all dimensions, while institutional terms like `President` and `National` demonstrated moderate toxicity levels below the average for political tweets.

Turning to the Mistral results for the top 10 most common political keywords (see Table 4), tweets containing the term `Obama` exhibited the highest levels of hate speech with a score of 0.185, followed closely by `America` (0.166) and `Trump` (0.152). Similar to the OpenAI results, `War` demonstrated the highest violence score (0.107), substantially exceeding all other keywords in this dimension. Tweets containing `Law` showed the second highest violence score (0.0411), closely followed by `Trump` (0.0407). Similar to the OpenAI findings, process-related terms like `Voting` consistently displayed the lowest toxicity scores across both hate speech (0.0345) and violence (0.0126) categories, while institutional terms such as `National` maintained relatively moderate toxicity levels.

3.4 The Relationship Between Tweet Toxicity and Account Reach on Twitter

In order to investigate whether accounts that produce toxic content on Twitter differ in their network reach compared to those producing non-toxic content, we conducted regression analyses examining how toxicity scores predict two key measures of account reach: *follower* count (a measure of audience size) and *friend* count (a measure of network connectivity).

Table 5 and Table 6 present the results of mixed-effects regression models examining these relationships using OpenAI's and Mistral's toxicity metrics, respectively, with year included as a random effect (treated as a categorical factor) to account for temporal dependencies in the data.[9] The results reveal complex relationships between different types of toxicity and account reach metrics.

When examining OpenAI's toxicity scores (Table 5), we observe that accounts posting content with higher *hate* speech scores tend to have significantly more *followers* ($\beta = 2685.9, p = .004$), suggesting that hateful content may attract larger audiences. In contrast, accounts posting content with higher *harassment* scores have substantially fewer *followers* ($\beta = -8256.1, p < .001$), indicating a negative relationship between harassing

[9]The results are nearly identical when using a fixed effect for year.

	Followers		Friends	
	Est.	*p*	Est.	*p*
(Intercept)	6840.6	<.001	1707.0	<.001
Hate	2685.9	.004	-41.4	.237
Harassment	-8256.1	<.001	-131.2	<.001
Violence	-2803.9	<.001	-167.7	<.001
Random effects				
Year		✓		✓
Observations	24,936,001		23,727,439	

Table 5: Regression analysis quantifying associations between OpenAI content moderation metrics (*hate*, *harassment*, *violence* scores) and Twitter network variables (*follower count*, *friend count*). Models with log-transformed user network variables can be found in the appendix.

content and audience size. Similarly, *violent* content is associated with reduced *follower* counts ($\beta = -2803.9, p < .001$).

For network connectivity, measured by *friend* count, both *harassment* and *violence* scores show statistically significant but directionally different relationships. Higher *harassment* scores are associated with fewer *friends* ($\beta = -131.2, p < .001$). Similarly, higher *violence* scores correlate with more *friends* ($\beta = 167.7, p < .001$). *Hate* speech scores show no significant relationship with *friend* count.

The Mistral model results (Table 6) demonstrate a pattern that partially aligns with previous observations while exhibiting distinct departures in several key aspects. Accounts posting content with higher combined *hate and harassment* scores have substantially fewer *followers* ($\beta = -5939.37, p < .001$), aligning with the *harassment* finding from the OpenAI model but contradicting the positive relationship between *hate* speech and *follower* count. *Violence* scores show a similar negative relationship with *follower* count ($\beta = -3576.90, p < .001$) as observed in the OpenAI results.

For *friend* counts, the Mistral results indicate that accounts posting content with higher *hate and harassment* scores were associated with higher *friend* count ($\beta = 42.85, p < .001$), while those posting content with higher *violence* scores were associated with a lower *friends* count ($\beta = -163.99, p < .001$). These findings partially contradict the OpenAI results, particularly regarding *violent* content's relationship with *friend* count.

