Insult Politics in the Age of Social Media*

Job Market Paper

Elliot Motte†

October 27, 2025

Click here for the latest version

Abstract

This paper studies the production of offensive content by politicians on Twitter/X. Applying state-of-the-art AI-based methods to the universe of tweets posted by members of the U.S. Congress (2017–2022), I measure the returns to offensive communication both in terms of online engagement and electoral support. I document that posting toxic tweets generates a sizable engagement premium and that these gains decrease with politicians' baseline toxicity. To examine how voters respond to offensive speech, I link the timing of toxic tweets to a large-scale electoral survey. Using an event study design around days marked by sharp spikes in representatives' toxic tweeting activity, I find that voting intentions for the politician increase in the following week. The increase is concentrated among ideologically aligned voters, while opponents display lower electoral support, widening the partisan voting gap. These findings highlight the presence of both online and electoral incentives to the production of offensive communication, which comes at the expense of growing polarization.

*I would like to thank my doctoral advisors Ruben Durante and Maria Petrova for their guidance and support throughout this project. I am also grateful towards Marie Beigelman, Julia Cagé, Ruben Enikolopov, Rosa Ferrer, Roberto Galbiati, Gianmarco León-Ciliotta, Mike McRae, Tom Pesso, Paul Seabright, Andrea Tesei, Stéphane Wolton, Ekaterina Zhuravskaya as well as participants of the CEPR Paris Symposium, the Digital Economy Workshop Berlin, the Monash-Paris-Warwick-Zurich Text-as-Data Workshop, the TSE Doctoral Workshop in the Economics of Digitization, the Catalan Economic Society Congress and the IEB Workshop in Political Economy for helpful comments and suggestions.

†Universitat Pompeu Fabra. Email: elliot.motte@upf.edu

1 Introduction

Offensive rhetoric has become a defining feature of contemporary political discourse. Hostile exchanges between politicians, insults directed at opponents, and demeaning comments about political or social groups now unfold daily on social media platforms, despite broad public disapproval.[1] At the same time, social media has transformed the ways politicians communicate with voters (Jungherr et al., 2020). These platforms have lowered the cost of direct communication, allowing politicians to bypass traditional media intermediaries and speak to the electorate continuously rather than episodically (Tumasjan et al., 2010; Leung and Yildirim, 2020). While the resulting information environment is richer, it also means that voter attention becomes scarcer, creating a market where politicians compete for visibility. This competition is strengthened by platform algorithms designed to reward content that generates online engagement (Narayanan, 2023; Beknazar-Yuzbashev et al., 2024). Together, these transformations may alter the incentives shaping politicians' use of offensive speech, which have yet to be systematically studied.

In the online attention economy, choosing offensive rhetoric over other communication strategies entails a trade-off. On the one hand, offensive speech may generate substantial visibility (Auter and Fine, 2016; Messing et al., 2017; Dai and Kustov, 2022), which could predict future campaign contributions or work towards building a large audience (Boken et al., 2023). Electorally, this strategy may also mobilize a politician's core supporters or depress opponent turnout (Ballard et al., 2023). On the other hand, such a strategy risks alienating moderate voters and galvanizing the opposition (Ansolabehere et al., 1994; Frimer and Skitka, 2018). While the new incentives created by the digital economy have received some attention in other areas (Eckles et al., 2016; Burtch et al., 2022; Mummalaneni et al., 2022), we know little about how they shape politicians' choice to use offensive rhetoric.

This paper studies the drivers of politicians' offensive online speech by measuring the returns to U.S. congress members' toxic communication on Twitter/X both in terms of public attention and in terms of electoral support. To do so, I construct a dataset of 3 million tweets posted by members of Congress between 2017 and 2022, enriched with measures of textual style extracted with AI-based methods, and linked to large-scale electoral survey responses. I document three main findings. First, toxic tweets generate markedly more engagement than non-toxic tweets, even after accounting for tweets' stylistic and content differences as well as differences between politicians. Second, favorable voting intentions increase in the week following a spike in toxic communication, and this increase depends on voters' ideological proximity with the politician. Third, the size of these returns decreases with politicians' baseline toxicity, indicating that the value of offensive speech declines with overuse.

[1] Recent survey evidence indicates that 70% of Americans believe elected officials should avoid aggressive language (Pew Research Center, 2024).

The analysis relies on the universe of tweets posted by all members of the U.S. Congress between June 2017 and December 2022. I measure offensive communication using two complementary approaches. The primary measure is the continuous toxicity score from Perspective API that captures a "rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion", an AI-based measure used in industry and research (Müller and Schwarz, 2023a; Ederer et al., 2024; Beknazar-Yuzbashev et al., 2025; Kalra, 2025). To complement this, I use a Large Language Model (LLM) to code each tweet according to a custom definition of toxicity which more specifically captures uncivil speech in political contexts and explicitly excludes respectful forms of criticism, even when expressed strongly. Both measures are highly correlated and the results are robust to the choice of measure.

To quantify the engagement returns to offensive speech, I compare tweets produced by the same politician within the same week. This design abstracts from unobserved differences across politicians that may correlate both with toxic communication and online engagement such as popularity, as well as within-politician changes over time, such as local news shocks. To isolate offensive style from critical substance and other rhetorical devices, I further control for the presence and strength of criticism, the target of the attack, and the use of sarcasm or emotional appeals, all extracted through LLM-based annotations that I validate with the help of human coders. I employ additional text analysis methods to account for differences in tweet sentiment, topic and lexical diversity.

There are strong engagement returns to offensive communication. Toxic tweets receive about 17% more likes than non-toxic ones, representing a sizable increase given that the average tweet receives 750 likes.[2] This engagement premium is robust to (i) considering other common tweet engagement metrics such as retweets, replies or quote tweets, and (ii) additionally accounting for unobserved differences in tweet topic within politicians. Toxicity emerges as the single most powerful driver of engagement when voicing criticism, exceeding other rhetorical devices such as appealing to emotions or using sarcasm, irony or humor.

Crucially, I provide evidence for *decreasing engagement returns* to toxicity. Offensive tweets generate a strong engagement premium for politicians with below-median baseline toxicity, but no additional benefit for those who are already highly toxic.[3] This pattern holds after netting out other politician characteristics that could affect differences in engagement returns, ensuring that the decreasing returns are not driven by these confounders.[4] The presence of decreasing returns points towards two distinct equilibrium strategies. First, it

[2] A back-of-the-envelope calculation using the likes-to-donations elasticity from Boken et al. (2023) shows that this engagement boost translates to a $0.57 increase in small campaign contributions per toxic tweet. While modest in isolation, these financial returns highlight that the engagement premium is economically meaningful, reinforcing incentives for politicians to use offensive rhetoric.

[3] Baseline toxicity is computed as the average toxicity in the first six months of a congressional session, and these results follow from regressions estimated over the eighteen remaining months.

[4] This includes politician ideology, electoral safety, congressional seniority, baseline tweeting activity (popularity and number of tweets), and sociodemographic characteristics.

suggests that low-toxicity politicians behave optimally given the trade-off they face, using incivility sparingly to generate sporadic attention shocks. For these politicians, the perceived electoral or reputational costs of a sustained toxic communication strategy outweigh immediate online visibility benefits. Second, for politicians with high baseline toxicity, additional toxic tweets no longer generate extra engagement. In spite of this, these politicians may still derive alternative and longer-term benefits from a constant stream of offensive language, such as mobilizing core supporters, signaling ideological purity, or maintaining a large online audience that contributes to a high national profile. These expected benefits may outweigh their reduced tweet-level engagement gains, explaining why they continue producing offensive rhetoric. These distinct strategies highlight that a politician's chosen level of toxicity results from trading-off short-term engagement with other political goals.

I then turn to investigating the conditions that magnify or attenuate the link between toxicity and online engagement by analyzing the presence of heterogeneous returns along political, demographic and rhetorical dimensions. First, politicians with strong baseline electoral support experience the largest engagement gains, suggesting that electorally-secure politicians may afford offensive rhetoric for visibility motives. Second, returns to toxicity are larger for younger politicians, indicating systematic differences across demographic groups. Third, engagement returns are stronger for centrist politicians which could indicate an additional salience premium of offensive rhetoric not entirely captured by pre-existing differences in baseline toxicity. Fourth, toxicity directed at individuals yields the strongest engagement rewards, relatively to attacks targeted towards political entities, social groups or broader issues and policies.

Some of these differences in engagement returns align closely with differences in average toxicity between politicians, suggesting a link between online incentives and communication choices. For instance, safer and younger politicians are more toxic than their more contested or older colleagues. In addition, individuals are the most frequent targets of offensive rhetoric. These cross-sectional differences map well with the heterogeneous engagement returns over these dimensions. This suggests that online engagement incentives shape politicians' rhetorical choices. Other cross-sectional differences, however, are not fully captured by differences in engagementspotif returns. Politicians at the ideological extremes are more toxic than moderates, even though engagement gains are strongest for the latter. This suggests that politicians additionally respond to other incentives beyond online visibility.

To understand the broader set of incentives that enter politicians' trade-off, the second part of the paper measures the electoral returns associated to offensive communication. To do so, I link representatives' patterns of toxic communication on Twitter/X to Nationscape, a large scale electoral survey tracking respondents' voting intentions in House elections. This allows to conduct an event study analysis at the district-by-day level, comparing respondents

interviewed in the days immediately following an unusually high toxicity spike by their representative to similar respondents surveyed in the days just before. To ensure tight comparisons, analyses are performed on the sample of respondents interviewed within a narrow time window of their representative's toxic spikes. Identification hinges on the assumption that, within this window, the timing of survey interviews in a district is as-good-as-random relative to the representative's toxic tweeting behavior. This is supported by balance checks showing that respondents' sociodemographic characteristics remain stable around toxic spikes.

Unusually offensive rhetoric yields measurable electoral gains and contributes to widening the partisan divide in electoral support. Voting intentions for the representative increase by 2.4 percentage points on average in the week following a toxic spike, and return to pre-spike levels within a week. The stability of voting intentions in the days preceding toxic spikes alleviates concerns about reverse causality or anticipation effects. Further heterogeneity analyses reveal that the increase in voting intentions is most pronounced among aligned voters, consistent with aggressive political rhetoric swaying the most persuadable supporters. Conversely, strong ideological opponents become less likely to vote for the politician, indicating that toxic communication strengthens partisan polarization in voter support. Mirroring patterns found in the analysis of online engagement, I show that the electoral returns to toxic communication decrease with a politician's baseline toxicity.

Turning to the mechanisms underpinning the presence of electoral returns, I provide evidence that (i) toxicity needs to be sufficiently salient to induce changes in voting intentions, and that (ii) exposure to social media amplifies voter responses. First, voting intentions increase following toxic spikes that generate unusually high levels of online engagement. In addition, there is no discernible change following high but milder instances of toxicity.[5] These findings indicate that salience in the visibility and intensity of toxicity needs to be sufficiently strong to generate changes in voting intentions. Second, the increase in voting intentions is concentrated among voters who rely exclusively on social media for political news, while it is absent among those who consume news solely through television, pointing to social media as the transmission channel.

To strengthen the validity of these results, I conduct several exercises addressing potential identification challenges. First, I rule out that the findings are due to high-frequency, party-specific shocks that jointly affect politicians' offensive communication and voter support. I show that the results are unchanged when adding party-by-day fixed effects, which absorb any unobserved event such as major national or international news developments, policy announcements or legislative activity, or reactions to statements by high-profile politicians (presidents, party leaders, governors). Second, placebo tests estimating changes in voter support following randomly generated spike dates show that the observed increase in voting

[5]Such instances are defined as days when a politician's toxicity ranks in the top 5%-10% rather than the top 1% of their own historical distribution of toxicity.

intentions lies in the extreme right tail of the placebo distribution. This result shows that the presence of electoral returns is not due to random fluctuations in the data or to coincident events unrelated to politicians' toxicity and voting intentions. Finally, I test whether voters react to toxicity itself or to any other salient deviation in politicians' online communication by constructing spikes for alternative rhetorical features such as irony, emotional intensity, sentiment, or tweet engagement. Voting intentions remain unchanged following any of these alternative spikes, confirming that the electoral response is specific to offensive rhetoric.

This paper contributes to a growing literature on politicians' use of social media. One strand of this work shows that politicians' online activity has real-world consequences, from increasing campaign funding (Petrova et al., 2021; Boken et al., 2023) to fueling offline hate or bias (Müller and Schwarz, 2023b; Cao et al., 2023; Grosjean et al., 2023), or enabling the reallocation of resources to marginal voters (Bessone et al., 2022). Another strand focuses on the online environment itself, documenting that politicians' communication polarizes the online public debate (Zhang et al., 2025) and induces greater toxicity among users (Müller and Schwarz, 2023a). In contrast to these outcome-oriented perspectives, this paper centers on the supply side by examining the incentives that shape politicians' choice to employ disrespectful and uncivil speech. I quantify the private returns to toxic communication in both online and offline domains and show how their heterogeneity across political, demographic, and rhetorical dimensions helps explain cross-sectional variation in toxicity use. The paper is more closely related to Algan et al. (2025), who combine online and offline data to study how negative emotions, in particular anger, shape policy views. Here, I use LLM-based tweet annotations to isolate the role of toxicity from emotions and other rhetorical devices in driving engagement and voter support, an outcome which has received little attention in this literature. In doing so, I provide new evidence on how politicians combine the private incentives they face when supplying toxic communication.

This paper also contributes to a recent and rapidly expanding literature on toxic communication in online environments (Beknazar-Yuzbashev et al., 2024; Jiménez Durán et al., 2024; Beknazar-Yuzbashev et al., 2025; Kalra, 2025). The consensus that emerges from these studies is that exposure to toxicity increases engagement, highlighting the trade-off between user welfare and platform revenues, and that algorithmic design can shape both online discourse and offline outcomes. This body of work, however, largely considers toxicity through the lens of user behavior and platform incentives. In contrast, I examine toxicity as a strategic choice by professional politicians. By quantifying both its online and offline private returns, and by documenting how these returns vary across political, demographic, and rhetorical dimensions, I show how toxicity functions as a tool in politicians' communication strategies. Importantly, the results also reveal a distinct negative externality specific to the political realm: bursts of toxic speech polarize voters.

Finally, the paper contributes to the longstanding debate on the electoral effects of negative political communication. Decades of research have produced mixed conclusions: some studies argue that negativity demobilizes voters (Ansolabehere et al., 1994, 1999), others find mobilizing or null effects (Freedman and Goldstein, 1999; Djupe and Peterson, 2002; Brooks and Geer, 2007; Finkel and Geer, 1998; Lau and Pomper, 2004), while more recent work highlights conditional effects depending on timing, target, or candidate characteristics (Brooks, 2010; Krupnikov, 2011; Galasso et al., 2023). Here, I shift the focus to systematically analyzing when toxic speech pays off. Additionally, I broaden the set of relevant incentives driving offensive strategy to online engagement, which offers a more complete perspective on politicians' decision-making process. In doing so, the paper reframes the debate from whether negativity "works" on average to understanding the conditions under which incivility pays off.

The remainder of this paper is structured as follows. Section 2 describes the data sources used in this study and presents key empirical facts about politicians' supply of toxic communication. Section 3 quantifies the engagement returns to toxic communication and examines how they differ across political, demographic, and rhetorical dimensions. Section 4 studies how voter support shifts following unusually toxic communication by their representatives. Finally, section 5 concludes.

2 Data

Tweet dataset

This paper relies on a large publicly available dataset containing the universe of tweets posted by U.S. congress members' office and campaign accounts between June 21, 2017 and December 31st, 2022.[6] This dataset contains 4,011,034 unique tweets posted by 729 unique congress members. This covers virtually all 741 congress members serving through the 115th (2017-2018) to 117th (2021-2022) congressional sessions.[7] Retweets are removed from the data in order to restrict focus on content that is originally produced by politicians. This yields a total of 3,000,164 tweets.

Available information is restricted to the full text of the tweet, the exact time it was posted and the author's screen name. I enrich this dataset in two main ways. First, I collect the engagement metrics of each tweet in the dataset, i.e. the number of likes, retweets, quote tweets and replies. This step was performed in November and December 2023, once these metrics had arguably stabilized.[8] This information was successfully collected for 96% of the

[6]The raw data files can be downloaded on the congresstweets Github repository.

[7]There are 12 congress members for which no tweet is recorded in the dataset: 4 Democrats representatives, 6 Republican representatives and 2 Republican senators.

[8]While it is still technically possible for tweets to continue generating engagement several months after

dataset (2,876,440 tweets).[9] Summary statistics for these metrics are displayed in Table A2.

Second, I use a suite of state-of-the-art natural language processing methods to measure the substantive content and style of tweets, as will now be discussed.

Toxicity measures

This paper makes use of two complementary toxicity metrics to measure the offensiveness of politicians' tweets. My primary measure relies on Perspective API's deep learning toxicity detection model developed in collaboration with Google to determine tweets' toxicity score – a metric also used in Müller and Schwarz (2023a), Jiménez Durán et al. (2024), Beknazar-Yuzbashev et al. (2025) and Kalra (2025) to determine the toxicity of social media posts. For each tweet text, Perspective API (PAI) returns a continuous measure ranging between 0 and 1 which can be interpreted as the fraction of individuals who would find the text to be a "rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion". The underlying machine-learning model built by PAI follows a transformer architecture and was trained on millions of user-generated online text instances. These instances were manually annotated in-house according to the aforementioned definition of toxicity.[10] Perspective API is now recognized as a leading tool for toxicity detection both in research and in industry, with national news outlets such as the *New York Times*, the *Wall Street Journal* and online fora such as Reddit relying on their measures for content moderation.

One of the key advantages of the PAI score is its granularity. As a continuous measure, it is particularly well-suited for identifying moments of extreme deviation from a politician's baseline toxicity level, a step I perform to determine voter responses to extreme toxicity. To provide some intuition for how well PAI toxicity proxies for the intensity of toxicity, Table A1 contrasts examples of tweets with low versus high toxicity scores, including additional gradation in the strength of toxicity among high toxicity tweets. As toxicity grows, stronger offenses are characterized in the text. Typically, low toxicity tweets relate to routine social media communication by politicians (e.g., announcing local on-the-ground activities or policy achievements), or to the courteous expression of disagreement. Conversely, tweets with high levels of toxicity consist of verbal attacks against other politicians, groups of people or fierce outrage on issues or policy items.

While PAI provides a granular measure of intensity, its general-purpose nature presents three specific measurement challenges in the context of political communication. First, the measure is built to identify content "likely to make someone leave a discussion" which, by

being posted, prior research has documented that the engagement life-cycle of tweets is counted in terms of days, not months (Pfeffer et al., 2023).

[9] Failures are due to tweets that were deleted between their posting date and the time of retrieval.

[10] While there are no statistics available on the pool of human annotators, PAI resorts to online crowd-sourcing platforms such as FigureEight for annotation.

design, is *negatively* correlated with engagement. To the extent that this paper documents a *positive* relationship between toxicity and engagement, this mechanical negative association between PAI toxicity and engagement would downward bias the estimates reported in section 3, thus limiting this cause for concern. Second, despite being self-contained and compact, the definition may be perceived as excessively broad and imperfectly suited to track the nuances of toxicity produced by politicians. Third, the measure could be conflating offensiveness with other features of political communication used to generate online engagement or to reap more votes (e.g. civil expressions of criticism or negativity, or the use of specific emotions such as anger).

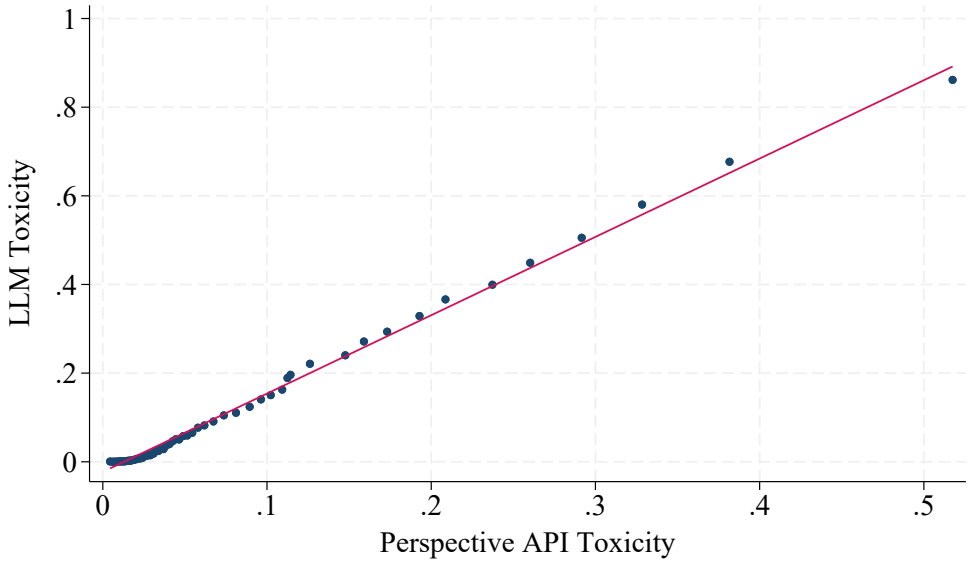
In order to address these challenges, I use a custom-built Large Language Model (LLM) annotation pipeline to extract the presence of specific features in the tweets' text. This includes classifying tweets as toxic using a detailed prompt designed to capture *political incivility*. A tweet is classified as toxic if it contains "disrespectful or aggressive language breaking norms of civil political debate". This includes insults, degrading comments, or vilification. Crucially, the prompt instructs the model not to classify civically expressed criticism as toxic, allowing for a cleaner separation between the substance of disagreement and the style of its delivery. The analysis relies on Deepseek-v3, which was selected as the best-performing model after a rigorous validation procedure that compares the performance of five leading LLMs against a human-annotated benchmark serving as the ground truth. The full prompt and details on the validation procedure are presented in Appendix C.1.

The primary purpose of this LLM-based binary measure of toxicity is to strengthen the validity of my findings by using an alternative measure tailored to instances of toxicity and incivility in political communication. Figure 1 provides a first step towards this goal by plotting the relationship between the two measures. The graph shows a strong, positive and monotonic relationship: as tweets' PAI toxicity score increases, so does their probability of being classified as toxic by the LLM. This provides systematic evidence for the intuition that higher PAI toxicity corresponds to stronger political incivility. This tight correlation brings additional confidence that PAI's continuous score serves as a reliable proxy for the intensity of toxic content.

Other Textual Features

To isolate the effect of toxicity from features of political communication that may be linked to engagement or electoral outcomes, I further leverage the LLM annotation pipeline to extract several other key dimensions from the tweets' text. This includes identifying the presence and strength of *criticism or disagreement*, a crucial step to distinguish the effect of toxic *style* from critical *substance*. For tweets containing criticism, the LLM also identifies the primary *target* of the attack, classifying it as an individual, a political entity, a social group,

Figure 1: Correlation between Perspective API Toxicity Score and LLM-Annotated Toxicity



Notes: The figure presents a binscatter plot of the LLM-annotated toxicity dummy (Y-axis) against the continuous Perspective API toxicity score (X-axis). The plot is constructed over 100 equal-sized bins of the PAI score. The solid red line shows the linear fit between the two variables.

or an issue. Additional stylistic features annotated via the LLM include the use of *sarcasm or irony*, the tweet's primary *rhetorical appeal* (cognitive or emotive), and the specific *emotion* conveyed in emotive tweets. Table A3 provides summary statistics for all LLM-annotated features. Notably, approximately 7.3% of tweets in the annotated sample are classified as containing political toxicity. Of the 34.6% of tweets expressing criticism or disagreement, the vast majority are directed at individuals (50.9%) or broader issues and policies (27.8%), while a much smaller fraction target social groups (2.4%). Political entities are the object of criticism or disagreement in 18.9% of critical tweets. About half of the tweets appeal to emotions (50.5%) and out of those, the dominant categories are anger (25.2%), hope (19.8%) and joy (19.6%). Sarcasm, irony or humor is rarely used as a stylistic device in politicians' online communication (5.8% of tweets). Appendix C.1 provides further details about the annotation methodology, including the prompts used for annotation, and a validation of the procedure against a human benchmark.

This set of LLM-generated variables is complemented by features extracted using more conventional natural language processing methods. I employ BERTopic to classify each tweet into one of 30 topics (Grootendorst, 2022), VADER for a continuous measure of sentiment (Hutto and Gilbert, 2014), and the MTLD metric for lexical diversity (McCarthy and Jarvis, 2010). I also tag tweets for the presence of a mention to another user, a hashtag, an emoji, a URL or any audiovisual media content (photo, audio, or video) as these features may also influence engagement and be linked to toxicity.

Finally, to assess whether observed changes in voting intentions are specific to abnormal deviations in politicians' toxic rhetoric of if they expand to unusual communication along other dimensions, I compute continuous scores for two additional stylistic features: irony and emotional intensity. The irony score is predicted by a pre-trained deep learning model specialized in irony detection, and the emotional intensity score is constructed by taking the maximum probability associated to a set of eleven emotions as predicted by a deep learning emotion classifier (Camacho-Collados et al., 2022). While the binary LLM annotations for sarcasm/irony/humor and for emotional appeals serve as the primary measures for these constructs in the main analysis, I use these continuous scores to proxy for the intensity of these features. As shown in Figure C8, both scores are strongly and positively correlated with their respective LLM-annotated counterparts, thus lending credence to their use as reliable intensity proxies for alternative communication dimensions. Appendix C.2 provides further methodological details and presents summary statistics of all these complementary features.

To facilitate a clearer interpretation of the relationship between toxicity and online engagement, I construct an indicator variable from the continuous PAI score. This is particularly relevant for the analysis conducted in section 3 as most other stylistic features are measured as binary or categorical variables. A tweet is classified as toxic if its PAI score is higher than 0.26. This threshold is chosen such that it optimizes a commonly used classification metric, the F1-score, between the PAI toxicity indicator and human-labelled PAI toxicity on a randomly selected validation sample. While the results presented in this paper are robust to threshold choice, two remarks are in order to provide more context. First, the threshold of 0.26 lies in a similar range as the thresholds used in previous studies of the impacts of toxicity on engagement.[11] Second, in the context of our sample, where the 95th percentile of PAI toxicity is 0.24, a threshold of 0.26 effectively isolates tweets on the extreme right tail of the toxicity distribution.[12] Appendix C.3 provides further details on the data-driven procedure used to select the threshold.

To build better intuition for the toxicity metric used in the analysis, Figure A1 displays the ten tweet-level features that are most strongly correlated with the PAI toxicity indicator. Expressions of negative emotions such as disgust and anger, and the presence of strong criticism are all strong predictors of toxicity. All else equal, tweets on politically divisive topics such as gun violence, abortion or Covid masks are more toxic. These results are robust to using the continuous PAI toxicity score as the outcome variable: the same ten features emerge as the strongest predictors, and their relative ordering is largely preserved. The binary formu-

[11] Beknazar-Yuzbashev et al. (2025) uses a threshold of 0.3 arguing that it aligns with recommendations from Perspective API, namely that scores above this level indicate content that is "suspect" of being toxic. Kalra (2025) determines the optimal threshold to binarize toxicity of Tik-Tok-like posts to be 0.2.

[12] In this respect, tweets by politicians contain much milder language than the set of refugee-related tweets analyzed in Jiménez Durán et al. (2024) where the mean toxicity is 0.41 points, versus 0.05 in the sample of politicians' tweets used in this paper.

lation is preferred here for interpretability, as it provides a clearer distinction between toxic and non-toxic tweets and allows for a more direct characterization of what makes a tweet toxic. These results indicate that controlling for these features is necessary to isolate the specific contribution of toxic language, as several have been shown to independently shape user engagement (Brady et al., 2017; Rathje et al., 2021; Frimer et al., 2023) and political attitudes (Algan et al., 2025).

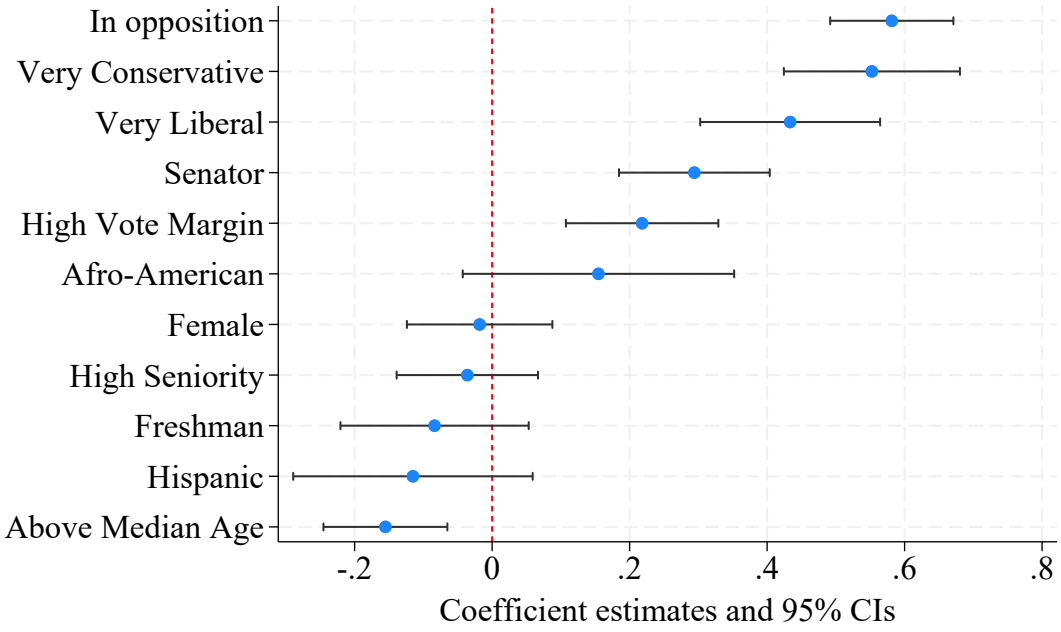
Congress members and constituency information

I combine several datasets in order to build a rich set of congress-member-by-session controls. Politicians' socio-demographics such as age, gender, race and political characteristics such as their party, seniority (number of terms served in Congress) and ideological leaning are taken from the Center for Effective Lawmaking. Politician ideology in that dataset is measured using the first dimension of the DW-Nominate score, which locates legislators on a continuous liberal–conservative scale based on their roll-call voting behavior in Congress (Poole and Rosenthal, 1985). In addition, I collect the vote margin of politicians in the previous election and a discrete indicator for how competitive a congressional race is (i.e. provided by the Cook ratings) from Ballotpedia. Descriptive statistics summarizing these characteristics as well as the characteristics related to congress members' tweet activity are presented in Table A4.

These characteristics allow to describe the variation in the supply of toxic speech across politicians. Figure 2 presents the results of a regression of a politician's standardized average toxicity in a session on their political and demographic characteristics, controlling for state and congress session fixed effects. This allows to identify which characteristics systematically predict the use of toxic rhetoric at the politician-level, and to shed insights on the determinants of toxic supply.

Several key patterns emerge. First, electoral and political features appear as the main predictors of toxicity. This suggests that the conditions of political competition (being in the party opposite to the executive, ideological positioning, strength of electoral competition in the constituency) shape the production of toxic online rhetoric and that politicians respond to such incentives. Second, within these purely political factors, being in the opposition and being positioned in the most liberal and most conservative quartiles of ideology stand out as the strongest predictors of toxicity. All else equal, opposition politicians' toxicity is 0.58 standard deviations higher than that of politicians whose party is in power. Politicians in the right-most and left-most quartiles of the ideology distribution display toxicity levels that are respectively 0.55 and 0.43 standard deviations higher than politicians in the two central quartiles. While both facts receive deeper empirical scrutiny in the following paragraphs, they suggest that: (i) politicians distill toxicity to remain visible when not in power ; and (ii)

Figure 2: Political and demographic predictors of toxicity



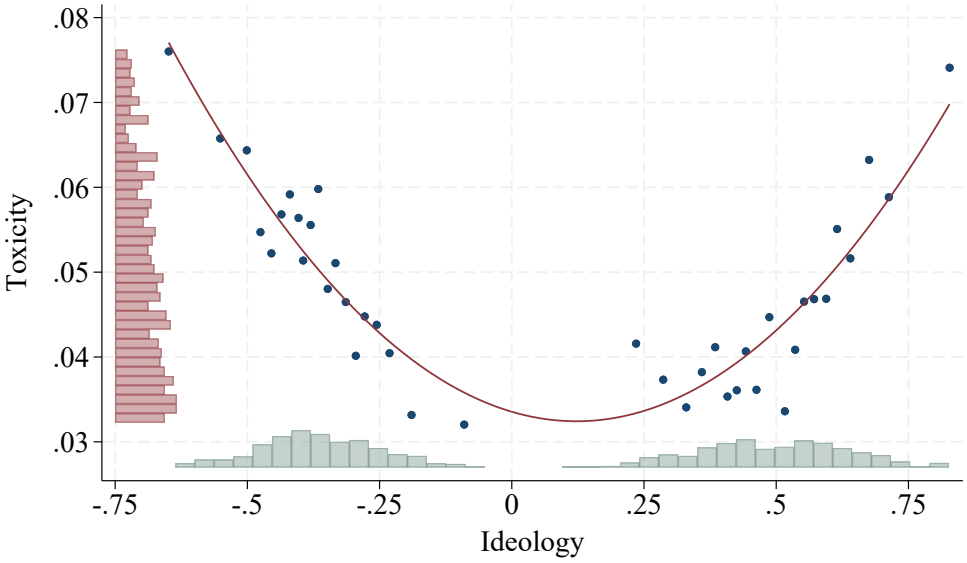
Notes: The figure plots estimates from a regression of the standardized member-session average PAI toxicity score on the full set of politician covariates. All characteristics are binary variables, save for 3-point politician ideology where the reference category groups politicians in the second and third quartiles of the DW-Nominate score, and politicians' race where the reference category is "White". The first dimension of the DW-Nominate score is a commonly-used measure of politician ideology that locates legislators on a continuous liberal–conservative scale based on their roll-call voting behavior in Congress (Poole and Rosenthal, 1985). State and congressional session fixed effects are included to absorb confounding geographic or time-related characteristics. Heteroskedastic-robust standard errors are used to compute CIs.

that toxicity is disproportionately employed by members at the ideological extremes. Third, politicians facing lower electoral competition—those elected with a high vote margin in the previous election—systematically use more toxic language. This pattern is consistent with prior evidence on the determinants of negative campaigning (Messing et al., 2017; Ballard et al., 2023; Frimer et al., 2023). Finally, politician demographics also play a role. For instance, older politicians are significantly less toxic than their younger colleagues which could be suggestive of the presence of double standards concerning the use of toxic rhetoric, or of broader "cultural" differences in approaches to political communication. While these descriptive findings could be rationalized by various models of toxic communication, they suggest that the supply of toxic speech responds to incentives tied to political competition.

Two key stylized facts about the production of toxic content merit further investigation due to their empirical salience. First, as shown in Figure 3, the use of toxic speech is non-linearly related to ideology. The U-shaped relationship indicates that toxic communication is disproportionately employed by politicians at the ideological extremes. This suggests that toxicity may serve as a strategic tool for polarization or could be used to signal extreme

stances on policy issues, highlighting the potential complementarity between politicians' ideological positioning and their communication style.

Figure 3: Relationship between congress member-session average toxicity and ideology



Notes: The figure plots the average toxicity of congress member-session units over ideology, aggregated over 40 bins. Ideology is measured using the first dimension of the DW-Nominate score of a congress member in the given congress session, computed from roll-call voting data following Poole and Rosenthal (1985). A quadratic fit of this bin scatter is plotted with a red line. Univariate histograms of congress member-session mean toxicity and ideology are also represented on the margins of the plot.

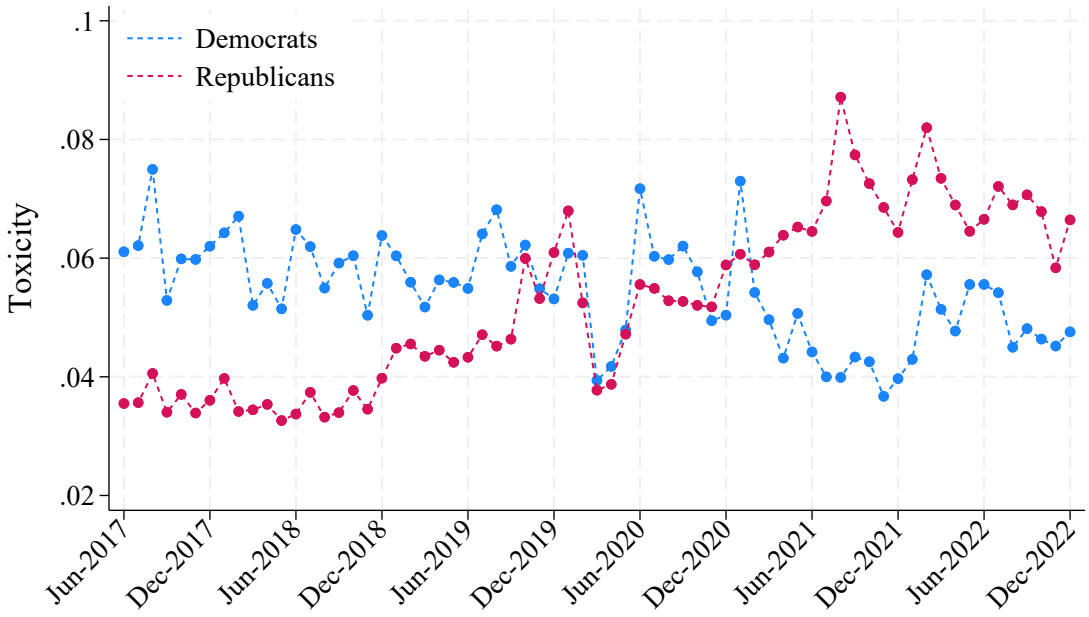
Second, the partisan gap in tweet toxicity evolves over time and tracks with changes in the political environment. Figure 4 plots the average toxicity of Democratic and Republican members over time. During the Trump presidency (2017-2020), Democratic members were, on average, more toxic than their Republican counterparts - the Covid pandemic period being a noteworthy exception with toxicity plummeting across both major parties.[13] This pattern reverses sharply following the 2020 election. Under the Biden presidency, Republican toxicity rises and consistently exceeds that of Democrats. This dynamic indicates that toxicity is a tool more often used by politicians in the opposition party.

Electoral surveys

Data on voting intentions and partisan affect come from the Nationscape electoral survey. The Nationscape survey is a large-scale electoral survey conducted on a period of 20 months from July 2019 to February 2021, with an average of 6,250 interviews carried out each week. As such, this electoral survey is one of the largest ever conducted about political attitudes at such high frequency, for such a prolonged period of time and over such a broad geographical

[13] As shown in Boxell et al. (2022), the Covid pandemic harbored a time of decrease in political polarization.

Figure 4: Evolution of tweet toxicity by party



Notes: The figure plots the average PAI toxicity score of tweets posted by Democrat congress members (in blue) and Republican congress members (in red) during each month between June 2017 and December 2022.

scope. Despite consisting of repeated cross-sections, the qualities of this dataset make it ideal to study how voter attitudes evolve around specific episodes in the electoral cycle, such as days with spikes in toxic content production by politicians.

The main variable of interest in the survey relates to respondents' voting intentions in the upcoming House elections. In addition to standard demographic variables and respondents' self reported ideology, the survey asks about their preferred candidate in the upcoming House elections. Answer choices are randomized and include "The Democratic candidate", "The Republican candidate", "Other" and "Don't Know". Using respondents' county of residence allows to build one of the main outcome measures, namely an indicator variable equal to one if the preferred candidate is the current representative, and zero otherwise.

Several restrictions are imposed on the Nationscape sample. First, respondents living in districts where the incumbent is not running for reelection are removed from the analysis on voting approval.[14] Second, the analysis is restricted to respondents who are interviewed after the congressional primaries have been held. This restriction is necessary because the identity of party nominees is usually unknown prior to primary elections.[15] This restriction

[14]This could be because the incumbent is running for another office (e.g. senator, governor or other state official), is running in another district, was deafeated in the House primary in their district, or is retiring from politics.

[15]The dates of primaries vary by state. I use those reported by the Federal Election Commission.

ensures that respondents are referring to their incumbent when answering the voting intention question. Third, respondents surveyed on or after the general election held on November 3, 2020 are removed from the sample as the voting intention question loses relevance in a post-election context. There are 98,684 respondents that satisfy these three restrictions.

Following the theoretical framework exposed in the introduction, I also proxy for a key feature that could play an important role in determining how large voting returns are, namely the degree of support towards the incumbent. Using respondents' self-reported partisan affiliation, I build a 5-point political alignment scale between a respondent and their representative. Respondents are categorized as strongly aligned (28.3% of the restricted sample), weakly aligned (24.4%), independent (12.0%), weakly opposed (17.6%) or strongly opposed (17.8%) to their representative. For instance, a respondent is classified as strongly aligned with their representative if they declare to be a strong Democrat (resp. strong Republican) and the representative is a Democrat (resp. Republican). Weakly aligned (resp. weakly opposed) include respondents who self-identify as weak partisans, or leaning towards a partisan affiliation.[16]

On average, 47.9% of respondents report they intend to vote for their House incumbent in the restricted sample. Unsurprisingly, voting intentions display a steep gradient in terms of ideological alignment with the incumbent with 95.1% of strongly aligned respondents reporting they intend to vote for the incumbent, compared to 78.1% of weakly respondents, 15.4% of independents, 7.8% of weakly opposed respondents and only 3.7% of strongly opposed respondents.

3 Effect of toxicity on online engagement

3.1 Empirical strategy

I study the relationship between politicians' tweets' toxicity and the level of engagement tweets generate from users on the platform by estimating panel regressions at the congress-member-by-week level. Specifications are of the following form:

$$TweetEngagement_{i,m,t} = \alpha + \beta Toxicity_{i,m,t} + \theta' X_{i,m,t} + \delta_{m,w(t)} + \eta_{dow(t)} + \epsilon_{i,m,t}, \quad (1)$$

where $TweetEngagement_{i,m,t}$ is the engagement received by tweet i (e.g.IHS-transformed number of likes/ retweets/ replies/ quote tweets) posted by congress member m at time t ; $Toxicity_{i,m,t}$ is an indicator variable equal to one if the tweet's Perspective API (PAI) score is above a data-driven optimal threshold (0.26), and zero otherwise. The vector $X_{i,m,t}$ contains a rich set of tweet-level controls designed to isolate the effect of toxicity from confounding

[16]The main results are robust to defining partisan affiliation differently.

linguistic features. This set includes controls such as a tweet's topic, its length in words, and indicators for the presence of media, URLs, hashtags, mentions, or emojis, as well as for whether the tweet is a quote tweet and whether it is posted by a politician's campaign account. Crucially, it also includes the detailed set of features extracted via the LLM annotation pipeline: an indicator for the presence of *criticism or disagreement*, and for whether such criticism is strongly expressed, the *target* of the criticism, an indicator for the use of *sarcasm, irony or humor*, the tweet's primary *rhetorical appeal* (cognitive or emotive), and the specific *emotion* conveyed. To account for the tone and linguistic complexity of the message, I also control for sentiment and lexical diversity using categorical variables indicating whether a tweet falls into the bottom quartile, interquartile range, or top quartile of these two features' distributions.