It is worth noting that toxicity scores for different dimensions are correlated, potentially complicating the interpretation of these regression results. However, all correlations between toxicity measures were below 0.35 (except for the correlation between OpenAI's *hate* and *harassment* values at 0.65, which is still moderate). This suggests that while there is some overlap between different toxicity dimensions, they are largely capturing distinct aspects of problematic content. The variance inflation factor (VIF) for each variable in every model was less than 2, indicating that multicollinearity is not a concern in our analyses. Further, we also conducted analyses using log-transformed *friends* and *followers* counts as these variables are highly skewed due to their nature as count data, which violates the normality assumption of linear regression. The log transformation was applied to address this skewness and better meet model assumptions, but the results remained largely the same as our primary analyses. The two main differences are apparent in the regression results based on the OpenAI endpoint values: (1) The association between *violent* speech and *follower* count is still negative but no longer statistically significant, and (2) the association between *hate* and *friend* count is positive but also statistically significant, which aligns with the association between *hate* speech and *follower* count, which further validates our findings. The Mistral results were robust to the log transformation. See the appendix for the full model results using the log-transformed user metrics.

These findings suggest that different forms of toxic political discourse are associated with distinct patterns of network reach on Twitter. While *hate* speech (as measured by OpenAI) may be associated with larger audiences, *harassment* and *violent* content generally correlate with reduced audience size. The relationship between toxicity and network connectivity (*friend* count) shows more variability across toxicity types and measurement models, suggesting a complex relationship between toxic content production and social network formation on the platform.

4 Conclusion

Social media networks have the potential to function as a "public square" for productive political deliberation. However, the capacity of social networks to facilitate political participation depends on the type of conversations taking place. We assessed more than 20M political tweets and 20M randomly selected English-language tweets to gauge

	Followers		Friends	
	Est.	*p*	Est.	*p*
(Intercept)	7065.73	<.001	1701.88	<.001
Hate / Har.	-5939.37	<.001	42.85	<.001
Violence	-3576.90	<.001	-163.99	<.001
Random effects				
Year		✓		✓
Observations	24,936,001		23,727,439	

Table 6: Regression analysis quantifying associations between Mistral content moderation metrics (*hate and harassment* and *violence* scores) and Twitter network variables (*follower count*, *friend count*). Models with log-transformed user network variables can be found in the appendix.

the prevalence of toxicity over time on Twitter. Our findings show that toxic speech is more common in political posts than English-language tweets at large. Our measurement strategy disaggregated toxicity into three dimensions—hate, harassment, and violent speech—and we show that harassment (or hate/harassment, when measured using Mistral's moderation endpoint) became more prevalent over time in political tweets compared to the broader discourse on Twitter. The amount of hate and violent speech remained relatively consistent over time in both political tweets and tweets at large. The highest sustained levels of toxicity in political speech appeared in 2020, following a divisive political election and the onset of a global pandemic.

Our findings lend credence to the idea that increasing concerns about political polarization on social media: Political discourse did, indeed, become more toxic over time. Past work suggests that this rising toxicity may cause some users to opt out of political discussions (Settle, 2018), potentially contributing to a vicious cycle of toxicity and selection that serves to increase the share of political discourse on social media that is dominated by toxic users.

Overall, the prevalence of toxic content in political tweets was low. Using the binary classifications from OpenAI's MODERATION endpoint, we find roughly 99% of political tweets had scores below OpenAI's threshold for violence (0.58) and hate (0.4) throughout the period studied (see Figure 6). The vast majority of tweets consistently fell below these thresholds, with only approximately 0.2-0.8% exceeding OpenAI's violence threshold and 0.4-1.5% exceeding the hate threshold, depending on the time period. Harassing tweets, however, were much more prevalent. Between 4-9% of political tweets exceeded the harassment threshold (0.44), with a notable increase during 2019-2020 when harassment peaked at nearly 9%.[10] The higher prevalence of harassment in political discourse compared to hate and violent speech reflects in part the ambiguity of the concept. Whereas violent and hate language are unambiguously prohibited by most networks' content policies and will be removed, harassment poses a more complex challenge. Harassment includes personal attacks and insults targeting individuals or groups, which are more common in online debates; hate speech requires explicit bias against protected groups, and violent speech involves direct depictions of harm. Because the prevalence of harassment on social media networks is a general social concern, future descriptive studies should both probe the measurement tools for harassment language and offer more nuanced descriptive accounts of how prevalent different manifestations of this concept are in online political discourse.