The model is estimated with a demanding set of fixed effects to account for unobserved heterogeneity. $\delta_{m,w(t)}$ are congress-member-by-week fixed effects which absorb all time-invariant characteristics of politicians as well as any unobserved shocks or trends specific to that politician occurring at a weekly-level. $\eta_{dow(t)}$ are day-of-the-week fixed effects to control for systematic daily patterns in user engagement. Finally, $\epsilon_{i,m,t}$ is the error term. Standard errors are clustered at the congress member level to account for potential correlation in the residuals for tweets from the same politician.

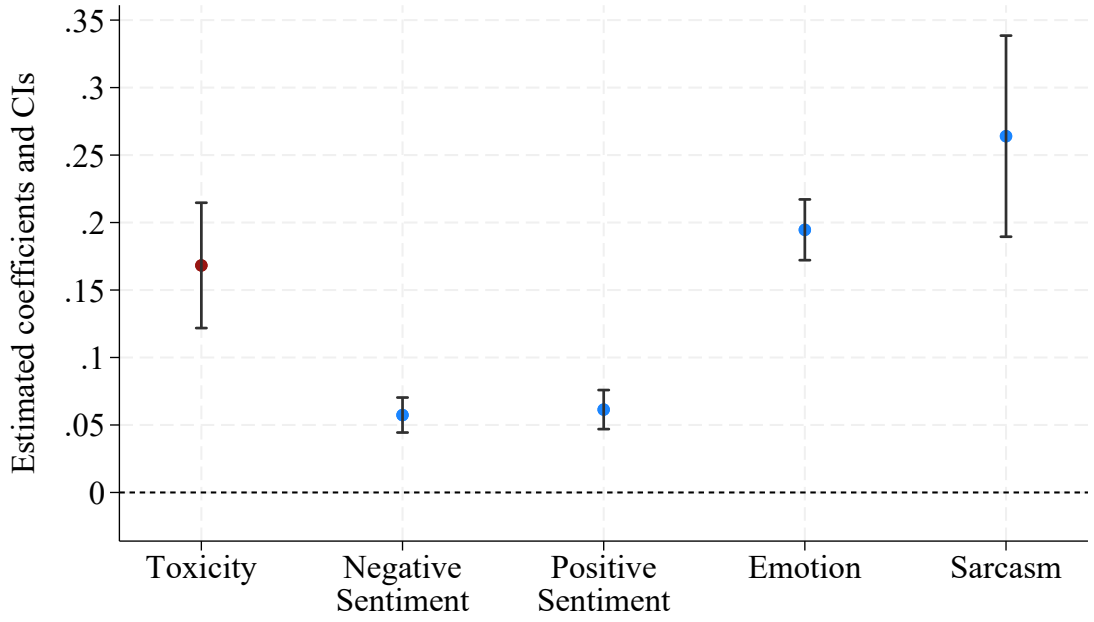
As such, the coefficient of interest β captures the residual correlation between tweet toxicity and tweet engagement once politician-specific time-varying factors and tweet-level characteristics are accounted for. In other words, I exploit variation in the toxicity of tweets posted by a given congress-member during a given week to estimate the association between engagement and toxicity, while controlling for a rich set of linguistic features such as the presence of criticism/disagreement, emotional communication or negative sentiment.

While this empirical strategy does not rely on purely random variation in tweet toxicity, the inclusion of tight congress-member-by-week fixed effects and detailed tweet-level controls allows to rule out a host of alternative explanations. First, they absorb any unobserved differences between politicians that might be simultaneously correlated with their engagement and their propensity to use toxic language (e.g., more popular politicians might also be more toxic). Second, the fixed effects account for any unobserved, politician-specific weekly events that may generate a temporary engagement boost while also inducing more toxic communication, such as a politician being featured in the national news, having a bill debated, or facing a local crisis. The coefficient β is therefore identified from the variation in a politician's communication style within a given week, netting out these potentially confounding between- and within-politician factors.

3.2 Main results

Figure 5 presents the main results, plotting the estimated coefficients from Equation 1. Several findings emerge. First, toxic communication produced by politicians is an important driver of online engagement. Toxic tweets are associated with a 16.8% increase in the number of likes relatively to non-toxic tweets, once conditioning on demanding fixed effects and other tweet-level drivers of engagement. This represents a sizable and economically significant increase given that the average tweet receives approximately 750 likes. This finding is consistent with recent evidence documenting that toxic content increases users' engagement and time spent on social media.[17]

Figure 5: Online engagement returns to politicians' tweet toxicity



Notes: Point estimates and 95% CIs associated to tweet toxicity and tweet textual features are plotted. Toxic tweets receive 16.8% more likes than similar non-toxic tweets written by the same politician in the same week. Estimates are obtained from a regression of the IHS-transformed number of likes on several textual features conditional on member-by-week and day-of-the-week fixed effects. The association between the number of likes and textual features is jointly estimated. Standard errors are clustered at the politician-level.

[17] Using a field experiment to hide toxic content from users' feed, Beknazar-Yuzbashev et al. (2025) find that suppressing toxicity reduces daily post clicks by around 13% on Twitter. While their experiment finds a null effect on aggregate user reactions (likes and retweets), my finding of a strong positive effect is not contradictory and can be attributed to two key distinctions. First, the two studies examine different content types: my analysis focuses on communication from high-profile political actors that may induce different expressive or identity signaling motives in users' reactions, whereas their study examines general user-generated content. Second, and more fundamentally, the studies address different economic questions. The hiding intervention on users' feeds estimates the average demand-side effect of removing all toxic content on users' social media activity. In contrast, my analysis focuses on the supply-side incentive for a producer to choose a toxic communication strategy.

Second, while toxicity provides a substantial engagement boost – about three times larger than that of expressing negative or positive sentiment – other dimensions of political communication such as the use of emotional appeals or the expression of sarcasm, irony or humor strongly drives user engagement. This is consistent with findings from previous studies which show that emotional language by politicians generates strong online engagement returns (Crockett, 2017; Brady et al., 2019). With respect to these studies, this paper's findings shed light on an under-explored yet economically relevant aspect of political communication, namely toxic style, and isolates it from related concepts that have received greater empirical scrutiny.

I provide additional evidence to gauge how important toxicity is as an online communication device for politicians by running a similar specification on the subsample of tweets that express strong criticism. This allows for a direct apples-to-apples comparison between toxic tweets and non-toxic tweets that share the same underlying critical intent, thereby cleanly isolating the marginal return to incivility. The results are presented in Figure 6. While the coefficient on toxicity remains relatively similar (0.207), the relative importance of other stylistic features, such as emotional appeals or the use of sarcasm, diminishes considerably. When expressing criticism or disagreement, the use of sarcasm/irony or emotions is not near as successful a strategy to generate engagement compared to criticizing or disagreeing in a disrespectful and uncivil way. This provides evidence that in the realm of political confrontation, the choice to employ a toxic communication style offers a very potent engagement advantage.

In light of the functioning of Twitter's recommendation algorithm that pushes popular content to users' feed, there is reason to believe that this induces toxic tweets to have a much larger reach than "ordinary" tweets, exposing a larger audience to toxicity.[18] This also highlights the important social rewards that toxic content generates on social media platforms and that may arguably distort congress members' tweeting behavior towards posting toxic content. As congress members learn that offensive tweeting reaps more online engagement, they may be lured into supplying violent online communication more often.[19]

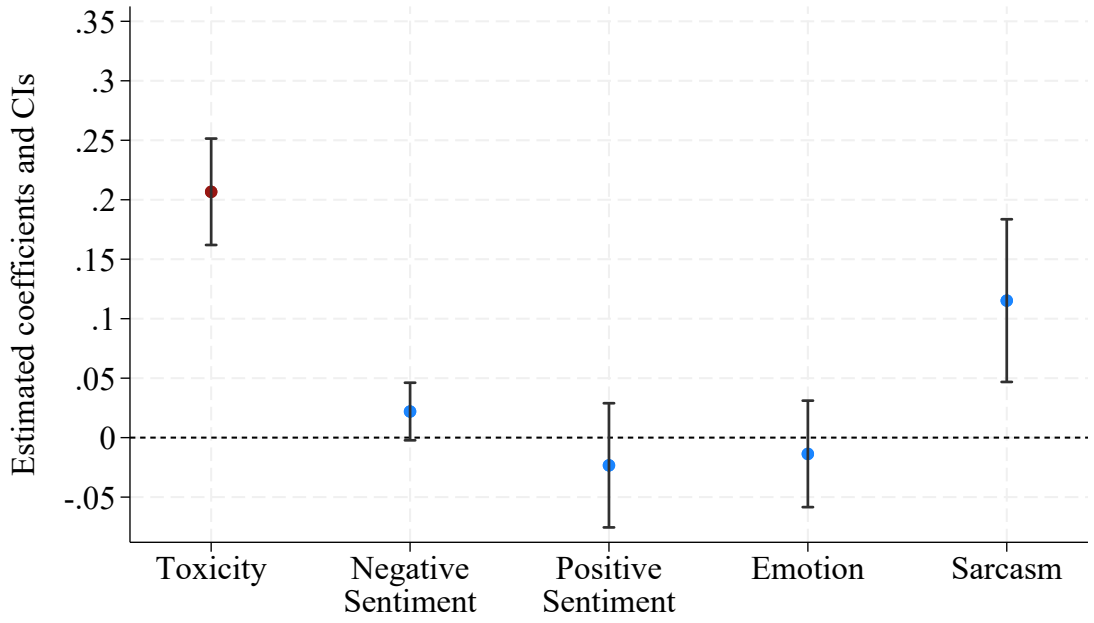
3.3 Robustness

Several robustness checks are carried out to consolidate the validity of this main finding. First, I show that the positive association between online engagement and toxicity holds across alternative definitions of toxicity. As displayed in Figure B1, the coefficient remains positive and statistically significant when using the custom LLM-based definition of toxicity

[18] As coded in Twitter's publicly released recommendation algorithm, liked content receives a large "boost" (i.e. 30 times) when the algorithm ranks content to display on users' feed.

[19] This mechanism is similar in spirit to the one exposed in Schöll et al. (2023) who study gendered differences in the social benefits to tweeting about gender issues among Spanish politicians.

Figure 6: Online engagement returns to politicians' tweet toxicity - Sample of strongly critical tweets



Notes: Point estimates and 95% CIs associated to tweet toxicity and various tweet textual features are plotted. Estimates are obtained from a regression of the IHS-transformed number of likes on several textual features conditional on member-by-week and day-of-the-week fixed effects. Only tweets expressing strong criticism or disagreement are included in this regression. The association between the number of likes and textual features is jointly estimated. The dependent variable is the inverse hyperbolic sine of the number of likes. Standard errors are clustered at the politician-level.

as the main regressor (point estimate = 0.054, p-value = 0.02) or when using the standardized measure of the continuous PAI toxicity score (point estimate = 0.074, p-value < 0.01). In addition, Figure B2 shows that the positive association between tweet toxicity and engagement is robust to using different thresholds to define the binary PAI toxicity variable. The point estimate gradually increases with the threshold, consistent with the idea that more extreme instances of toxicity generate stronger engagement returns.

Second, I test whether the finding is specific to proxying for online engagement through the number of likes by re-estimating the main specification using other measures of user engagement as the dependent variable. Table B1 shows that the positive association holds whether engagement is measured by the number of retweets, replies, quote tweets, or by the total volume of engagement generated by a tweet (measured as the sum of likes, retweets, replies and quote tweets). The consistency of the result across different forms of engagement confirms that toxicity generates a broad-based attention response. Due to the high correlation between user engagement metrics (Goda et al., 2020), the remainder of this paper thus focuses on the number of likes as the primary proxy for user engagement.

Third, I provide evidence that the result is not an artifact of my estimation framework by varying the functional forms and fixed effects used to estimate equation 1. Table B2 shows that the positive association still holds when using a Poisson regression to model the relationship between the raw count of likes and tweet toxicity. The coefficient on toxicity remains positive and highly significant (point estimate = 0.245, p-value < 0.01). Finally, Table B3 displays the stability of the baseline OLS coefficient across progressively more demanding specifications. The estimate remains remarkably stable when moving from a model where politician fixed effects and week fixed effects enter additively (point estimate = 0.168, p-value < 0.01) to the preferred specification with politician-by-week fixed effects (point estimate = 0.168, p-value < 0.01). There is barely any movement in the coefficient when additionally controlling for politician-by-tweet-topic fixed effects (point estimate = 0.169, p-value < 0.01). Combining the stability of the main coefficient across specifications to the increase in the R-squared (from 0.59 to 0.65 to 0.67) suggests that unobserved within-politician factors are unlikely to be driving the association between tweet toxicity and online engagement.

Taken together, these results confirm that the positive association between online engagement and tweet toxicity is not sensitive to alternative definitions of toxicity and engagement, nor to the choice of statistical model, thus strengthening the credibility of my main finding.

3.4 Heterogeneity

Having established a clear connection between tweet toxicity and online engagement, I next investigate the conditions that influence the strength of the returns to tweet toxicity. Understanding such patterns is crucial for explaining the variation in the supply of toxic speech across politicians. To do so, I rerun a variant of equation 1 that interacts the toxicity indicator with a comprehensive set of politician, constituency, and time-related characteristics. The specification is as follows:

$$\begin{aligned} IHS(Likes)_{i,m,t} = \alpha &+ \beta Toxicity_{i,m,t} + \lambda M_{m,t} \\ &+ \gamma(Toxicity_{i,m,t} \times M_{m,t}) \\ &+ \theta' X_{i,m,t} + \delta_{m,w(t)} + \eta_{dow(t)} + \epsilon_{i,m,t}, \end{aligned} \tag{2}$$

where $M_{m,t}$ is a vector of characteristics describing the political environment a tweet is produced in. These characteristics are grouped into two broad categories: (i) politicians' demographic and political characteristics (e.g., race, gender, age by quartiles, party, ideology by quartiles, vote margin in the previous election by quartiles, and whether it's the politician's first term in Congress) ; (ii) politicians' baseline tweeting behavior (e.g., toxicity, popularity

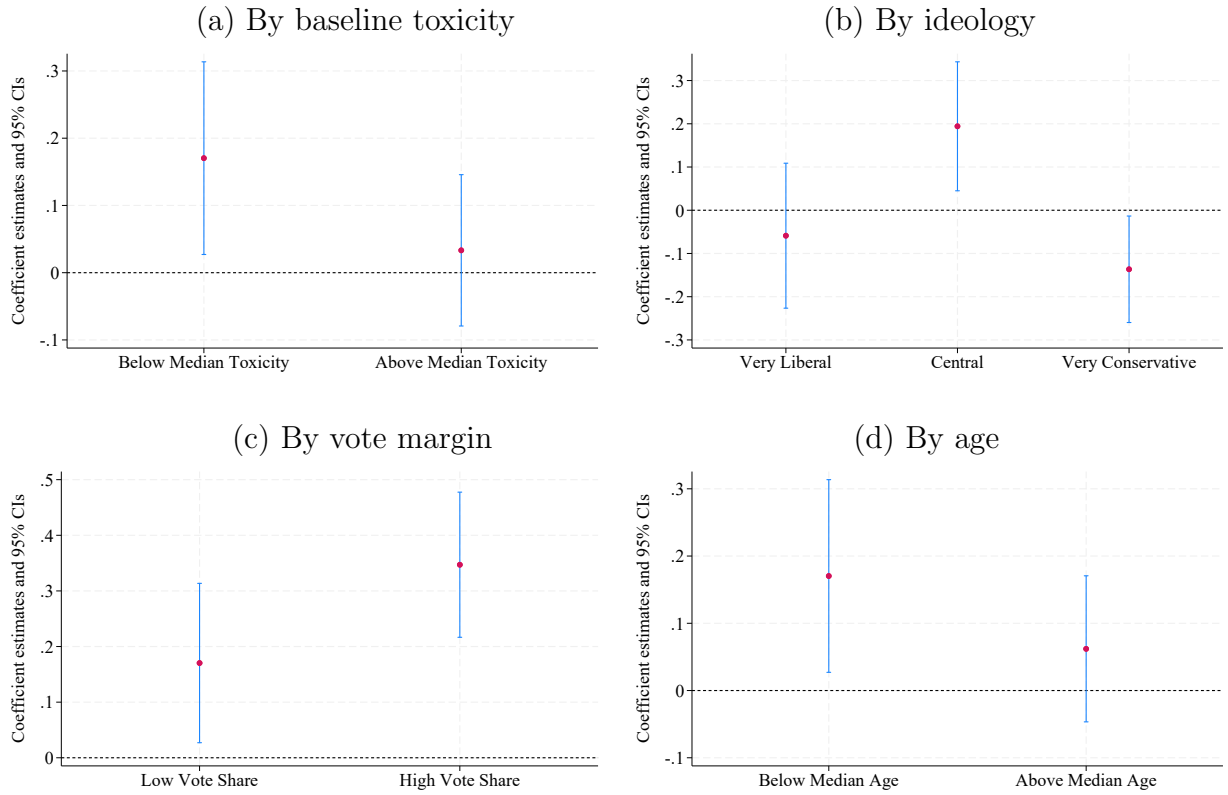
as proxied by the average number of likes, and productivity as measured by the total number of tweets). Characteristics linked to politicians' baseline tweeting behavior toxicity are computed as the average of that characteristic in the first six months of the congressional session (baseline period). Politicians are classified into whether they fall below or above the median of each baseline measure. To avoid data leakage, the specification is estimated on the sample of tweets posted after the first six months of the congressional session. The vector also includes an indicator for whether the tweet is produced by a politician's campaign account and whether it was posted in the three months leading to a congressional election. As before, this specification includes politician-by-week fixed effects $\delta_{m,w(t)}$, day-of-the-week fixed effects $\eta_{dow(t)}$ and the full set of tweet-level controls. Standard errors are clustered at the congress member level.

Figure 7 plots the results. The most striking pattern that emerges is the presence of *decreasing returns* to a politician's toxic content production, displayed in panel (a). It plots the marginal engagement return of a toxic tweet depending on whether a politician's baseline toxicity is below or above the median of the cross-politician baseline toxicity distribution. The engagement premium is largest for politicians who are not very toxic in the baseline period and diminishes for politicians whose baseline toxicity is above the median. The difference between the two coefficients is statistically significant (p-value = 0.016). This finding suggests that breaking norms of civil political debate generates greater attention gains when done parsimoniously.

Beyond politicians' baseline toxicity, other factors related to the political environment or to politicians' demographics also influence the size of attention returns. Panel (c) shows that the engagement premium is significantly larger for politicians with a high vote share in their previous election (p-value for the difference in coefficients = 0.008). This suggests that electorally-secure incumbents may afford offensive slur for visibility motives. In addition, the engagement premium is larger for younger politicians, i.e. those below the median age, which suggests the presence of double standards in toxicity production – users perceiving toxicity differently when produced by older politicians – or greater social media communication skills of younger politicians when distilling toxicity in their online rhetoric. Other demographic characteristics like gender and race, or political factors like seniority, do not appear to significantly affect the size of engagement returns. Taken together, these patterns align closely with the cross-sectional variation in toxicity between politicians presented earlier. The fact that electorally safer and younger politicians—the two groups who receive the highest online returns—are more toxic than their more electorally-contested and older counterparts is consistent with politicians responding to these engagement incentives.

However, the heterogeneity by ideology, shown in panel (b), cannot be explained by engagement incentives alone. While the descriptive data clearly shows that ideologically ex-

Figure 7: Heterogeneous engagement returns to politicians' tweet toxicity



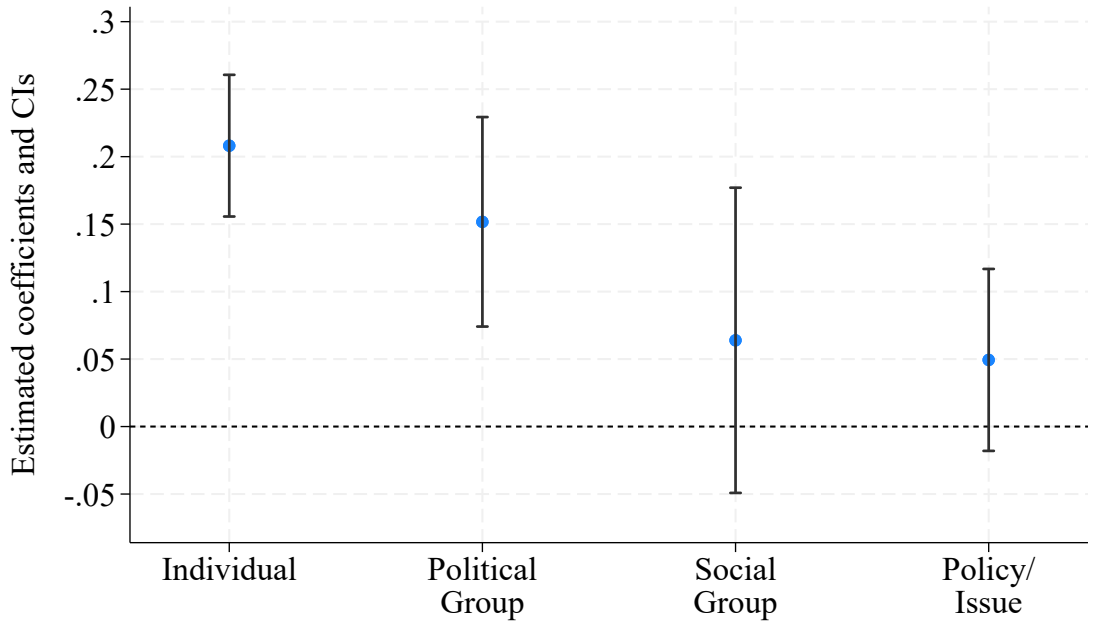
Notes: All four panels display point estimates and 95% CIs associated to the interaction between tweet toxicity and a given characteristic. Estimates are derived from the fully interacted regression specified in Equation 2. Panel (a) displays differential engagement returns by politicians' baseline toxicity; panel (b) by politician ideology, with central politicians corresponding to politicians in the interquartile range of the DW-NOMINATE score while very liberal politicians (resp. very conservative) correspond to politicians in the first (resp. fourth) quartile; panel (c) by vote margin of the incumbent in the previous election; and panel (d) by politicians' age. The regression is estimated on all tweets posted after the first six months of the congressional session and includes the full set of tweet-level controls and politician-by-week fixed effects. Standard errors are clustered at the politician-level.

treme politicians are the most toxic, engagement returns are significantly lower than those for centrists. This misalignment suggests that for politicians at the ideological extremes, the decision to use toxic language is likely driven by objectives other than maximizing broad online engagement. As the subsequent analysis on voting intentions will suggest, these politicians may be prioritizing the mobilization of their ideological base and the signaling of partisan purity, goals for which toxicity can be an effective tool even if it fails to generate widespread attention online. This underscores the need to consider both online and offline incentives to fully understand the strategic calculus behind political incivility.

Finally, I investigate whether engagement returns differ depending on the target of toxic tweets. Using the target categories annotated by the LLM pipeline, I estimate the marginal engagement return of a toxic tweet for each of the four categories: (i) individuals; (ii) collec-

tive political entities (e.g., a party, a political group, a media organization); (iii) non-political entities or minorities (e.g., social, racial, religious minorities); (iv) issues, policies, ideologies (e.g., a societal problem, a piece of legislation)). The results, shown in Figure 8, reveal a clear hierarchy of engagement rewards. Toxic tweets targeting individuals generate the largest engagement boost (point estimate = 0.208, p-value < 0.01), while toxic statements aimed at social groups/minorities (point estimate = 0.064, p-value = 0.27) or abstract policies and issues (point estimate = 0.049, p-value = 0.15) are associated with much smaller increases.

Figure 8: Engagement returns to toxicity by target



Notes: Point estimates and 95% CIs associated to the target of toxic tweets are plotted. Estimates are obtained from a regression of the IHS-transformed number of likes on a categorical variable for tweets' toxicity target. Targets are identified via LLM annotation for toxic tweets that express criticism. The definition of each category is provided in Appendix C.2. The reference category is a non-toxic tweet. The regression includes the full set of tweet-level controls as detailed in Section 3 and politician-by-week fixed effects. Standard errors are clustered at the politician-level.

These findings help to rationalize some of the variation observed in the supply of toxic speech between politicians. Consistent with a simple model of political communication, politicians direct their toxic rhetoric towards the targets that yield the highest attention rewards. Indeed, 63% of targeted toxic tweets are aimed at individuals, the most rewarding target category, whereas only 5.7% are directed towards social groups. Taken together, the alignment between the magnitude of online engagement returns and politicians' observed use of toxic communication suggests that politicians adjust their communication strategies to online incentives embedded in the social media platform.

4 Effect of toxicity on voting intentions

4.1 Empirical strategy

Identifying the causal effect of toxic political communication on voting intentions – let alone actual electoral behavior – is empirically challenging. A naive correlation between a politician's toxicity and their electoral support is likely to be confounded by omitted variables such as candidate quality or ideology. Reverse causality may also prevent interpreting this correlation causally as toxic speech may be used out of desperation by politicians trailing in the polls.

To overcome these challenges, I employ an event study design that investigates how voters' attitudes evolve around salient shocks in their House representative's communication on Twitter/X. The underlying idea is that online rhetoric must stand out relatively to the online flow of information voters are exposed to in order to influence their offline political attitudes. I therefore focus on comparing voters interviewed before and after days on which their representative's online communication is unusually toxic – referred to as "toxic spikes" henceforth. The continuous nature of the Perspective API score is particularly well-suited for this task, as it allows for the identification of these high-intensity events. Specifically, a toxic spike is defined as a day where the representative's average PAI score exceeds the 99th percentile of their own historical daily toxicity distribution.[20] Figure B3 provides visual examples of this spike identification for representatives of both parties and who vary in terms of average toxicity.

This empirical strategy compares the voting intentions of individuals living in the same district who are interviewed in the days immediately before and immediately after their representative's toxic spike. Formally, I estimate the following specification:

$$\begin{aligned}\text{Voting Intention}_{i,d,t} = \alpha + \beta \text{Post Toxic Spike}_{d,t} \\ + \theta' X_i + \psi_d + \phi_t + \eta_{dow(t)} + \epsilon_{i,d,t},\end{aligned} \tag{3}$$

where $\text{Voting Intention}_{i,d,t}$ is an indicator equal to one if respondent i living in district d and interviewed on day t intends to vote for the incumbent serving their district in the upcoming House elections, and zero otherwise (e.g., voting for another candidate or not knowing who to vote for). $\text{Post Toxic Spike}_{d,t}$ is an indicator equal to one if the interview occurs in the week following a toxic spike, and zero if it occurs in the week before. X_i is a vector of respondent-level controls (ideological alignment with the incumbent, age, gender,

[20] This distribution is calculated over the pre-2020 election period (July 2019-November 2020), omitting the first six months of the 116th congressional session to allow for the construction of baseline tweeting behavior variables.

race, ethnicity, and education). ψ_d are district fixed effects; ϕ_t are day fixed-effects; $\eta_{dow(t)}$ are day-of-the-week fixed effects; and $\epsilon_{i,d,t}$ is the error term. Standard errors are clustered at the electoral district level. Specifications are estimated on the sample of respondents surveyed after the congressional primary has occurred in their state, before the general election is held, and in districts where the incumbent is on the House ballot.[21] It only includes respondents interviewed within a one week window of their incumbent's toxic spike in order for the control group (respondents interviewed before) to be as comparable as possible to the treated group (respondents interviewed after).

The core identifying assumption is that the timing of a respondent's interview within a district is as-good-as-random with respect to the exact day of the representative's toxic spike, conditional on time fixed effects and respondent sociodemographics. I argue that this is the case as survey roll-out and completion rates are unlikely to be affected by representatives' toxic tweeting behavior. Empirical support for this argument is provided in Table B4. This table shows that the observable demographic and political characteristics of respondents interviewed in the week before a spike are statistically indistinguishable from those interviewed in the week after.

4.2 Main Results

Results for the effect of incumbents' toxic spikes on voting intentions are presented in Table 1. Respondents display a 2.4 percentage point increase in their voting intentions for the incumbent in the week following a toxic spike by the incumbent, as shown in column (1). This represents a 4.6% increase relative to the pre-spike mean level of support for the incumbent.

To further understand the source of this shift, I decompose the outcome into finer margins: intending to vote for the incumbent versus being undecided, and intending to vote for the incumbent versus the opponent. The results, respectively shown in Columns (2) and (4), indicate that the effect is driven by the persuasion of undecided voters, not of opposing partisans. Support for the incumbent increases by 3.6 percentage points relative to being undecided, with no evidence of voters switching their support from the opponent. Interestingly, column (4) shows that support for the opponent increases by 3.6 percentage points relative to being undecided. Taken together, these results suggest that toxic spikes do not persuade opponents but instead act as polarizing events, pushing previously undecided voters towards their respective partisan camps with a small net benefit accruing to the incumbent.[22]

[21] Restricting to respondents interviewed after the primary election ensures that the identity of the party nominees are known and that respondents have appropriate information to correctly answer the voting intention question.

[22] The reason why the estimates in columns (2) and (4) do not cancel each other out in the main specification is because the dependent variable in Column (1) represents a three-way choice, i.e. (i) preferring the incumbent, (ii) the opponent, or (iii) being undecided. The 2.4 percentage point net increase in support for the incumbent reflects the aggregate shift among all respondents, whereas columns (2) to (4) analyze shifts

Table 1: Changes in voting intentions following incumbents' toxic spike

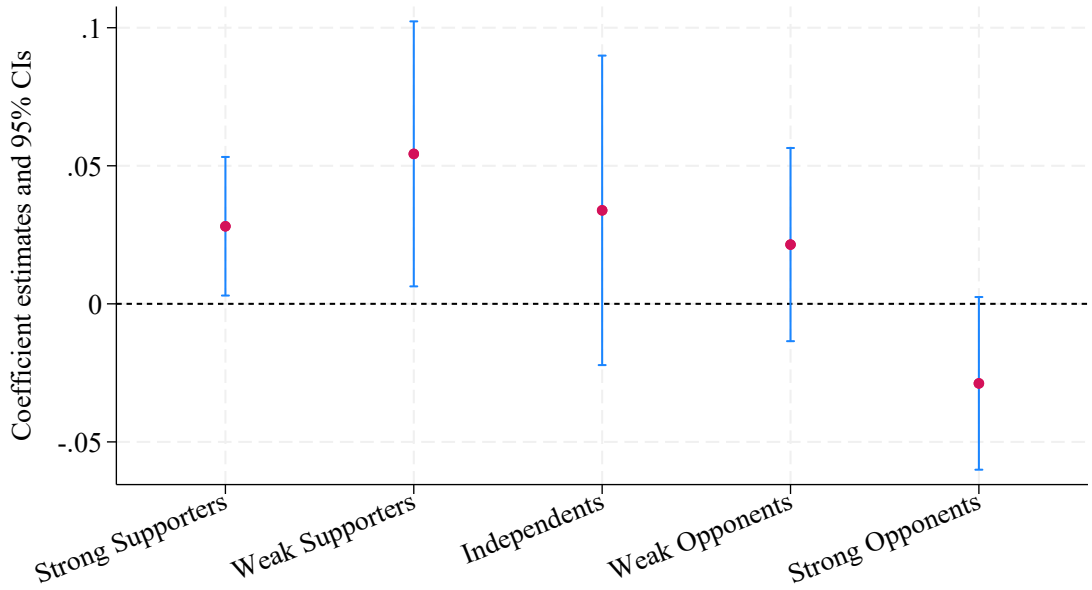
	(1) Vote Incumbent vs. Rest	(2) Vote Incumbent vs. Undecided	(3) Vote Incumbent vs. Opponent	(4) Vote Opponent vs. Undecided
Post Toxic Spike=1	0.024***	0.036***	0.007	0.036**
	(0.009)	(0.011)	(0.009)	(0.016)
Respondent controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
Day FEs	Yes	Yes	Yes	Yes
Day of week FEs	Yes	Yes	Yes	Yes
Observations	6,117	4,216	5,075	2,921
Number of districts	197	197	197	194
R squared	0.65	0.51	0.77	0.57
Mean dep. var.	0.52	0.75	0.63	0.65

Notes: The table presents results from regressions of respondents' voting intentions on an indicator for whether the respondent is surveyed in the week following their House representative's toxic spike. In column (1), the dependent variable is an indicator variable equal to 1 if the respondent intends to vote for their House incumbent, and equal to 0 otherwise (i.e. if the respondent intends to vote for another candidate or they are undecided). The dependent variable in columns (2) and (3) are similarly coded but are equal to 0 if the respondent is undecided, resp. intends to vote for the incumbents' opponent. In column (4), the dependent variable is coded as 1 if the respondent intends to vote for the opponent, and equal to 0 if they are undecided. All regressions include respondents' 5-point ideological alignment with the incumbent, socio-demographic controls, district, day and day of the week fixed effects. Standard errors are clustered at the electoral district level. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure 9 presents additional supporting evidence for the divergence in voting intentions following unusually toxic communication by politicians. The figure displays heterogeneous responses of voters to toxic spikes depending on their ideological alignment with the incumbent. It shows that the increase in support for the incumbent is driven by their own partisans. The effect is largest among weakly aligned voters, who become 5.4 percentage points more likely to support the incumbent (p-value = 0.027), a substantial effect given their 76.3% pre-spike baseline support. In contrast, the spike appears to backfire among the opposition. Strongly opposed voters are 2.9 percentage points less likely to support the incumbent post-toxic-spike, effectively hardening their opposition (p-value = 0.071). There is no discernible shift in voting intentions following a toxic spike for respondents who self-identify as independents or weak opponents. Taken together, these results suggest that toxic spikes serve as a tool to mobilize a politician's electoral base while alienating the opposition, thereby contributing to intensifying polarization.

within specific two-way comparisons.

Figure 9: Changes in voting intentions following incumbents' toxic spike by respondent ideological alignment



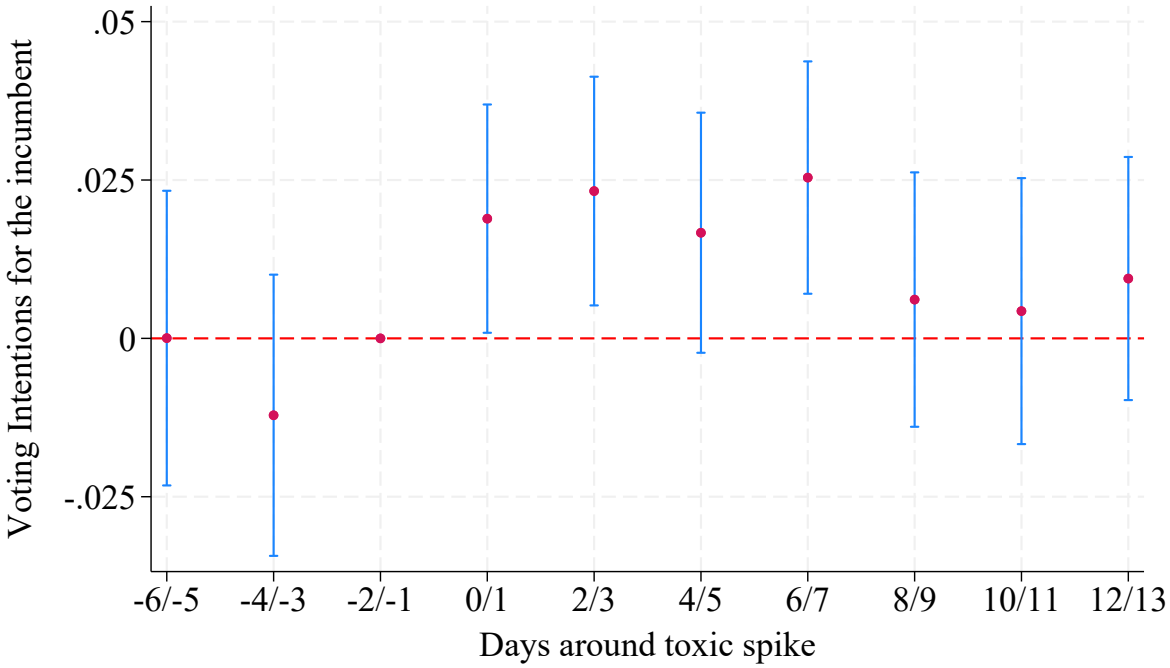
Notes: The figure displays coefficients and 95% confidence intervals of a regression of respondents' voting intentions on an indicator for whether the respondent is surveyed in the week following their House representative's toxic spike interacted with respondents' 5-point ideological alignment with the incumbent. Regressions include respondents' baseline ideological alignment with the incumbent, socio-demographic controls, district, day and day of the week fixed effects. Standard errors are clustered at the electoral district level.

Effect Dynamics How long does the increase in voting intentions following a politician's toxic spike on Twitter last? To answer this question, I estimate the event-study version of equation 3 where indicator variables for day groups are included instead of the Post Toxic Spike indicator variable. Days are pooled in groups of two to ensure coefficients are estimated with sufficient precision. I also extend the sample to include respondents surveyed up to two weeks following a toxic spike in order to better measure when the effects fade away.

Figure 10 displays the associated event study graph corresponding to this specification. It allows to analyze the dynamics of toxicity's incidence on voting intentions over time. The increase in voting intentions is manifest starting on the day of the toxic spike and the following, and lasts a total of 7-8 days. Voting intentions then gradually revert to their pre-toxic spike levels. As such, the effect of extreme toxicity on voting intentions seems fairly short-lived and in accordance with temporary news shocks. It is also reassuring to observe that voting intentions are remarkably stable in the days leading up to politicians' toxic spikes.

This supports the view that the empirical design is picking up relevant variation in politicians' aggressive communication strategy. Specifically, it reduces concerns that politicians are using toxicity to react to changes in voting intentions in their constituency. Had this been the case, one would have expected clear monotonic patterns in the days preceding a toxic spike. It also rules out the presence of anticipation effects by voters to incumbents' use of toxic language. This helps to attenuate concerns about the presence of unobservable factors in the interaction between politicians and voters that could bias the main estimates.

Figure 10: Event study of voting intentions around toxic spikes by House representatives



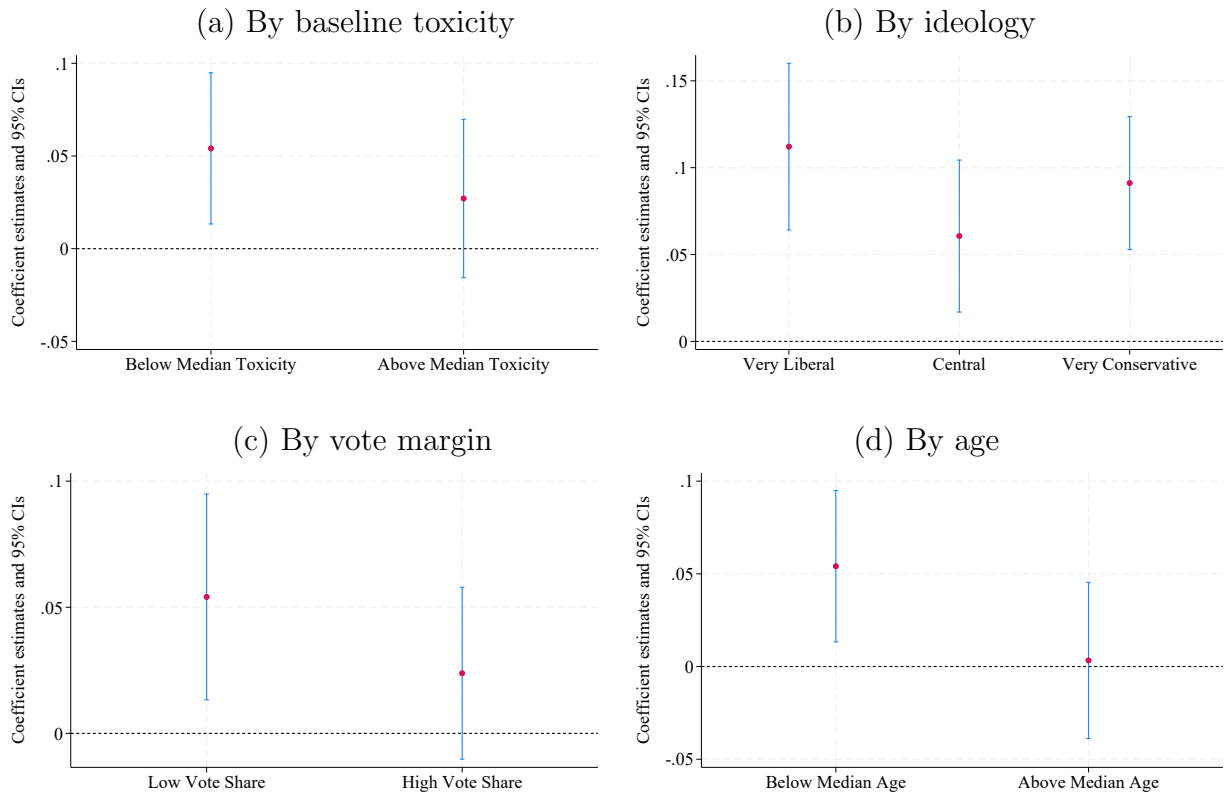
Notes: The figure displays coefficients and 95% confidence intervals (CIs) of a regression of respondents intent to vote for their House representative in the 2020 elections on leads and lags around spikes in the toxic content produced by their representative. Each lead and lag pools together two days to ensure precise estimation. The two days prior to the toxic spike are used as the reference time unit. Regressions include respondents' 5-point ideological alignment with the incumbent, socio-demographic controls, district, day and day of the week fixed effects. Standard errors are clustered at the electoral district level.

Decreasing and heterogeneous returns Mirroring the pattern of decreasing returns found in the analysis of online engagement, I find evidence that gains in voting intentions after toxic spikes depend on a politician's prior communication style. To do so, I interact the Post Toxic Spike indicator with a dummy for whether the incumbent's baseline toxicity is above the sample median, jointly with interactions of the indicator with a full set of politician

and constituency characteristics.[23]

Figure 11 presents the results. Panel (a) indicates that the increase in voting intentions is concentrated among politicians with low baseline toxicity. Conversely, voting intentions remain unaffected following toxic spikes by politicians whose baseline toxicity is high. While the difference between these two point estimates is at the not statistically significant (p-value=0.191), the pattern is clear and directionally consistent with the online engagement findings.

Figure 11: Heterogeneous political returns to toxicity



Notes: All four panels display point estimates and 95% CIs associated to the interaction between being interviewed in the week following a toxic spike and a given characteristic. Coefficients for the interaction with the full set of politician and constituency characteristics are jointly estimated. Panel (a) displays differential political returns by politicians' baseline toxicity; panel (b) by politician ideology, with central politicians corresponding to politicians in the interquartile range of the DW-NOMINATE score while very liberal politicians (resp. very conservative) correspond to politicians in the first (resp. fourth) quartile; panel (c) by vote margin of the incumbent in the previous election; and panel (d) by politicians' age. Regressions include respondents' 5-point ideological alignment with the incumbent, socio-demographic controls, district, day and day of the week fixed effects. Standard errors are clustered at the electoral district level.