Our regression analysis of the relationship between toxicity and account reach reveals intriguing patterns that warrant further exploration. Accounts posting content with higher hate speech scores (as measured by OpenAI) tend to have significantly more followers and might also have more friends, while those posting harassing or violent content generally have fewer followers. This suggests a complex relationship between different forms of toxicity and audience engagement. These patterns may reflect a form of strategic behavior where certain types of toxic content are deployed to build following, while other forms alienate potential audience members. Future work should explore how these dynamics relate to political polarization and the potential for toxic political discourse to become normalized within online communities. The Mistral model results show somewhat different patterns, with both hate/harassment and violent content negatively associated with follower counts. However, this discrepancy could be a function of Mistral combining hate and harassment into a single metric; if decomposed, we might find similar positive associations for hate speech as observed in the OpenAI results.

[10] Comparable results are seen when using Mistral's *hate and harassment* binary classifications.

5 Limitations

Our study faces several methodological limitations. First, while we employed two state-of-the-art toxicity detection systems (OpenAI's OMNI-MODERATION endpoint and Mistral's MODERATION model) to enhance methodological robustness, these models may not fully capture the contextual nuances of political discourse. The variation in how these systems identify and score different dimensions of toxicity reflects broader challenges in computational content analysis of subjective phenomena. Though our dual-model approach provides stronger validation than single-model studies, the lack of standardized operational definitions across AI classification systems remains a significant constraint. Different training data compositions, mathematical representations of toxicity constructs, and algorithmic approaches to feature extraction all influence how these systems classify content. These methodological challenges underscore the importance of continued refinement in computational approaches to measuring toxic speech, particularly when examining politically charged language where context and intent significantly impact interpretation.

Building on these technical limitations, our approach also faces broader interpretive challenges. The inherently subjective and culturally contingent nature of toxicity means that automated scoring systems have fundamental precision limits, especially with rapidly evolving online vernacular. Political discourse presents particular difficulties as rhetorical strategies, ideological framing, and domain-specific knowledge significantly shape how content should be interpreted. To address these limitations, future research could implement mixed-methods approaches that integrate human annotation on strategically sampled tweet subsets, allowing researchers to validate machine classifications and investigate contexts where model predictions diverge from human judgment.

Finally, our keyword sampling methodology, while comprehensive, may exhibit both false positive and false negative classification errors, which may miss relevant political discourse or erroneously include non-political content containing our designated keywords. Although we implemented a rigorous multi-stage methodology for keyword generation and validation to mitigate these concerns, subsequent research should further refine the political discourse identification protocols. Additionally, our dataset's temporal boundary (November 2022) precludes analysis of more contemporary trends in political toxicity. Finally, as Twitter's user demographic composition and content moderation policies evolved throughout our study period, observed fluctuations in toxicity metrics may partially reflect platform-level governance interventions rather than organic shifts in political discourse dynamics.

References

Neil Fasching and Yphtach Lelkes. 2025. Model-dependent moderation: Inconsistencies in hate speech detection across llm-based systems. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics. Forthcoming.

Christian Fuchs. 2015. Social media and the public sphere. In *Culture and economy in the age of social media*, pages 315–372. Routledge.

Puck Guldemond, Andreu Casas Salleras, and Mariken Van der Velden. 2022. Fueling toxicity? studying deceitful opinion leaders and behavioral changes of their followers. *Politics and Governance*, 10(4):336–348.

Rhett Hanscom, Tamara Silbergleit Lehman, Qin Lv, and Shivakant Mishra. 2024. The toxicity phenomenon across social media. *arXiv preprint arXiv:2410.21589*.

Gabriela Juncosa, Taha Yasseri, Julia Koltai, and Gerardo Iniguez. 2024. Toxic behavior silences online political conversations. *arXiv preprint arXiv:2412.05741*.

Jin Woo Kim, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, 71(6):922–946.

Gary King, Benjamin Schneer, and Ariel White. 2017. How the news media activate public expression and influence national agendas. *Science*, 358(6364):776–780.

Svetlana Kiritchenko, Georgina Curto Rex, Isar Nejadgholi, and Kathleen C. Fraser. 2023. Aporophobia: An overlooked type of toxic language targeting the poor. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 113–125, Toronto, Canada. Association for Computational Linguistics.

Samara Klar and Yanna Krupnikov. 2016. *Independent politics*. Cambridge University Press.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic

content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318.

Michalis Mamakos and Eli J Finkel. 2023. The social media discourse of engaged partisans is toxic even when politics are irrelevant. *PNAS nexus*, 2(10):pgad325.