This result is consistent with a "surprise" effect. Breaking norms of civic political debate is most effective at mobilizing voters when it is most salient, i.e. when used by politicians

[23]Baseline toxicity is computed as the average PAI score of a politician's tweets in the first six months of the 116th congressional session (January-June 2019).

who usually adhere to those norms. For representatives who frequently employ aggressive language, toxic spikes may simply be normalized by voters and fail to register as a significant event. Combined with the presence of decreasing online engagement returns, this helps to rationalize why the supply of offensive speech is still used rather sparingly by politicians.

Beyond a politician's own communication history, the electoral returns to toxicity also vary systematically with their ideological positioning and electoral conditions. Panel (b) reveals a U-shaped relationship between ideology and electoral returns. The effect of a toxic spike on voting intentions is large for both very liberal and very conservative politicians, while it is smaller for centrists. This finding helps resolve the puzzle identified in the online engagement analysis. While extremists receive lower engagement returns from toxicity, they receive the highest electoral returns. This suggests that their supply of toxicity responds more to electoral incentives, e.g. mobilizing their ideological base, and that these outweigh the goal of maximizing online engagement.

In contrast, the returns vary significantly with electoral competition, as shown in Panel (c). The positive effect on voting intentions is concentrated among politicians in districts where competition is high, i.e. those who won by a tight margin in the previous election while returns are smaller for more secure incumbents (p-value for the difference in coefficients = 0.077). This presents another clear trade-off. Electorally safe politicians receive higher online engagement gains from toxicity but smaller electoral returns. In the cross-sectional comparison to politicians serving more contested constituencies, our earlier findings showed they displayed stronger average levels of toxic communication. This suggests that, free from immediate electoral pressure, electorally safe politicians are more responsive to online incentives and may use toxicity as a tool to build visibility and attention. Conversely, politicians in more competitive constituencies, while gaining less online traction, may choose to distill toxicity for offline mobilization.

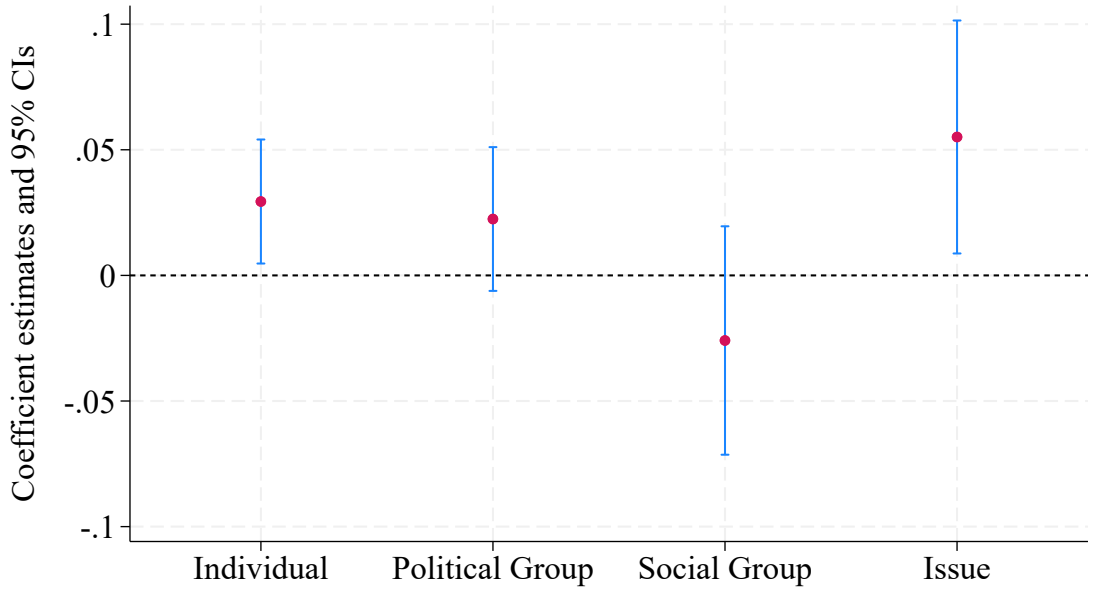
Finally, Panel (d) shows a steep age gradient in electoral returns that mirrors the online engagement pattern. The positive effect of a toxic spike is driven entirely by politicians below the median age. This alignment of both online and offline incentives rationalizes the descriptive fact that younger politicians are more toxic than their older colleagues. For this group, toxic communication appears to be a winning strategy on both fronts, providing a strong incentive for its use.

Heterogeneous returns by target of toxic spikes Just as changes in political views following toxic spikes depend on its *producer*, they also depend on the *target* of such spikes. To investigate how political returns differ along this content-related dimension, I rerun the main specification including an interaction between the Post Toxic Spike indicator and the

target category of the most toxic tweet on the day of a toxic spike.

Figure 12 presents the results and reveals a hierarchy of political returns. Toxic spikes targeting issues or policies are associated with the largest increase in voting intentions (5.5 percentage points; p-value = 0.020), followed by targeting individuals (2.9 percentage points; p-value = 0.020) and political groups (2.2 percentage points; p-value = 0.124). In stark contrast, spikes targeting social groups are associated with a qualitative decrease in voter support (-2.6 percentage points; p-value = 0.262) which is statistically different from the positive returns of targeting any of the three alternative categories, marking it as a relatively ineffective strategy.

Figure 12: Effect of toxicity by target of the toxic spike



Notes: The figure displays coefficients and 95% confidence intervals of a regression of voting intentions on an indicator for being surveyed in the week following a toxic spike interacted with the target of the toxic spike. Spikes are defined as days in which the average PAI toxicity score of a politician exceeds the 99th percentile of their historical daily distribution. The target of a spike is defined as the target of the most toxic tweet on the spike day. Regressions include respondents' ideological alignment with the incumbent, socio-demographic controls, district, day and day of the week fixed effects. Standard errors are clustered at the electoral district level.

Combined with the heterogeneous online engagement results by target type, these findings rationalize the variation in targeting strategies observed in politicians' toxic tweets. The data appears to be consistent with a model of strategic communication where politicians trade-off online and offline incentives. A majority of toxic tweets target individuals (63%) precisely because this category is associated with the strongest engagement returns and relatively high electoral support. Conversely, targeting social groups is the least frequent strategy (6%) as it generates minimal online engagement and is associated, if anything, with a decrease in

voting intentions. Finally, while targeting issues or policies is associated with the strongest increase in voter support, it produces the lowest online engagement returns. This combination explains why this targeting strategy is used sparingly (14%) as politicians may be trading off the rewards of an electorally superior strategy against its limited ability to capture online attention.

4.3 Robustness Checks

I perform several tests to ensure the main finding is robust to various issues, including potential threats to identification, alternative measurement choices, and concerns about sample composition.

First, I address the concern that the documented increase in voting intentions following toxic spikes may merely be due to random fluctuations in the data. It could be that, by "pure luck", respondents surveyed after the spike happen to be slightly more supportive of the incumbent for reasons entirely unrelated to the politician's toxic rhetoric. It could also be that toxic spikes coincidentally occur on the same day as a local news shock that favors the incumbent (e.g., win of a local sports team).

To rule out such alternative explanations, I conduct the following placebo test. For each district in the sample, I generate 1,000 sets of placebo spike dates. These are randomly drawn outside a four-week window around any true toxic spike in the district to prevent contamination of respondents in the placebo samples. I then re-estimate the main specification for each of these 1,000 samples. Figure B4 plots the distribution of these 1,000 placebo coefficients against the main estimate represented by the vertical red line. The 2.4 percentage point net increase in voting intentions for the incumbent lies in the extreme right tail of the placebo distribution, with only one draw out of 1000 producing an estimate above that value. This procedure is reproduced for the three binary choice outcomes and results largely corroborate my main findings. The associated placebo p-values are 0.001 (i.e. one draw out of 1000) for the choice of voting for the incumbent versus being undecided, 0.143 for the choice of voting for the incumbent versus the opponent and 0.004 for voting for the opponent versus being undecided. These results enable to rule out that the observed changes in voting intentions following toxic spikes are mere artifacts due to spurious correlations or random fluctuations in the survey data.

Second, and relatedly, one may still be concerned about the presence of omitted variables such as concomitant shocks that co-determine the evolution of voting intentions and the production of online toxic language by politicians. For instance, toxic spikes could be produced in reaction to specific events covered in the news, or to tweets by other prominent politicians (e.g. presidents, governors or candidates to presidential and gubernatorial offices) and voting intentions could change following such events. Despite the inclusion of day fixed

effects removing such co-variation between national events and voting intentions, there could still be residual variation between parties that may bias the main estimates upward. For instance, there could be a differential response of politicians in terms of toxicity production to these events depending on their party. This could be the case if Republican incumbents follow suit on Trump's toxic tweeting, and that voting intentions for Republican incumbents improve due to Trump's tweeting. In order to rule out differential unobserved party trends between politicians that may relate both to toxicity production and voting intentions, I estimate the main specifications additionally including party-by-day fixed effects. The results are displayed in Table B6 and are qualitatively similar to the main results. This suggests that unobserved differential party trends do not play a role in explaining the main effect of toxicity production on voting intentions.

Third, to ensure the finding is not an artifact of a single measurement choice, I test its robustness to an alternative definition of a toxic spike based on the custom LLM annotations. I define a spike as a day where a politician's share of LLM-classified toxic tweets exceeds the 99th percentile of their own historical distribution of this daily share. The event study for this alternative measure is presented in Figure B5. The results show a similar dynamic pattern to the main analysis, with a positive jump in voting intentions following the spike. The average post-spike effect is attenuated to 0.9 percentage points (p=0.072), which is expected given that this spike definition, derived from a binary classification, is a less precise measure of communication intensity than the one based on the continuous PAI score. Nevertheless, the consistency of the dynamic pattern provides confidence that the result is not specific to the PAI measure.

Finally, to address concerns about statistical power and the representativeness of the survey data at the district level, I examine the stability of the main effect in districts with a sufficiently large number of survey respondents. Figure B6 plots the main coefficient when the sample is progressively restricted to districts with at least 20, 30, 40, and 50 respondents interviewed in the two-week window around a spike. The point estimate remains stable and, if anything, increases in magnitude as the sample is restricted to higher-density districts. While the confidence intervals mechanically widen with the narrowing of the sample size, the stability of the estimate confirms that the main result is not driven by noisy, low-sample districts and holds in the parts of the sample where statistical power is greatest.

4.4 Mechanisms

Having established that voting intentions display a net increase in the week following a politician's toxic spike, this section turns to investigating the mechanisms driving this result. I proceed by first testing the specificity of toxicity against alternative forms of salient communication, and then provide suggestive evidence that consuming news through social media

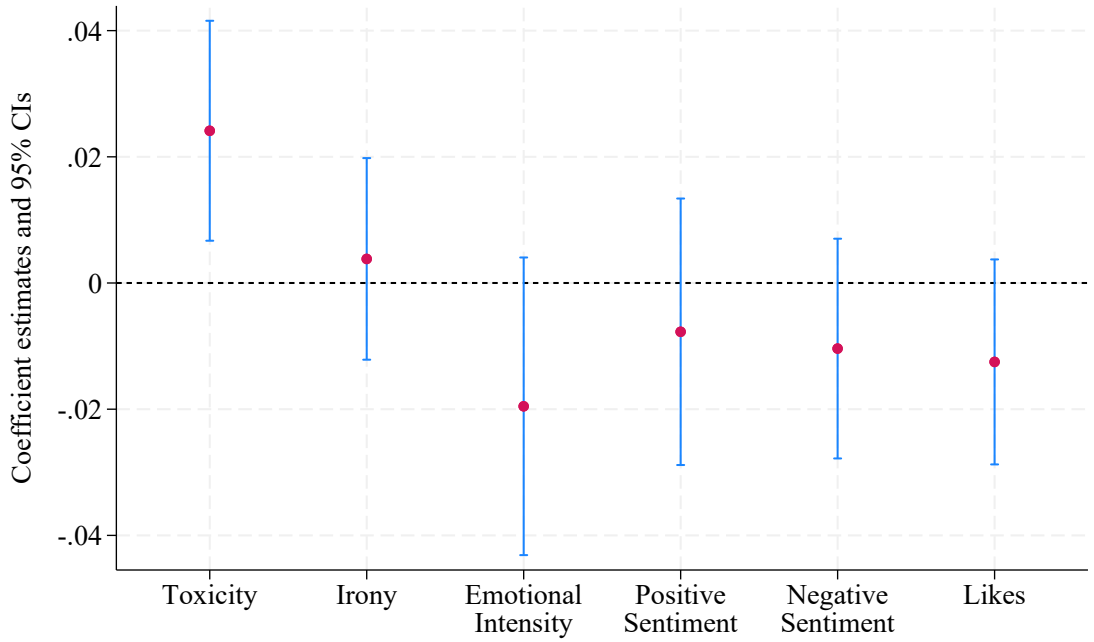
acts as a transmission channel between toxic spikes and changes in political attitudes.

Salience of toxic rhetoric One leading explanation for the main finding is that voters respond not to toxicity *per se*, but to any unusually salient deviation from a politician's typical communication. To test this hypothesis, I construct "placebo spikes" for other dimensions of online rhetoric, including irony, emotional intensity, and sentiment. Consistently with the methodology used to identify toxic spikes, I define such spikes as days where the continuous score associated to the respective feature exceeds the 99th percentile of the politician's own daily distribution. Symmetrically, a spike in negative sentiment is defined as days in which the VADER sentiment score is below the 1st percentile of the politician's historical daily distribution. I then re-estimate the main specification, replacing the Post Toxic Spike indicator with a similar indicator for each alternative spike.

Figure 13 presents the results of this exercise. The estimates for spikes in irony, emotional intensity, and both positive and negative sentiment are statistically indistinguishable from zero. Furthermore, to disentangle the effect of toxic *content* from the online *visibility* it generates, I also test for the effect of a spike in a politician's daily average number of likes. According to this alternative interpretation, it could be that the changes in voting intentions are not due to salient toxic communication but are the result of the visibility boost generated by toxicity. By examining how voting intentions change following a spike in the number of likes received by tweets – regardless of stylistic features – I can test this alternative hypothesis. In doing so, I find, again, a null effect. Taken together, these results suggests that voters are not merely responding to unusual political communication or engagement, but rather to distinct violations of civic norms by political officials.

However, that voting intentions remain unchanged following a spike in the number of likes does not imply that visibility is completely irrelevant. Instead, I provide suggestive evidence that engagement amplifies the response of voters to abnormally toxic communication. To do so, I rerun the main specification including an interaction term between the Post Toxic Spike indicator and a measure for the strength of the online engagement generated by the toxic spike. Specifically, I build a variable measuring the difference between the number of likes generated by the most toxic tweet in a toxic spike and the politician's average number of likes at baseline. Figure B7 shows the effect of a toxic spike by tertiles of this variable, i.e. depending on whether the level of engagement generated by the spike is low, medium or strong compared to politicians' baseline engagement. While the point estimates are not statistically different from each other, the pattern is qualitatively clear. Toxic spikes built on relatively low-engagement tweets have a minimal effect, while the impact on voting intentions grows monotonically with the relative engagement level of the toxic message. This result suggests that toxic spikes garnering higher online visibility induce stronger shifts in political attitudes.

Figure 13: Changes in voting intentions following incumbents' spikes in other textual features



Notes: The figure displays coefficients and 95% confidence intervals of regressions of respondents' voting intentions on an indicator for whether the respondent is surveyed in the week following their House representative's spikes in other textual features of tweets. Spikes are defined as days in which the average feature use in tweets by the incumbent exceeds the 99th percentile of their historical daily distribution. Spikes in negative sentiment are defined as days in which the VADER sentiment score is below the 1st percentile of the politician's historical daily distribution. Regressions include respondents' ideological alignment with the incumbent, socio-demographic controls, district, day and day of the week fixed effects. Standard errors are clustered at the electoral district level.

That toxic communication needs to be salient in order to move voters is further supported by voting intentions changing only following the most intense toxic spikes. Figure B8 shows that when a spike is defined using a milder threshold, namely the 95th percentile of a politician's historical daily toxicity distribution, the subsequent change in voting intentions is substantially attenuated (point estimate = 0.008; p-value = 0.046). It even vanishes when toxic spikes are defined using the 90th percentile of a politician's historical distribution (point estimate = 0.002; p-value = 0.585). Taken together, these findings indicate that the salience of politicians' toxic rhetoric needs to be strong enough whether in terms of visibility or intensity to induce changes in voting intentions.

Social media exposure as the transmission channel So far, the estimates presented in this section capture an intention-to-treat effect of toxicity on voter attitudes, as accurate information about voters' Twitter usage and exposure to their representative's toxicity remains unobserved. In order to test whether the observed changes in voting intentions are mediated by voters' exposure to politicians' toxic communication on social media, I use information

about respondents' news diet contained in the survey.

Respondents are asked to self-report the sources they use for political news through the following multiple answer choice question: "Have you seen or heard news about politics on any of the following outlets in the past week?". Choices include "social media (e.g. Facebook, Twitter)" as well as 11 other news sources ranging from specific Cable TV channels (CNN, MSNBC, FoxNews) to local TV stations, public TV broadcast networks, local newspapers, national newspapers, and radio. 71.4% of respondents report that one of their news sources is social media meaning that it is often bundled together with other news sources in respondents' news diets. In order to isolate social media consumption away from other types of news sources, I build a three-point categorical variable to proxy for the intensity of social media use for political news with groups corresponding to: (i) no use of social media for political news; (ii) some use when the respondent lists social media as one of several sources; and (iii) exclusive use when respondents' only source is social media. I include this variable in the main estimating equation and report the estimates related to its interaction with the indicator variable for being interviewed in the week following a toxic spike. To benchmark the moderating influence of social media against other major sources of news, I run a similar specification using the intensity of exposure to TV for political news. This is particularly relevant as television sources – whether cable, public broadcast or local – are present in 84.3% of respondents' news diets.[24]

Figure 14 presents the results of these interacted specifications. The findings reveal two key patterns that support a social media-based transmission mechanism when taken together.

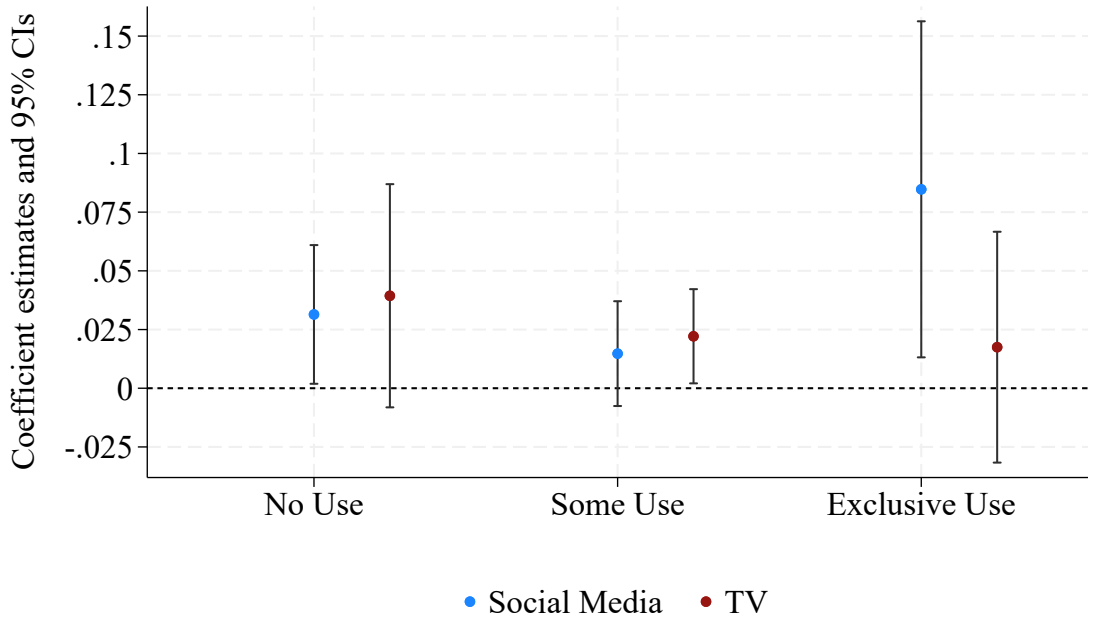
First, the increase in voting intentions following a toxic spike is largest for respondents who report getting their political news exclusively from social media. For this group, a toxic spike increases the intention to vote for the incumbent by 8.5 percentage points, more than double that of respondents who do not use social media for news purposes or only partly. This indicates that social-media single-homers, i.e. voters who are less exposed to countervailing information from traditional news sources, are most responsive to politicians' toxic online communication, consistent with the social media exposure mechanism.[25]

Second, the results for TV consumption serve as a compelling placebo, allowing to rule out a an alternative "single-homing" explanation where any voter who relies on a single news

[24]Radio and newspapers are reported as sources of political news in respectively 30.8% and 47.9% of respondents' news diets. Because they are more marginal sources, I focus solely on comparing social media to television.

[25]Interestingly, however, the response does not appear to be an increasing function of "treatment intensity". The change in voting intentions for respondents with a mixed news diet that includes social media is indistinguishable from zero (point estimate = 0.015; p-value = 0.194), and qualitatively smaller than for those who do not rely on social media for political news (point estimate = 0.031; p-value = 0.037). This pattern is consistent with a model where voters in a richer information environment are able to contextualize or discount online rhetoric against information from several offline sources. The fact that respondents with a mixed social media diet report on average 5.24 sources for political news, versus 2.98 for individuals who do not use social media for information purposes, supports this argument.