Amy Mitchell, Mark Jurkowitz, J Baxter Oliphant, and Elisa Shearer. 2020. Americans who mainly get their news on social media are less engaged, less knowledgeable. *Pew Research Center*, 30.

Kevin Munger. 2023. Temporal validity as meta-science. *Research & Politics*, 10(3):20531680231187271.

Jaime E Settle. 2018. *Frenemies: How social media polarizes America*. Cambridge University Press.

Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

Elizabeth Suhay, Emily Bello-Pardo, and Brianna Maurer. 2018. The polarizing effects of online partisan criticism: Evidence from two experiments. *The International Journal of Press/Politics*, 23(1):95–115.

Wentao Xu. 2024. Characterization of political polarized users attacked by language toxicity on twitter. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, pages 185–189.

A Political Keywords

The keyword creation process involved a multi-step approach to ensure comprehensive coverage of political discourse on Twitter. First, we used the New York Times (NYT) Archive API to extract all article headlines published between 2012 and 2023. From these headlines, we generated n-grams and trained a logistic regression model to predict whether a headline belonged to the "Politics" section. We then identified the top 5% of n-grams most strongly associated with political content and manually reviewed each for relevance. This step provided an initial set of keywords related to political events, policies, and broader political discourse.

Next, we employed GPT-4 and DeepSeek V3 to classify a random sample of 2 million tweets as political or non-political. Using the tweets classified as political by both models, we trained another logistic regression model to identify the words that are most predictive of political content. This step allowed us to expand our keyword list with terms that were highly indicative of political discourse on Twitter. To ensure our keyword list was as comprehensive as possible, we selected a sample of tweets that were classified as political by GPT-4 or DeepSeek V3 but did not contain any of the keywords from our initial list. For these tweets, we again applied logistic regression to capture the most predictive words and also conducted a manual review of a subset of those tweets to identify more keywords. This process not only led to the inclusion of additional keywords but also prompted us to add the names and Twitter handles of all U.S. presidents and vice presidents who held office since 2012.

Keywords: Politicians

- Trump, Obama, Clinton, Biden, Abbott, Abrams, Klobuchar, Cuomo, Andrew Yang, Bannon, Barr, Barrett, Bernie, Blinken, Boehner, Bolton, George Bush, Jeb Bush, Buttigieg, Carly Fiorina, Ben Carson, Cheney, Chris Christie, Ted Cruz, Feinstein, Franken, Gillibrand, Gingrich, Ginsburg, Giuliani, Gorsuch, Graham, Hagel, Nikki Haley, Harris, Harry Reid, Ilhan Omar, Jan Panel, John Kelly, John Lewis, Kaine, Kasich, Kavanaugh, Kennedy, Khashoggi, Lynch, Malley, Manafort, Manchin, Martin Malley, Mattis, McCain, McCarthy, McConnell, Bloomberg, Mnuchin, Mueller, Nunes, Palin, Pelosi, Pence, Pompeo, Paul Ryan, Sanders, Santorum, Scalia, Schiff, Schumer, Jeff Sessions, Scott Walker, Warnock, Warren, John Roberts, Roger Stone, Romney, @barackobama, @mittromney, @hillaryclinton, @realdonaldtrump, @joebiden, @kamalaharris, @govtimwalz, @jdvance, @timkaine, @speakerryan, @mike_pence

Keywords: General political terms

- POTUS, GOP, Congressman, Congresswomen, DJT, Secretary, Amendment, Policies, Minister, President, Administration, Aide, Aides, Ambassador, Candidate, Conservative, Delegate, Diplomat, Evangelical, First Lady, Lobbying, Lobbyist, Nominee, Partisan, Progressives, Political Party, Political Parties, Presidency

Keywords: Institutions

- Congress, Senate, White House, Whitehouse, CIA, FBI, Legislative, State Dept, State Department, Supreme Court, #supreme court,

Federal, FEMA, #FEMA, Homeland Security, Military, Navy, Air Force, Cabinet, Committee, Agencies, Amtrak, World Bank

Keywords: Issues

- Immigration, Climate Change, #climate, Global Warming, Healthcare, Taxes, Economy, Black Lives Matter, MAGA, Stop the Steal, Law, War, National, Abortion, Agriculture, Criminal Justice, DACA, #DACA, Deportation, Drug Prices, Food Stamps, Health Law, Medicare, NAFTA, National Security, Stock Market, Transportation, Travel Ban, Treasury, Infrastructure

Keywords: Processes

- Election, #election, Politic, Policy, Vote, #vote, Voting, Appointed, Political Appointment, Approval Rating, Assassination, Battleground, Bipartisan, Caucus, Debate, Deficit, Disinformation, Donation, Drone Strike, Electoral College, Endorsement, Espionage, Exit Polls, Gerrymandering, Hearings, Impeachment, Briefing, Inaugural, Inauguration, Investigation, Investigator, Nomination, Primaries, Recount, Retaliations, Special Counsel, Spending Bill, Subpoenas, Super PAC, Super Tuesday, Surveillance, Swing State, Sworn In, Tariff, Taxpayers, Tech Giant, Ballot, Campaign, Convention, CPAC, Establishment, Filibuster, First Black, First Draft, Government, Gridlock, House Approves, House Passes, House Seat, Midterm, Negotiations, News Conference, News Media, Oath of Office, Oligarch, Opposition To, PAC, #PAC, Pardons, Polling, Watergate, Withdraws From

Keywords: Locations

- America, China, Europe, European, Mexico, South Korea, Arizona, Capitol, Iowa, Iran, Iraq, Israel, Maine, Maryland, Massachusetts, Nevada, New Hampshire, New York, North Dakota, Ohio, Pennsylvania, Rhode Island, South Dakota, Texas, Ukraine, Virginia, Wyoming, Guantánamo, Guantanamo

B Comparison of political and random tweets

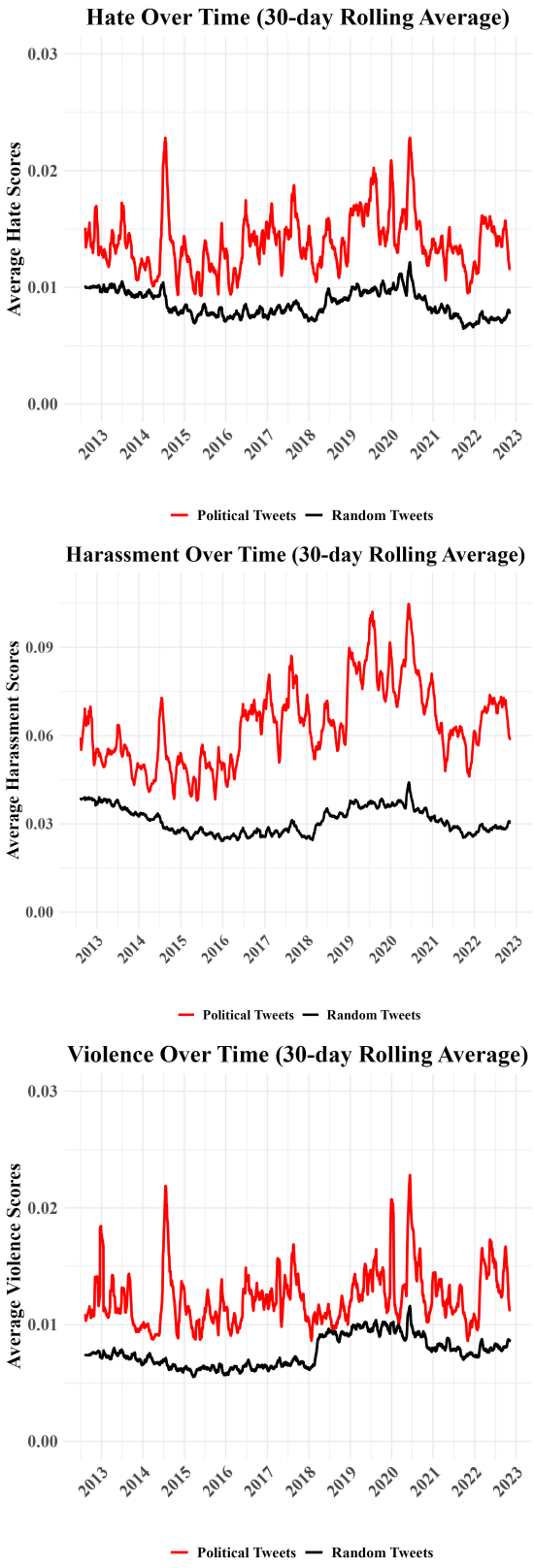

Figure 4: Comparison of political tweets versus all Twitter discussion across three dimensions of toxicity: hate, harassment, and violence measured with the OpenAI endpoint.