Figure 14: Changes in voting intentions following incumbents' toxic spike by respondents' news diet



Notes: The figure displays coefficients and 95% confidence intervals of regressions of respondents' voting intentions on an indicator for being surveyed in the week following a toxic spike interacted with exposure intensity to social media (in blue), or to television news (in red). For each medium (social media or TV), respondents are classified as not relying on the medium for political information purposes, reporting the medium as one of their information sources among others, or relying exclusively on the medium for news. Regressions include respondents' baseline exposure intensity to the medium, the number of total sources used for political information, ideological alignment with the incumbent, socio-demographic controls, district, day and day of the week fixed effects. Standard errors are clustered at the electoral district level.

source is more easily persuaded. In stark contrast to social media, respondents who rely solely on television for political news show no significant change in their voting intentions following a toxic spike on Twitter. This null result for exclusive TV viewers reinforces the specificity of the social media channel. The effect is not driven by media diet concentration alone, but by exposure to the specific medium on which politicians supply toxic communication.

5 Conclusion

Social media is transforming political communication. By lowering the cost of direct outreach and rewarding content that captures attention, these platforms alter how politicians communicate with voters and modify the incentives that govern their rhetorical choices, namely the use of offensive language.

This paper quantifies the strength of these incentives and examines how political context and both supply- and demand-side factors shape their magnitude. I identify two key sources of private returns: (i) online attention gains, reflected in substantially higher engagement generated by toxic posts, and (ii) offline electoral rewards evidenced by an increase in voting intentions following toxic spikes in politicians' tweeting activity. Taken together, these findings highlight that toxicity is a rhetorical tool that politicians may use to generate attention on social media, and that they can trade this for electoral support, albeit at the cost of increased polarization.

The magnitude of these returns varies systematically with political market conditions—such as the competitiveness of races, politicians' demographic profiles, and the rhetorical style of their communication—in ways that mirror observed usage patterns of offensive rhetoric. For instance, engagement and electoral gains are highest for younger and electorally-secure politicians, who are also more likely to employ toxic language. Yet, offline incentives alone cannot fully rationalize the cross-sectional variation in toxicity. This pattern points to an equilibrium where politicians optimize over multiple objectives, balancing immediate attention gains against the costs and benefits of electoral persuasion.

Beyond increasing incentives for attention, the rise of social media is reshaping the informational relationship between voters and politicians. By enabling continuous and often personal interactions between politicians and voters, these platforms generate real-time feedback that can shape politicians' communication behavior. Understanding whether such feedback mechanisms enhance political accountability or instead reinforce populist and attention-maximizing strategies is a promising direction for future research. Pursuing it will require research designs able to isolate learning dynamics from endogenous communication choices.

Finally, while this paper documents the private returns accruing to politicians, the long-run social costs of normalizing offensive rhetoric remains an open question. Negative externalities may extend beyond the polarization of voter attitudes. For instance, the sizable attention brought to incivility may erode citizens' trust in democratic institutions, or alter political selection by creating barriers to entry for candidates unwilling to engage in such behavior. Understanding the broader welfare impacts of offensive speech requires quantifying these systemic consequences.

References

Algan, Y., E. Davoine, T. Renault, and S. Stantcheva (2025): "Emotions and policy views," Tech. rep., Harvard University Working Paper.

Ansolabehere, S., S. Iyengar, A. Simon, and N. Valentino (1994): "Does attack advertising demobilize the electorate?" *American political science review*, 88, 829–838.

Ansolabehere, S. D., S. Iyengar, and A. Simon (1999): "Replicating experiments using aggregate and survey data: The case of negative advertising and turnout," *American political science Review*, 93, 901–909.

Auter, Z. J. and J. A. Fine (2016): "Negative campaigning in the social media age: Attack advertising on Facebook," *Political Behavior*, 38, 999–1020.

Ballard, A. O., R. DeTamble, S. Dorsey, M. Heseltine, and M. Johnson (2023): "Dynamics of polarizing rhetoric in congressional tweets," *Legislative Studies Quarterly*, 48, 105–144.

Beknazar-Yuzbashev, G., R. Jiménez-Durán, J. McCrosky, and M. Stalinski (2025): "Toxic content and user engagement on social media: Evidence from a field experiment," Tech. rep., CESifo Working Paper.

Beknazar-Yuzbashev, G., R. Jiménez-Durán, and M. Stalinski (2024): "A model of harmful yet engaging content on social media," in *AEA Papers and Proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 114, 678–683.

Bessone, P., F. R. Campante, C. Ferraz, and P. Souza (2022): "Social Media and the Behavior of Politicians: Evidence from Facebook in Brazil," Tech. rep., National Bureau of Economic Research.

Boken, J., M. Draca, N. Mastrorocco, and A. Ornaghi (2023): "The returns to viral media: the case of US campaign contributions," *CEPR DP18337*.

Boxell, L., J. Conway, J. N. Druckman, and M. Gentzkow (2022): "Affective Polarization Did Not Increase During the COVID-19 Pandemic," *Quarterly Journal of Political Science*, 17, 491–512.

Brady, W. J., J. A. Wills, D. Burkart, J. T. Jost, and J. J. Van Bavel (2019): "An ideological asymmetry in the diffusion of moralized content on social media among political leaders." *Journal of Experimental Psychology: General*, 148, 1802.

Brady, W. J., J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel (2017): "Emotion shapes the diffusion of moralized content in social networks," *Proceedings of the National Academy of Sciences*, 114, 7313–7318.

Brooks, D. J. (2010): "A negativity gap? Voter gender, attack politics, and participation in American elections," *Politics & Gender*, 6, 319–341.

Brooks, D. J. and J. G. Geer (2007): "Beyond negativity: The effects of incivility on the electorate," *American Journal of Political Science*, 51, 1–16.

Burtch, G., Q. He, Y. Hong, and D. Lee (2022): "How do peer awards motivate creative content? Experimental evidence from Reddit," *Management Science*, 68, 3488–3506.

Camacho-Collados, J., K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. Espinosa-Anke, F. Liu, E. Martínez-Cámara, et al. (2022): "TweetNLP: Cutting-Edge Natural Language Processing for Social Media," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, U.A.E.: Association for Computational Linguistics.

Cao, A., J. M. Lindo, and J. Zhong (2023): "Can social media rhetoric incite hate incidents? Evidence from Trump's "Chinese Virus" tweets," *Journal of Urban Economics*, 137, 103590.

Crockett, M. J. (2017): "Moral outrage in the digital age," *Nature human behaviour*, 1, 769–771.

Dai, Y. and A. Kustov (2022): "When do politicians use populist rhetoric? Populism as a campaign gamble," *Political Communication*, 39, 383–404.

Dell, M. (2025): "Deep learning for economists," *Journal of Economic Literature*, 63, 5–58.

Djupe, P. A. and D. A. Peterson (2002): "The impact of negative campaigning: Evidence from the 1998 senatorial primaries," *Political Research Quarterly*, 55, 845–860.

Eckles, D., R. F. Kizilcec, and E. Bakshy (2016): "Estimating peer effects in networks with peer encouragement designs," *Proceedings of the National Academy of Sciences*, 113, 7316–7322.

Ederer, F., P. Goldsmith-Pinkham, and K. Jensen (2024): "Anonymity and identity online," *arXiv preprint arXiv:2409.15948*.

Finkel, S. E. and J. G. Geer (1998): "A spot check: Casting doubt on the demobilizing effect of attack advertising," *American journal of political science*, 573–595.

Freedman, P. and K. Goldstein (1999): "Measuring media exposure and the effects of negative campaign ads," *American journal of political Science*, 1189–1208.

Frimer, J. A., H. Aujla, M. Feinberg, L. J. Skitka, K. Aquino, J. C. Eichstaedt, and R. Willer (2023): "Incivility is rising among American politicians on Twitter," *Social Psychological and Personality Science*, 14, 259–269.

Frimer, J. A. and L. J. Skitka (2018): "The Montagu Principle: Incivility decreases politicians' public approval, even with their political base." *Journal of Personality and Social Psychology*, 115, 845.

Galasso, V., T. Nannicini, and S. Nunnari (2023): "Positive spillovers from negative campaigning," *American Journal of Political Science*, 67, 5–21.

Goda, S., N. Agata, and Y. Matsumura (2020): "A stacking ensemble model for prediction of multi-type tweet engagements," in *Proceedings of the Recommender Systems Challenge 2020*, 6–10.

Grootendorst, M. (2022): "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*.

Grosjean, P., F. Masera, and H. Yousaf (2023): "Inflammatory political campaigns and racial bias in policing," *The Quarterly Journal of Economics*, 138, 413–463.

Hutto, C. and E. Gilbert (2014): "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, 216–225.

Jiménez Durán, R., K. Müller, and C. Schwarz (2024): "The effect of content moderation on online and offline hate: Evidence from Germany's NetzDG," *Available at SSRN 4230296*.

Jungherr, A., G. Rivero, G. R. Rodríguez, and D. Gayo-Avello (2020): *Retooling politics: How digital media are shaping democracy*, Cambridge University Press.

Kalra, A. (2025): "Hate in the Time of Algorithms: Evidence on Online Behavior from a Large-Scale Experiment," *arXiv preprint arXiv:2503.06244*.

Korinek, A. (2023): "Generative AI for economic research: Use cases and implications for economists," *Journal of Economic Literature*, 61, 1281–1317.

Krupnikov, Y. (2011): "When does negativity demobilize? Tracing the conditional effect of negative campaigning on voter turnout," *American Journal of Political Science*, 55, 797–813.

Lau, R. R. and G. M. Pomper (2004): *Negative campaigning: An analysis of US Senate elections*, Rowman & Littlefield.

Leung, B. T. K. and P. Yildirim (2020): "Competition, Politics, & Social Media," *arXiv preprint arXiv:2012.03327*.

Ludwig, J., S. Mullainathan, and A. Rambachan (2025): "Large language models: An applied econometric framework," Tech. rep., National Bureau of Economic Research.

McCarthy, P. M. and S. Jarvis (2010): "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment," *Behavior research methods*, 42, 381–392.

Messing, S., P. van Kessel, A. G. Hughes, N. Judd, R. Blum, and B. Broderick (2017): "Partisan Conflict and Congressional Outreach," *Pew Research Center*.

Müller, K. and C. Schwarz (2023a): "The Effects of Online Content Moderation: Evidence from President Trump's Account Deletion," *Available at SSRN 4296306*.

——— (2023b): "From hashtag to hate crime: Twitter and antiminority sentiment," *American Economic Journal: Applied Economics*, 15, 270–312.

Mummalaneni, S., H. Yoganarasimhan, and V. Pathak (2022): "How Do Content Producers Respond to Engagement on Social Media Platforms?" *Available at SSRN 4173537*.

Narayanan, A. (2023): "Understanding social media recommendation algorithms," .

Petrova, M., A. Sen, and P. Yildirim (2021): "Social media and political contributions: The impact of new technology on political competition," *Management Science*, 67, 2997–3021.

Pew Research Center (2024): "Most Americans say elected officials should avoid heated or aggressive speech," https://www.pewresearch.org/short-reads/2024/01/31/most-americans-say-elected-officials-should-avoid-heated-or-aggressive-speech/, accessed: 2025-09-23.

Pfeffer, J., D. Matter, and A. Sargsyan (2023): "The half-life of a tweet," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, 1163–1167.

Poole, K. T. and H. Rosenthal (1985): "A spatial model for legislative roll call analysis," *American journal of political science*, 357–384.

Rathje, S., J. J. Van Bavel, and S. Van Der Linden (2021): "Out-group animosity drives engagement on social media," *Proceedings of the National Academy of Sciences*, 118, e2024292118.

Schöll, N., A. Gallego, and G. Le Mens (2023): "How Politicians Learn from Citizens' Feedback: The Case of Gender on Twitter," *American Journal of Political Science*.

Tumasjan, A., T. Sprenger, P. Sandner, and I. Welpe (2010): "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Proceedings of the international AAAI conference on web and social media*, vol. 4, 178–185.

Zhang, L., G. Gratton, P. Grosjean, and H. Yousaf (2025): "Adding Fuel to the (Gun) Fire: How Politicians Polarize the Public Debate," *Center for Law & Economics Working Paper Series*, 1.

A Tweet and congress member sample descriptions

Table A1: Examples of non-toxic and highly toxic tweets

Text	Toxicity (PAI)	Toxicity (LLM)	Target
If you can't win, cheat. This tactic is completely **despicable** and prevents the actual Census – the one required by our Constitution – from doing its job.	0.33	1	Issue/Policy
Another day, another attempt by @HouseDemocrats to conceal the truth from the American people. This impeachment inquiry is a kangaroo court. Imagine being accused of a crime and not having the right to defend yourself or the ability to confront your accuser. **Just shameful!**	0.36	1	Political Group
Barr said the words "no collusion, no collusion" straight out of the mouth of Trump. No attempt to hide his continuing defense of the President using the President's own words. How much more of a **lackey** can he be? **FAKE AG!**	0.44	1	Individual
Another of the thousands upon thousands of illegal alien felons who have reentered the U.S. We must secure America's borders and protect our citizens from **child-molesting scum** like this guy.	0.69	1	Social Group
Excited to announce a $4,772,000 @USDA grant for the Village of Heuvelton. This funding will rehabilitate and upgrade the village's wastewater treatment facilities. Proud to advocate for #NY21! Read more	0.01	0	NA
Let's keep our successful COVID-19 response up and cases down. Remember to wear your mask and social distance this long weekend.	0.03	0	NA
Not only does this bill not include water infrastructure, but it also only prioritizes urban areas - leaving rural communities like ours in the dust.	0.02	0	Issue/Policy
I read the whistleblower complaint this morning. The report clearly outlines a blatant attempt to cover-up President Trump's communications with Ukraine and a pattern of obstruction at the White House. Anyone involved in this cover-up should be immediately investigated.	0.04	0	Individual

Notes: The table presents a selection of tweets written by U.S. congress members with varying degrees of toxicity as measured by Perspective API's toxicity detection model. LLM annotations for the presence of toxicity and the target category are also displayed.

Table A2: Summary statistics of tweet engagement metrics

	Count	Mean	St.Dev.	Min	P10	P25	P50	P75	P90	Max
Likes	2.87M	875.7	7046.6	0	3	8	24	122	830	2.95M
Replies	2.87M	95.0	761.1	0	0	1	5	20	99	206K
Retweets	2.87M	193.0	1351.5	0	1	2	7	30	193	498K
Quotes	2.87M	18.3	202.9	0	0	0	1	3	16	156K
Total Engagement	2.87M	1182.1	8828.8	0	5	13	39	184	1.2K	3.70M

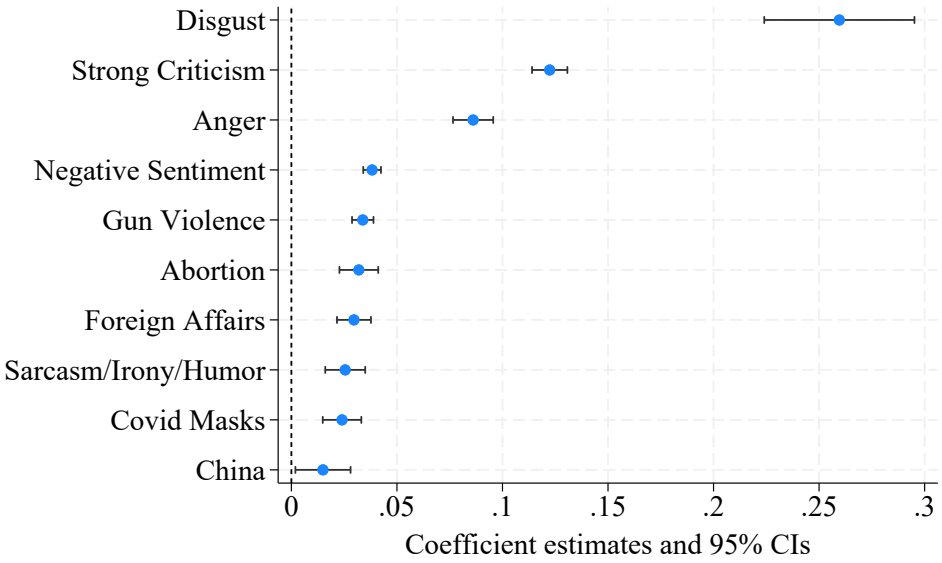
Notes: The sample includes all tweets posted by members of the U.S. congress between June 21st, 2017 and December 31st, 2022. Total engagement of a tweet is computed as the sum of the counts across all four engagement metrics.

Table A3: Summary statistics of LLM-annotated tweet features

Variable	Mean	N
Toxicity (LLM)	0.073	744643
Criticism/Disagreement	0.346	741280
Strong Criticism	0.447	245802
Target of Criticism		
Individual	0.509	250351
Political Group	0.189	250351
Social Group	0.024	250351
Issue/Policy	0.278	250351
Sarcasm/Irony/Humor	0.058	743547
Emotive Appeals	0.505	738504
Emotion Category		
Anger	0.252	371633
Hope	0.198	371633
Joy	0.196	371633
Pride	0.142	371633
Gratitude	0.117	371633
Sadness	0.057	371633
Fear	0.031	371633
Disgust	0.006	371633

Notes: This table presents the distribution of the seven LLM-annotated tweet features, following the procedure detailed in Appendix C.1. Annotation is performed on all tweets posted by House Representatives between January 3, 2019 and November 2, 2020. When annotated as "unsure", variables are set to missing. The strength and target of criticism are annotated only for tweets marked as expressing criticism/disagreement. Emotion category is annotated only for tweets expressing an emotive appeal.

Figure A1: Strongest Tweet-Level Predictors of Toxicity



Notes: The figure displays coefficients from a regression of the PAI toxicity indicator on all observed tweet-level textual features, save for LLM-annotated toxicity. Only estimates for the 10 strongest positive predictors of PAI toxicity are displayed. The dependent variable is a binary measure equal to one if the Perspective API (PAI) score of a tweet is above an optimal data-driven threshold (0.26). Regressions include member fixed effects and week fixed effects. Standard errors are clustered at the congress member level. The full list of features includes: presence and strength of criticism, presence of sarcasm, expression of emotive appeals, detailed emotion in case of emotional appeal, positive and negative sentiment, high and low lexical diversity, tweet topic, tweet being posted on a weekend, tweet being posted during the last three months of the electoral campaign, indicators for whether the tweet contains a hashtag, a mention, an emoji, a URL, media content, whether it is a quote tweet, whether it is posted by the campaign or the official account, and the number of words and verbs.

Table A4: Congress member summary statistics

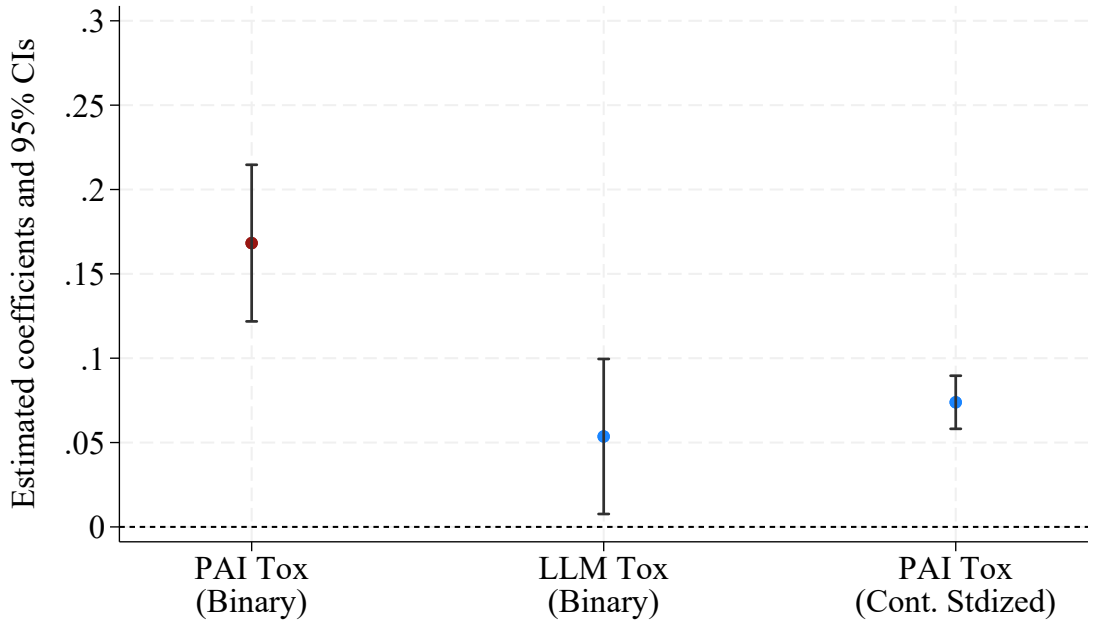
	Count	Mean	St.Dev.	Min	P10	P25	P75	P90	Max
Politician features									
Female	1639	24.6							
Afro-American	1639	10.6							
Hispanic	1639	7.3							
Age	1634	59.0	11.5	26	43	51	67	73	88
Terms served	1639	5.5	4.4	1	1	2	7	12	27
Republican	1628	50.4							
Vote margin	1593	30.5	22.8	0	6	14	41	61	100
Tweet activity									
Toxicity score	1639	0.05	0.03	0.01	0.02	0.03	0.06	0.08	0.22
Total tweets	1639	1823.0	1573.9	1	366	722	2500	3845	15010
Mean likes	1586	565.9	2142.9	0	12	21	201	1017	30901
Mean replies	1586	69.2	228.9	0	2	5	32	123	3422
Mean retweets	1586	134.0	497.2	0	4	7	53	242	8464
Mean quotes	1586	13.6	43.8	0	1	1	6	28	601
Campaign account (%)	1639	19.9	23.0	0	0	2	30	53	100

Notes: The table displays summary statistics for congress members' main socio-demographic and political characteristics, as well as features related to their tweeting activity. Variables are aggregated at the congress-member-by-congress-session level.

B Additional Results

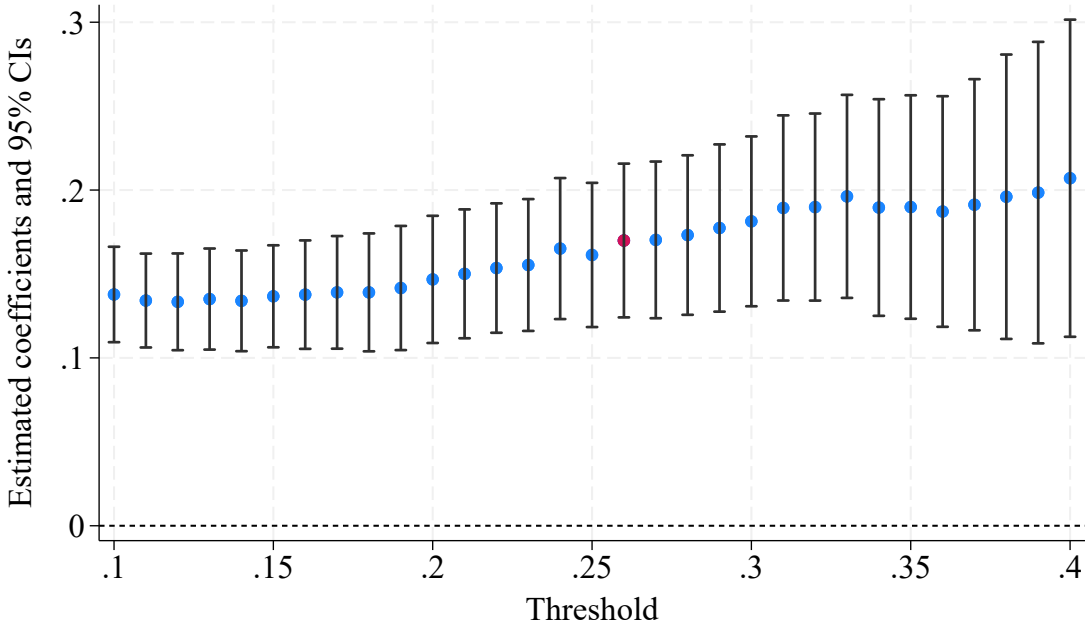
B.1 Online engagement and toxicity

Figure B1: Toxicity and online engagement - Robustness to alternative definitions of toxicity



Notes: Point estimates and 95% CIs associated to alternative definitions of tweet toxicity are plotted. The first coefficient corresponds to the main toxicity indicator, i.e. a binary measure equal to one if the Perspective API (PAI) score of a tweet is above an optimal data-driven threshold (0.26). The second corresponds to the toxicity indicator variable outputted by the custom LLM annotation for political incivility. The third corresponds to the standardized continuous PAI score. The dependent variable is the IHS-transformed number of likes. All regressions are separately estimated using the preferred specification, including the full set of tweet-level controls as detailed in Section 3 and politician-by-week fixed effects. Standard errors are clustered at the politician-level.

Figure B2: Toxicity and online engagement - Robustness to using alternative thresholds to discretize PAI toxicity



Notes: Point estimates and 95% CIs associated to the binary Perspective API (PAI) toxicity variable are plotted. Each point is estimated from a separate regression where the toxicity indicator is defined using a different PAI score threshold, ranging from 0.10 to 0.40. The estimate related to the toxicity indicator using the data-driven optimal threshold (0.26) is marked in red. The dependent variable is the IHS-transformed number of likes. All regressions are estimated using the preferred specification, including the full set of tweet-level controls as detailed in Section 3 and politician-by-week fixed effects. Standard errors are clustered at the politician-level.

Table B1: Tweet engagement and tweet toxicity – Alternative engagement metrics

	(1) IHS(Engagement)	(2) IHS(Retweets)	(3) IHS(Replies)	(4) IHS(Quotes)
Toxicity=1	0.162***	0.141***	0.174***	0.147***
	(0.023)	(0.024)	(0.022)	(0.025)
Tweet controls	Yes	Yes	Yes	Yes
Day of week FEs	Yes	Yes	Yes	Yes
Congress member × Week FEs	Yes	Yes	Yes	Yes
Observations	669,726	669,726	669,726	669,726
Number of congress members	386	386	386	386
R squared	0.65	0.64	0.64	0.57
Mean raw DV	1017.92	178.78	72.59	15.48

Notes: The table presents results from regressions of tweets' engagement metrics on an indicator for tweet toxicity. The toxicity indicator is equal to one if a tweet's Perspective API (PAI) score is above an optimal data-driven threshold (0.26). Dependent variables are IHS-transformed total number of engagement, i.e. sum of likes, retweets, replies and quote tweets (1), retweets only (2), replies only (3), and quote tweets only (4) received by tweets. All regressions are estimated using the preferred specification, including the full set of tweet-level controls as detailed in Section 3 and politician-by-week fixed effects. Standard errors are clustered by politician. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B2: Tweet engagement and tweet toxicity – Poisson specification

	(1)	(2)	(3)
	Number of likes		
Toxicity=1	0.229***	0.245***	0.257***
	(0.054)	(0.042)	(0.041)
Tweet controls	Yes	Yes	Yes
Congress member FEs	Yes	No	No
Week-by-year FEs	Yes	No	No
Day of week FEs	Yes	Yes	Yes
Congress member × Week FEs	No	Yes	Yes
Congress member × Topic FEs	No	No	Yes
Observations	670,888	669,713	669,169
Number of congress members	386	386	386
Pseudo-R^2	0.72	0.78	0.79
Mean Likes	752	751	751

Notes: The table presents results from Poisson regressions of the number of likes received by a tweet on an indicator for tweet toxicity. The toxicity indicator is equal to one if a tweet's Perspective API (PAI) score is above an optimal data-driven threshold (0.26). The dependent variable is the number of likes generated by the tweet. All specifications include the full set of tweet-level controls as detailed in Section 3, including tweet features such as the use of criticism, the target of criticism, the use of sarcasm/irony/humor, appeals to emotions, positive and negative sentiment. Column (1) includes politician and week fixed effects. Column (2) includes politician-by-week fixed effects. Column (3) adds politician-by-topic fixed effects. Standard errors are clustered at the politician level. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

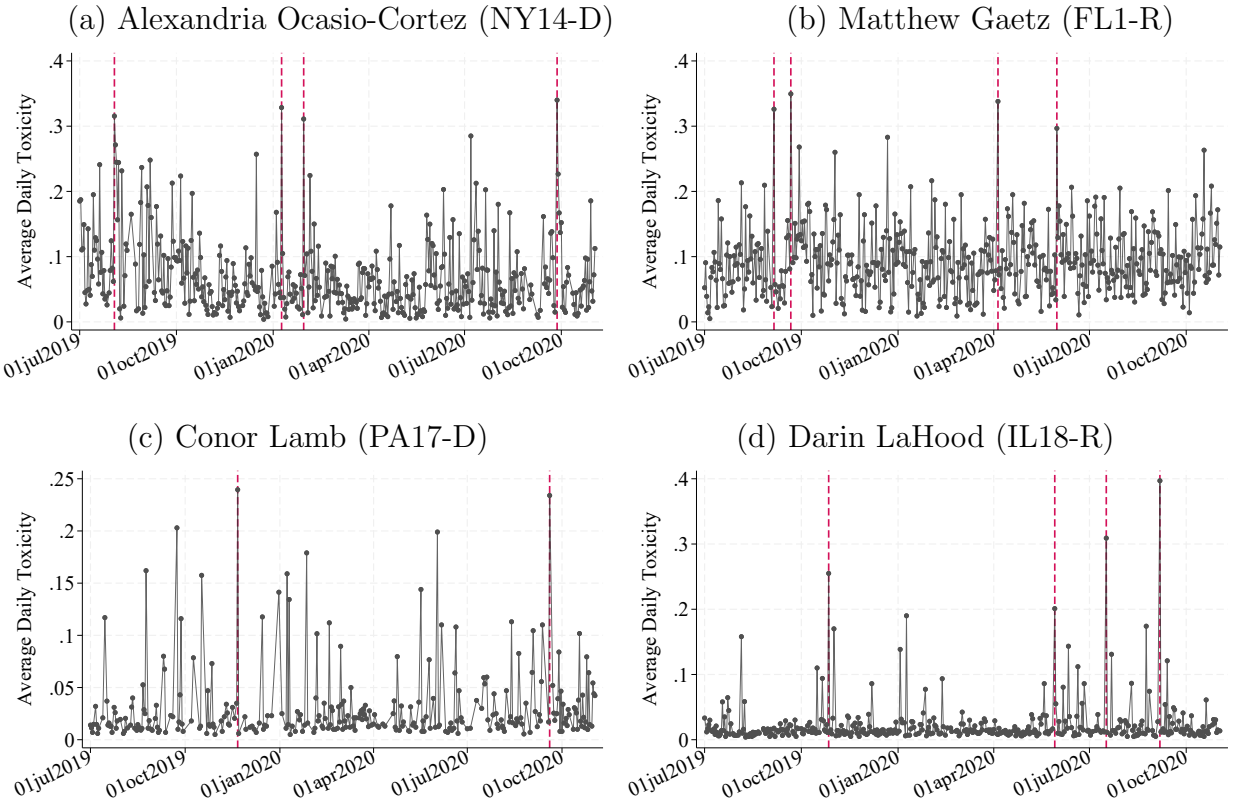
Table B3: Tweet engagement and tweet toxicity – Alternative specifications

	(1)	(2)	(3)
		IHS(Likes)	
Toxicity=1	0.168***	0.168***	0.169***
	(0.026)	(0.024)	(0.021)
Tweet controls	Yes	Yes	Yes
Congress member FEs	Yes	No	No
Week-by-year FEs	Yes	No	No
Day of week FEs	Yes	Yes	Yes
Congress member × Week FEs	No	Yes	Yes
Congress member × Topic FEs	No	No	Yes
Observations	670,888	669,726	669,194
Number of congress members	386	386	386
R squared	0.59	0.65	0.67
Mean Likes	751.97	751.07	751.13

Notes: The table presents results from regressions of the IHS-transformed number of likes on an indicator for tweet toxicity. The toxicity indicator is equal to one if a tweet's Perspective API (PAI) score is above an optimal data-driven threshold (0.26). All specifications include the full set of tweet-level controls as detailed in Section 3, including tweet features such as the use of criticism, the target of criticism, the use of sarcasm/irony/humor, appeals to emotions, positive and negative sentiment. Column (1) includes politician and week fixed effects. Column (2) includes politician-by-week fixed effects. Column (3) adds politician-by-topic fixed effects. Standard errors are clustered at the politician level. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

B.2 Voting and toxicity

Figure B3: Evolution of daily toxicity and identification of toxic spikes



Notes: The figure displays the daily average Perspective API (PAI) toxicity score for four selected House Representatives between July 1, 2019 and November 2, 2020. The red vertical lines indicate toxic spikes defined as days where the average toxicity exceeds the 99th percentile of the politician's own historical distribution over the period. The four representatives were selected to represent both parties equally and their average toxicity over the period spreads across the entire distribution. Panels (a) and (b) displays the evolution of daily toxicity for two high toxicity representatives while panels (c) and (d) displays this for two low toxicity representatives.

Table B4: Balance in respondents' socio-demographics following a toxic spike

	(1) Age	(2) Female	(3) College	(4) White	(5) Black	(6) Other race	(7) Hispanic
Post Toxic Spike=1	0.249	0.007	0.004	-0.012	0.003	0.009	-0.006
	(0.461)	(0.014)	(0.012)	(0.010)	(0.008)	(0.009)	(0.009)
District FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day of week FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,525	6,525	6,525	6,525	6,525	6,525	6,525
Number of districts	197	197	197	197	197	197	197
R squared	0.12	0.13	0.13	0.18	0.18	0.16	0.17
Mean dep. var.	44.95	0.60	0.37	0.73	0.12	0.14	0.13

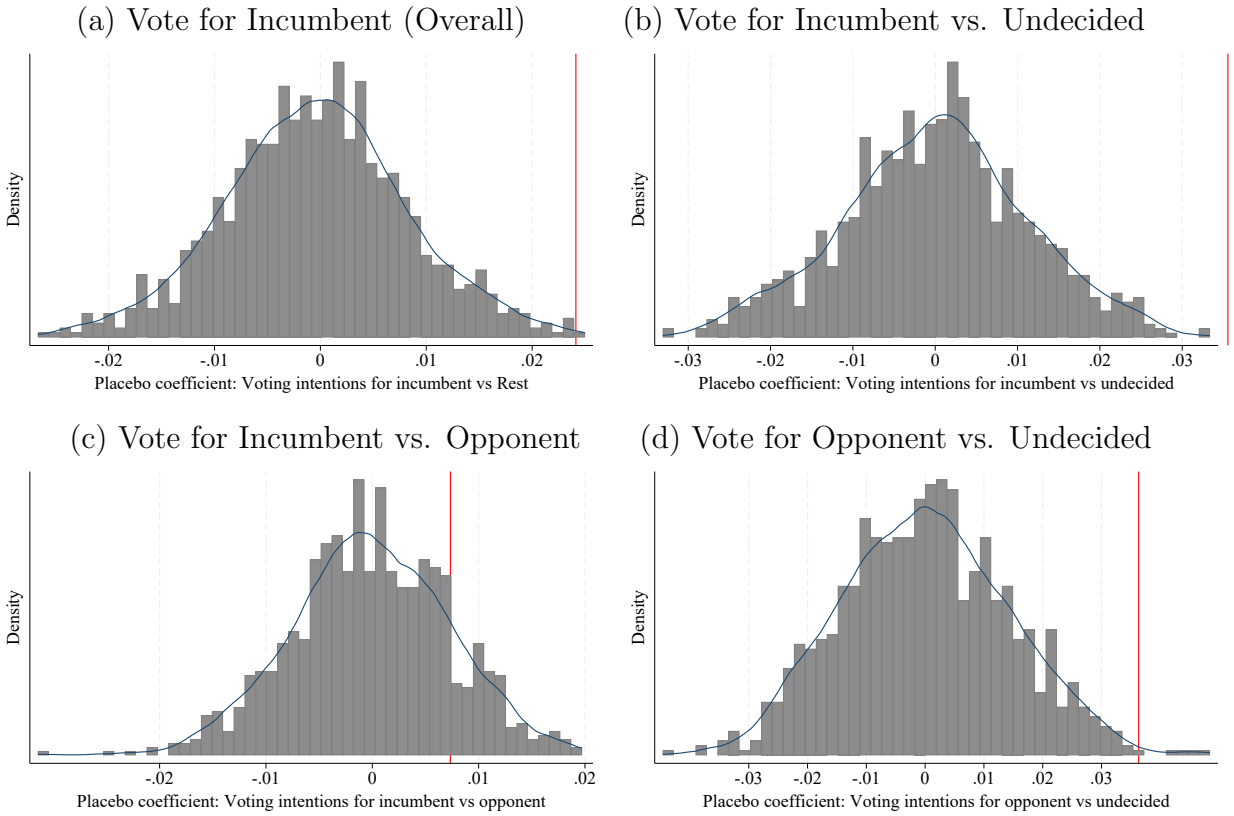
Notes: The table presents results from regressions of respondents' socio-demographic characteristics on an indicator for whether the respondent is surveyed in the week following their House representative's toxic spike. All regressions include district, day and day of the week fixed effects and do not include any socio-demographic controls. Only respondents interviewed within a one week window of toxic spikes are included in the regressions. Standard errors are clustered at the electoral district level. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B5: Balance in respondents' ideological alignment with their incumbent following a toxic spike

	(1) Strong supporter	(2) Weak supporter	(3) Independent	(4) Weak opponent	(5) Strong opponent
Post Toxic Spike=1	0.013	0.009	-0.022***	0.008	-0.008
	(0.012)	(0.011)	(0.008)	(0.011)	(0.011)
District FEs	Yes	Yes	Yes	Yes	Yes
Day FEs	Yes	Yes	Yes	Yes	Yes
Day of week FEs	Yes	Yes	Yes	Yes	Yes
Observations	6,134	6,134	6,134	6,134	6,134
Number of districts	197	197	197	197	197
R squared	0.09	0.08	0.06	0.08	0.08
Mean dep. var.	0.30	0.25	0.11	0.17	0.18

Notes: The table presents results from regressions of respondents' socio-demographic characteristics on an indicator for whether the respondent is surveyed in the week following their House representative's toxic spike. All regressions include district, day and day of the week fixed effects and do not include any socio-demographic controls. Only respondents interviewed within a one week window of toxic spikes are included in the regressions. Standard errors are clustered at the electoral district level. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure B4: Placebo Test: Distribution of Coefficients from Random Spike Dates



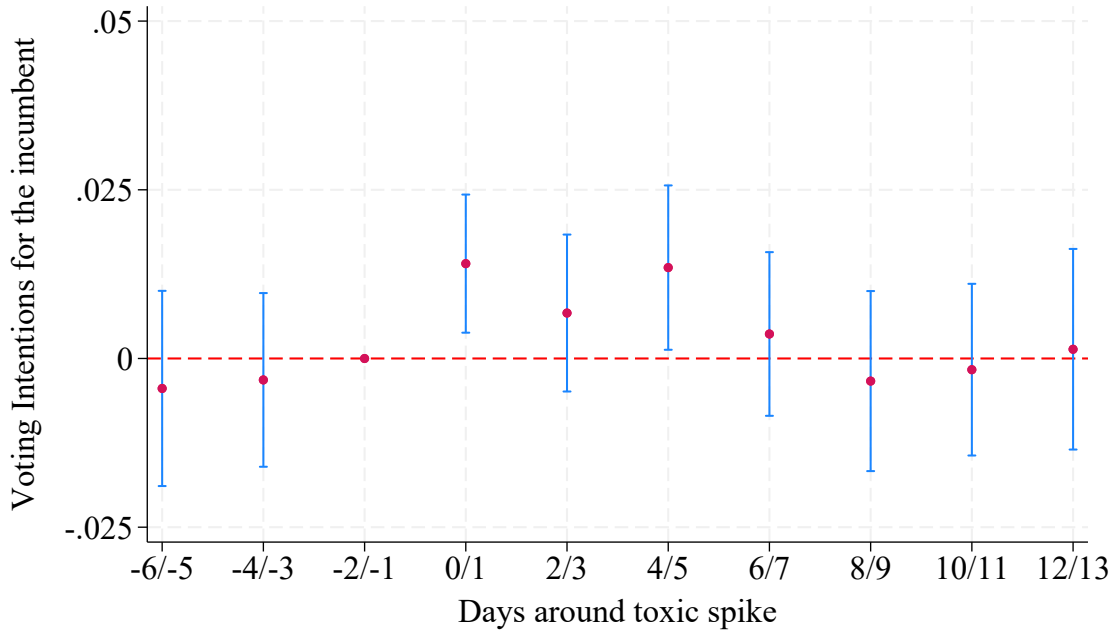
Notes: The figure displays the distribution of 1,000 placebo coefficients generated as follows. For each repetition, a set of placebo "spike" dates is randomly generated for each district, drawn from outside a four-week window of any real toxic spike to prevent contamination. The main specification is then re-estimated using these placebo dates. The vertical red line in each panel indicates the point estimate obtained using the true toxic spike dates. Each panel corresponds to a different binary outcome variable: (a) intending to vote for the incumbent versus any other option; (b) intending to vote for the incumbent versus being undecided; (c) intending to vote for the incumbent versus the opponent; and (d) intending to vote for the opponent versus being undecided. The placebo p-values associated to each test are: 0.002 (a), 0.001 (b), 0.164 (c) and 0.015 (d).