B.1 Alternative political classifications

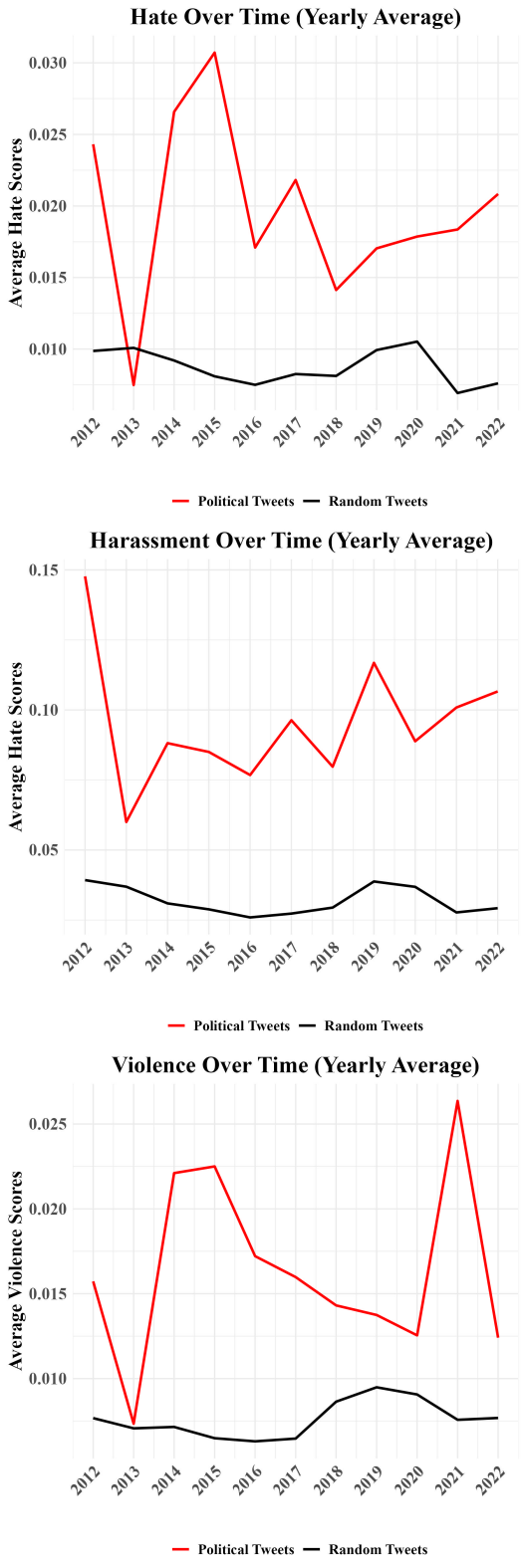

Figure 5: Replication of Figure 4 using GPT-4O-MINI for political classifications rather than keywords. The sample consists of 10,000 randomly sampled from each year. These data are much more sparse than the main sample, we aggregate within year.

C Distribution of Toxicity Scores

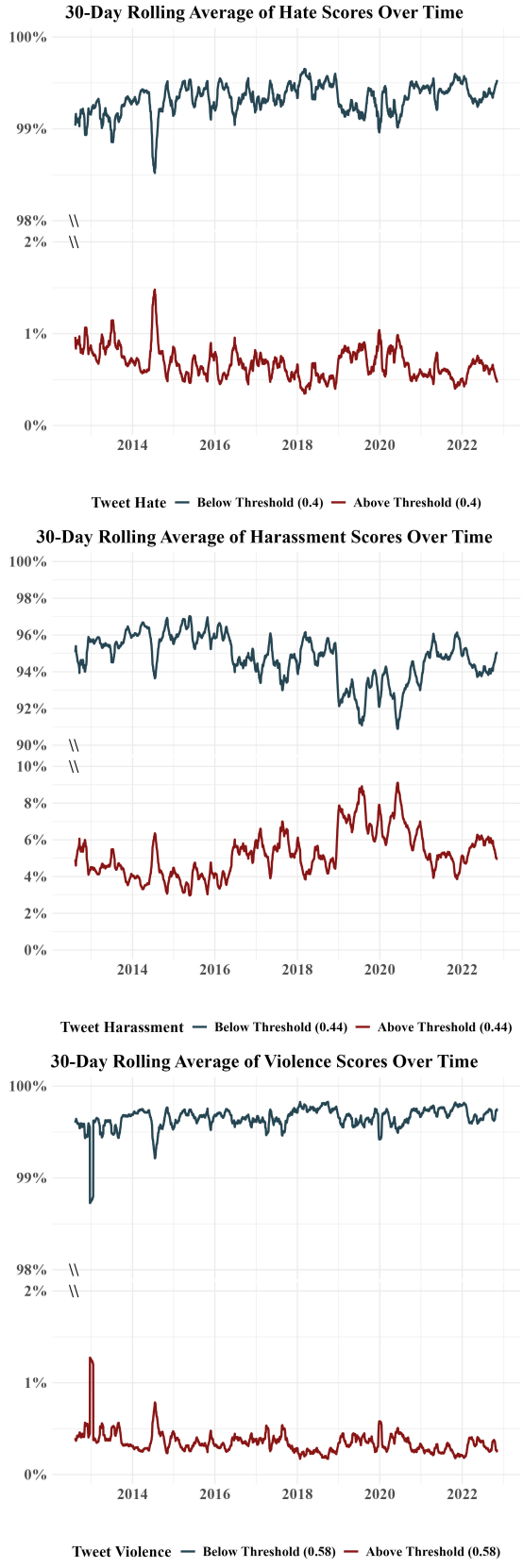

Figure 6: Rolling average of tweets exceeding toxicity thresholds measured with the OpenAI endpoint (2013-2023).

D Toxicity by Keyword Category

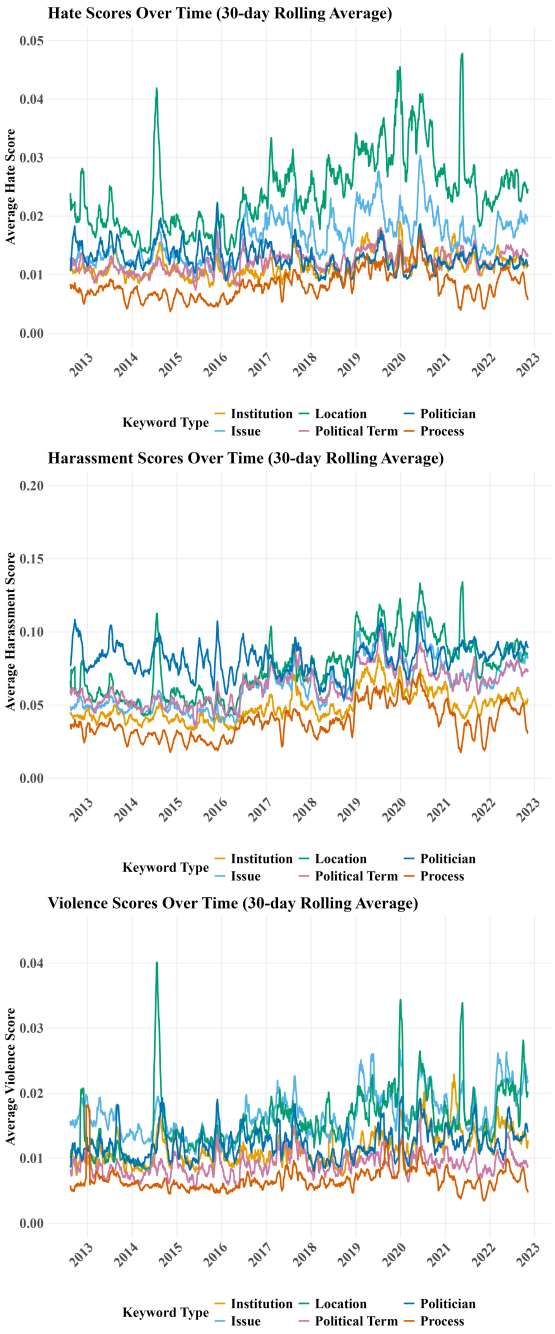

Figure 7: Toxicity scores for six keyword categories over time across three toxicity dimensions measured with the OpenAI endpoint: hate, harassment, and violence.

E Human Validation

Category	Acc.	Prec.	Rec.	F1	Kappa
Harassment	0.94	0.44	0.57	0.50	0.46
Hate	0.99	0.80	0.27	0.40	0.40
Violence	0.96	0.83	0.50	0.62	0.61

Table 7: Performance metrics by category. The Human validation task is based on a sample of 1000 tweets which was stratified to include 500 tweets from the random sample and 500 tweets from the keyword sample.

	Harass.	Hate	Violence
Political (keyword)	0.0370**	0.0054**	0.0053**
	(0.0048)	(0.0015)	(0.0012)
Fixed-effects			
Year	X	X	X
Fit statistics			
Obs.	109,981	109,981	109,981
R^2	0.00363	0.00079	0.00070
Within R^2	0.00212	0.00026	0.00027

Clustered (year) std-errors in parentheses
*Signif. Codes: **: 0.01, *: 0.05*

Table 8: Regression analysis predicting harassment, hate, and violence score (measured with the OpenAI endpoint) based on whether a tweet is classified as political using the lexicon-based approach, controlling for year.

	Harass.	Hate	Violence
Political (GPT-4o)	0.0656**	0.0103**	0.0080**
	(0.0047)	(0.0012)	(0.0017)
Fixed-effects			
Year	X	X	X
Fit statistics			
Obs.	109,981	109,981	109,981
R^2	0.00770	0.00140	0.00100
Within R^2	0.00620	0.00088	0.00057

Clustered (year) std-errors in parentheses
*Signif. Codes: **: 0.01, *: 0.05*

Table 9: Regression analysis predicting harassment, hate, and violence score (measured with the OpenAI endpoint) based on whether a tweet is classified as political using GPT-4O-MINI.

	Political (keyword)	Political (GPT)
Variables		
Hate score	-0.0025	-0.0717***
	(0.0064)	(0.0198)
Harassment score	0.0066	0.0853***
	(0.0043)	(0.0193)
Violence score	0.0033	0.0053
	(0.0044)	(0.0112)
Keyword: Politician	0.6537***	0.7227***
	(0.0163)	(0.0175)

Continued in Table 11

Table 10: Fixed-effects OLS regressions predicting whether a tweet is political (part 1).

	Political (keyword)	Political (GPT)
Keyword: Political term	0.5753***	0.4241***
	(0.0154)	(0.0358)
Keyword: Pol. inst.	0.6193***	0.3967***
	(0.0321)	(0.0336)
Keyword: Pol. issue	0.7558***	0.2415***
	(0.0394)	(0.0243)
Keyword: Pol. process	0.7958***	0.3200***
	(0.0253)	(0.0247)
Keyword: Location	0.7876***	0.2129***
	(0.0241)	(0.0249)
Fixed-effects		
Year	X	X
Fit statistics		
Obs.	109,981	109,981
R^2	0.867	0.344
Within R^2	0.866	0.339

Clustered (year) std-errors in parentheses
*Signif.: ***: 0.01, **: 0.05, *: 0.1*

Table 11: Fixed-effects OLS regressions predicting whether a tweet is political (part 2).

The human validation task was conducted to assess the reliability of our automated toxicity detection. We randomly sampled 1000 tweets, with equal representation from our political and random datasets. Human annotators labeled these tweets for harassment, hate speech, and violent content, allowing us to calculate agreement metrics between human judgments and our automated classification systems.

F Regression Results with Log-Transformed DV

	Log of Followers		Log of Friends	
	Est.	*p*	Est.	*p*
Hate	0.69	<.001	0.18	<.001
Harassment	-0.39	<.001	-0.03	<.001
Violence	-0.01	.686	-0.11	<.001
Random effect				
Year	X		X	
Observations	24,936,001		23,727,439	

Table 12: Regression analysis quantifying associations between OpenAI content moderation metrics (hate, harassment, violence scores) and log-transformed Twitter network variables (follower count, friend count).

	Followers		Friends	
	Est.	*p*	Est.	*p*
Hate and Harass.	-0.05	<.001	0.17	<.001
Violence	-0.13	<.001	-0.05	<.001
Random effect				
Year	X		X	
Observations	24,936,001		23,727,439	

Table 13: Regression analysis quantifying associations between Mistral content moderation metrics (*hate and harassment* and *violence* scores) and log-transformed Twitter network variables (follower count, friend count).