Table B6: Changes in voting intentions following incumbents' toxic spike - Party-by-day Fixed Effects

	(1) Vote Incumbent vs. Rest	(2) Vote Incumbent vs. Undecided	(3) Vote Incumbent vs. Opponent	(4) Vote Opponent vs. Undecided
Post Toxic Spike=1	0.026***	0.041***	0.008	0.036**
	(0.009)	(0.011)	(0.009)	(0.016)
Respondent controls	Yes	Yes	Yes	Yes
District FEs	Yes	Yes	Yes	Yes
Party × day FEs	Yes	Yes	Yes	Yes
Day of week FEs	Yes	Yes	Yes	Yes
Observations	6,101	4,199	5,059	2,900
Number of districts	197	197	197	194
R squared	0.67	0.53	0.78	0.59
Mean dep. var.	0.52	0.76	0.63	0.65

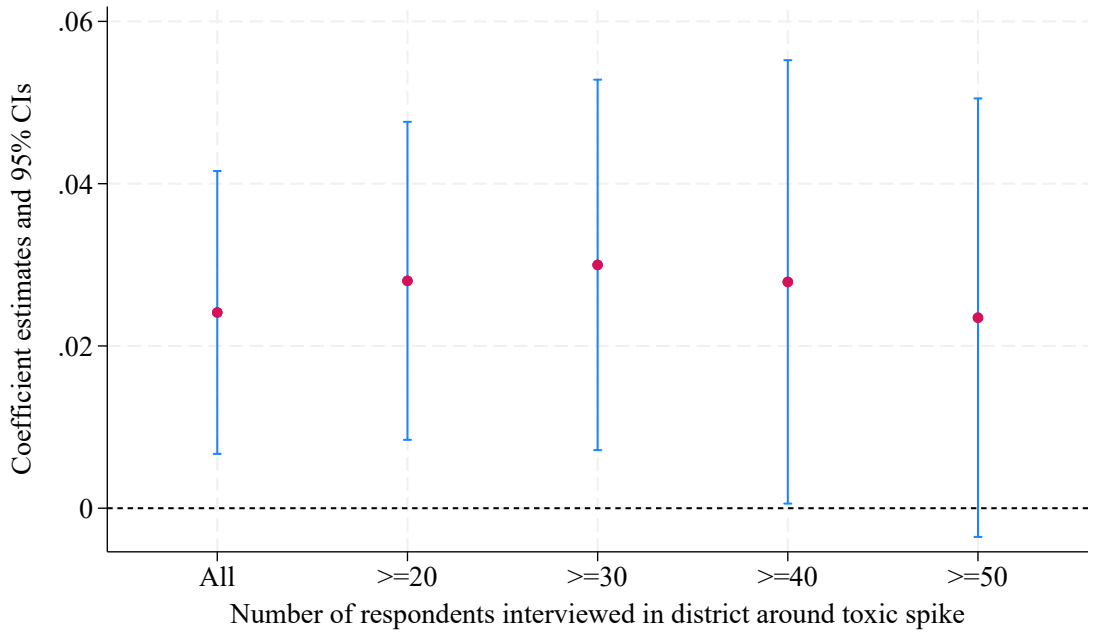
Notes: The table presents results from regressions of respondents' voting intentions on an indicator for whether the respondent is surveyed in the week following their House representative's toxic spike. In column (1), the dependent variable is an indicator variable equal to 1 if the respondent intends to vote for their House incumbent, and equal to 0 otherwise (i.e. if the respondent intends to vote for another candidate or they are undecided). The dependent variable in columns (2) and (3) are similarly coded but are equal to 0 if the respondent is undecided, resp. intends to vote for the incumbents' opponent. In column (4), the dependent variable is coded as 1 if the respondent intends to vote for the opponent, and equal to 0 if they are undecided. All regressions include respondents' 5-point ideological alignment with the incumbent, socio-demographic controls, district, party-by-day and day of the week fixed effects. Standard errors are clustered at the electoral district level. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figure B5: Event study of voting intentions around toxic spikes – Alternative toxic spike definition



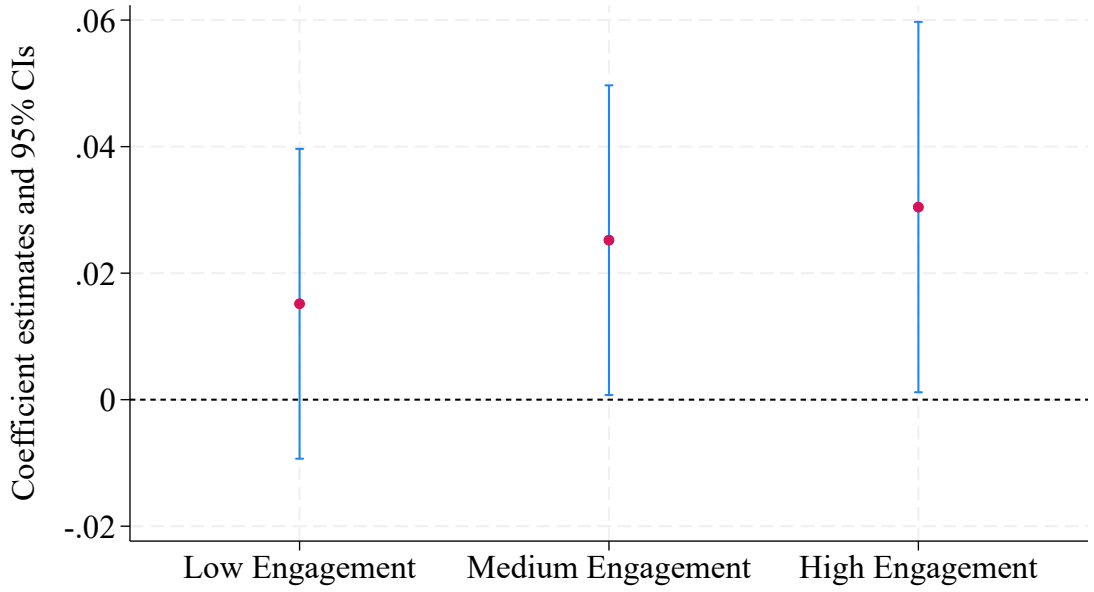
Notes: The figure displays coefficients and confidence intervals of a regression of respondents intent to vote for their House representative in the 2020 elections on leads and lags around spikes in the toxic content produced by their representative. A toxic spike is defined as a day where a politician's share of tweets classified as toxic by the custom LLM exceeds the 99th percentile of their own historical daily distribution of this share. Each lead and lag pools together two days to ensure precise estimation. The two days prior to the toxic spike are used as the reference time unit. Regressions include respondents' 5-point ideological alignment with the incumbent, socio-demographic controls, district, day and day of the week fixed effects. Standard errors are clustered at the electoral district level.

Figure B6: Changes in voting intentions following incumbents' toxic spike by number of respondents in district



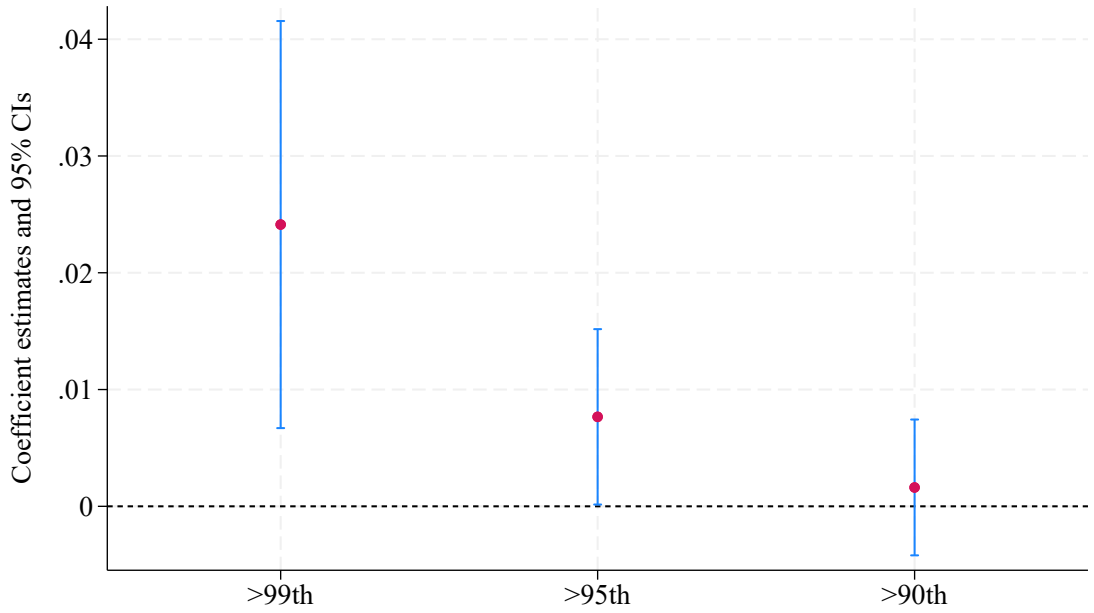
Notes: The figure displays coefficients and 95% confidence intervals of regressions of respondents' voting intentions on an indicator for being surveyed in the week following their representative's toxic spike. Each estimate is obtained from running the Equation 3 on different subsamples. The first point ("All") represents the baseline coefficient using the full estimation sample. Subsequent points progressively restrict the sample of respondents to those living in districts with at least 20, 30, 40, and 50 respondents surveyed within the two-week window around a toxic spike. A toxic spike is defined as a day where a politician's average Perspective API tweet toxicity score exceeds the 99th percentile of their own historical daily distribution. Regressions include respondents' ideological alignment with the incumbent, socio-demographic controls, district, day and day of the week fixed effects. Standard errors are clustered at the electoral district level.

Figure B7: Effect of toxic spike by engagement



Notes: The figure displays coefficients and 95% confidence intervals of regressions of respondents' voting intentions on an indicator for being surveyed in the week following a toxic spike interacted with a measure of a toxic spike's online engagement. Spikes are defined as days in which the average PAI toxicity score of a politician exceeds the 99th percentile of their historical daily distribution. The measure of toxic spike engagement is the difference between the number of likes generated by the most toxic tweet in a toxic spike and the politician's average number of likes during the first six months of the congressional session, binned into tertiles. Regressions include respondents' ideological alignment with the incumbent, socio-demographic controls, district, day and day of the week fixed effects. Standard errors are clustered at the electoral district level.

Figure B8: Effect of toxic spike by intensity



Notes: The figure displays coefficients and 95% confidence intervals of regressions of respondents' voting intentions on an indicator for whether the respondent is surveyed in the week following their House representative's toxic spikes. Spikes are defined according to various intensities of toxicity. They are identified as days in which the average PAI toxicity score of a politician exceeds the 99th percentile of their historical daily distribution, the 95th percentile or the 90th percentile. Regressions include respondents' ideological alignment with the incumbent, socio-demographic controls, district, day and day of the week fixed effects. Standard errors are clustered at the electoral district level.

C Textual Feature Extraction

This appendix provides detailed information on the methods used to extract the textual features employed in this paper. This includes details concerning: (i) the LLM annotation pipeline and validation; (ii) the use of more conventional natural language processing methods to extract tweet topic, sentiment and lexical diversity.

C.1 LLM Annotation Pipeline

C.1.1 Objectives

The primary toxicity measure used in this paper, the Perspective API (PAI) score, offers several advantages, including its granularity as a continuous measure of intensity and its status as a widely used benchmark in both industry and academic research (Müller and Schwarz, 2023a; Jiménez Durán et al., 2024; Beknazar-Yuzbashev et al., 2025; Kalra, 2025). However, its general-purpose nature presents several measurement challenges in the specific context of political communication. The PAI definition of toxicity capturing a "rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion" is broad and may be imperfectly suited to identify instances of *political incivility*. Another concern is that the measure may conflate incivility with related but distinct linguistic features, such as strong but civil criticism, negative sentiment, or specific emotions like anger and disgust – all of which are known to influence online engagement independently (Brady et al., 2017; Rathje et al., 2021; Frimer et al., 2023; Algan et al., 2025).

To address these challenges, I develop an custom tweet annotation pipeline relying on state-of-the-art Large Language Models (LLMs). This pipeline serves three main objectives. First, it allows to validate this paper's main findings by using a second, context-aware measure of toxicity tailored to political discourse. Second, it enables to create a rich set of control variables that allows to more cleanly isolate relevant variation in toxicity as well as benchmarking the role played by toxicity in driving online and offline outcomes against other key components of political communication. Third, it allows to conduct novel heterogeneity analyses based on the target of toxic content produced by politicians.

Specifically, this LLM annotation pipeline performs seven tasks: (i) identifying the presence of political toxicity; (ii) detecting the expression of criticism or disagreement in the tweets' content; (iii) discriminating between strong instances of criticism or disagreement and more mild ones ; (iv) identifying, summarizing and categorizing the targets of criticism in tweets expressing this feature ; (v) marking the use of sarcasm, irony or humor; (vi) classifying the rhetorical appeal used in tweets (cognitive or emotive), and (vii) the specific emotion expressed in the tweet if it is classified as appealing to emotions.

Due to resource constraints, the detailed LLM annotations were performed on a specific,

politically significant subset of the data: the universe of tweets from the 116th Congress' House Representatives running for reelection in November 2020 ($N = 753,806$). In addition to describing the universe of tweets over a two-year period, another advantage of gathering LLM annotations on this sample is that they allow to identify spikes in politicians' toxic rhetoric using an alternative measure when studying offline political returns. This analysis is conducted using a large-scale electoral survey fielded during the 116th Congress.

As shown in Table C1, this sampling strategy induces some statistically significant differences between the annotated and non-annotated samples. Most notably, the annotated sample exhibits lower average engagement. This is due to the sampled tweets being posted by House Representatives who enjoy a lower social media profile than Senators. In addition, following the doubling of the tweet length limit in November 2017, tweets in the annotated sample are longer and more lexically diverse than those in the non-annotated sample – which comprises tweets written between June and November 2017. Critically, however, the average PAI toxicity score is statistically indistinguishable between the two samples, and the differences in other stylistic features of communication (sentiment, emotional intensity and irony) are qualitatively negligible. If anything, these differences suggest that the estimates derived from this sample in section 3 likely represent a conservative measure of the engagement returns to toxicity across the full study period, where the average engagement levels are higher but communication features similar.

C.1.2 LLM Validation Procedure

The reliability of the LLM-based annotations was established through a rigorous validation procedure, consistent with best practices for the use of LLMs in social science research (Ludwig et al., 2025). The procedure was designed to select the best-performing model from a set of high-quality candidates by benchmarking their performance against a human-coded gold standard. It follows several steps.

First, a ground truth dataset was created by drawing a random sample of 300 tweets from the main dataset. To ensure sufficient variation for evaluating the classification of rare features such as toxicity, this sample was stratified. Specifically, the sample was composed of one-third of tweets with high PAI scores (above the 90th percentile of the toxicity distribution, i.e. 0.15) and two-thirds with scores below this threshold. This dataset was then independently annotated by two trained research assistants based on a detailed codebook for all seven textual features of interest. Because comparing predictive models to individual, sometimes disagreeing, human annotations can be ambiguous, I further restrict the set of tweets used for the main validation exercise to those where the two human annotators independently provided the exact same classification. This process isolates a set of unambiguous instances that provide a stable and reliable gold standard, allowing for a cleaner assessment

Table C1: Balance in tweet-level characteristics

	(1) Not LLM-annotated		(2) LLM-annotated		(2)-(1)
Variable	N	Mean	N	Mean	Difference
PAI Toxicity score	2213069	0.05	745947	0.05	-0.00
		(0.00)		(0.00)	
Likes	2166273	909.74	708328	771.79	-137.95***
		(4.85)		(8.03)	
Replies	2166273	102.19	708328	73.11	-29.07***
		(0.53)		(0.80)	
Retweets	2166273	197.09	708328	180.54	-16.54***
		(0.90)		(1.69)	
Quotes	2166273	19.22	708328	15.62	-3.60***
		(0.15)		(0.18)	
Total Engagement	2166273	1228.23	708328	1041.07	-187.16***
		(6.04)		(10.28)	
Sentiment	2238254	0.25	753806	0.24	-0.00***
		(0.00)		(0.00)	
Emotional intensity	2238254	0.65	753806	0.65	0.00*
		(0.00)		(0.00)	
Irony	2238254	0.42	753806	0.40	-0.02***
		(0.00)		(0.00)	
Lexical diversity	2238254	76.70	753806	80.48	3.78***
		(0.05)		(0.08)	
Number of words	2238254	30.10	753806	31.51	1.41***
		(0.01)		(0.01)	
Campaign account	2234042	0.28	753806	0.27	-0.01***
		(0.00)		(0.00)	

Notes: The table compares the means of tweet-level characteristics between the sample of tweets used for LLM annotation (Column 2) and the remaining tweets in the full dataset (Column 1). The LLM-annotated sample consists of all tweets from House Representatives running for reelection during the 116th Congress (January 2019 - November 2020). Column (3) reports the difference in means between the two groups. Robust standard errors are reported in parentheses. The significance of the difference is calculated from a simple OLS regression of each characteristic on an indicator for the tweet being in the LLM-annotated sample. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

of model performance.

Against this benchmark, I evaluated a set of five high-performing and widely-used LLMs. This set is composed of four proprietary models (Gemini 2.0 Flash, Gemini 2.0 Flash Lite,

GPT-4.1-mini, GPT-4o-mini) and one open-source model (Deepseekv3). I traded-off several criteria upon selecting these models: (i) performance on text-related tasks as reported on the Language Model Arena leaderboard, (ii) affordability and parsimony, (iii) speed of annotation; and (iv) diversity of providers in order to reflect the rapidly evolving landscape of generative AI models (Korinek, 2023).[26] Such diversity is crucial as it ensures that annotation performance is optimized over different model makes, instead of tying one's hands to one given model from the onset without further testing.

The primary metric used for this evaluation is the F1-Macro score. The F1-Macro calculates the F1-score for each category independently and then computes their unweighted average over all categories in the distribution.[27] This metric is chosen over simpler alternatives like accuracy for its robustness in handling imbalanced distributions such as toxicity or sarcasm, which are common in our data (Dell, 2025). In addition, it is particularly well-suited for the prediction of categorical variables as it evaluates performance across all categories, rather than focusing only on a single "positive" class. I adopt the F1-Macro score for all seven annotation tasks to maintain a consistent evaluation framework.[28]

Figure C2 displays the F1-Macro score for each of the five LLMs, as well as for a random classifier baseline that predicts labels according to their empirical distribution in the restricted validation set. Across the seven annotation tasks, Deepseekv3 achieves the highest average F1-Macro score overall (0.849), outperforming the other LLMs (Gemini 2.0 Flash: 0.845; GPT-4o-mini: 0.818; GPT-4.1-mini: 0.786; Gemini 2.0 Flash Lite: 0.770) and substantially exceeding the random baseline (0.419). It also achieves the highest F1-Macro score on three out of seven individual annotation tasks, namely detecting the target category, classifying criticism/disagreement, and identifying sarcasm/irony.

Having selected Deepseekv3 as the best-performing model, I provide a more detailed assessment of its performance to further build confidence in the quality of its annotations.

[26] While state-of-the-art reasoning models (e.g. Gemini 2.5 Pro, GPT-o3, Deepseek-R1) achieve the highest performances on text related tasks, their expected gains on relatively simple text classification tasks are too low with respect to the surge in cost they entail over more standard non-reasoning models. Standard models only suffer from a 1.2% to 7.5% performance loss on text-related tasks compared to their reasoning counterparts but are priced 4 to 17 times cheaper as of July 30, 2025. Refer to the pricing information for Gemini, GPT, and Deepseek.

[27] In binary classification tasks, the F1 score is a commonly used metric. It is the harmonic mean of precision and recall. Precision measures the share of predicted positive cases that are truly positive (a low precision implies a high rate of Type I errors, or false positives). Recall measures the share of true positive cases that are correctly identified (a low recall implies a high rate of Type II errors, or false negatives). The F1-score provides a single metric that balances this trade-off by penalizing models with low precision or low recall.

[28] The results of the model selection exercise are qualitatively identical when using a combination of the standard F1-score for binary tasks and the F1-Weighted score for multi-class tasks.

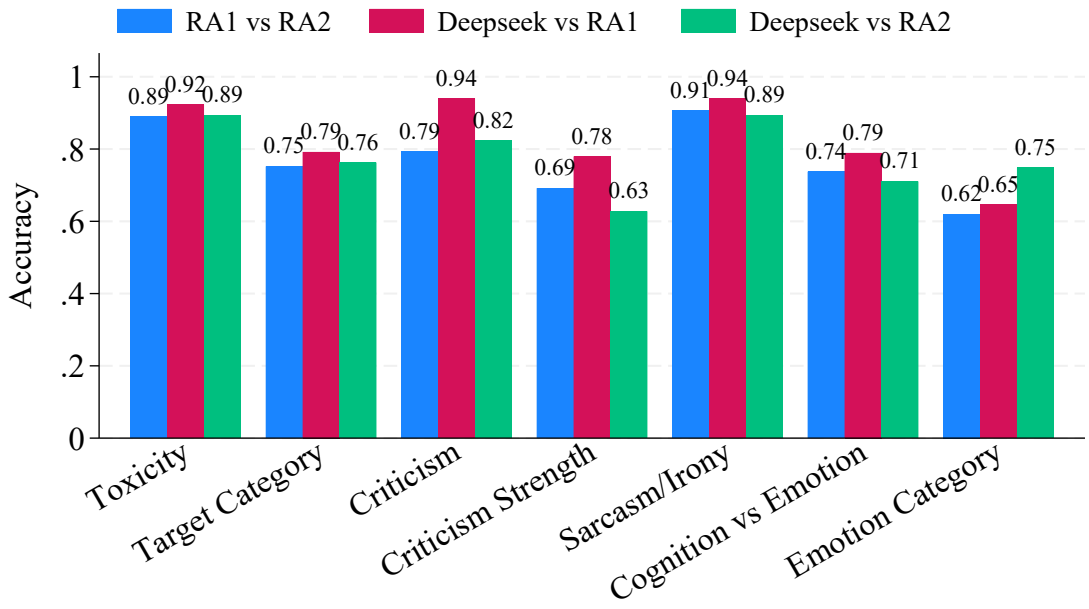
Table C2: LLM Annotation Performance

LLM	Avg.	Toxicity	Targ. Cat.	Crit.	Crit. Strength	Sarcasm	Rhet. Appeal	Emo. Cat.
Deepseekv3	**0.849**	0.867	**0.866**	**0.979**	0.785	**0.832**	0.827	0.788
Gemini 2.0 Flash	0.845	**0.898**	0.727	0.970	0.857	0.812	0.801	**0.850**
GPT-4o mini	0.818	0.777	0.683	0.958	**0.906**	0.788	0.775	0.842
GPT-4.1 mini	0.786	0.698	0.767	0.937	0.813	0.727	**0.829**	0.733
Gemini 2.0 Flash Lite	0.770	0.717	0.534	0.962	0.862	0.793	0.810	0.712
Random Classifier	0.419	0.466	0.323	0.501	0.563	0.483	0.468	0.126

Notes: The table displays the F1-Macro score for each of the five LLM candidates and a random classifier baseline that predicts labels according to their empirical distribution in the validation set. Performance is measured on the subset of tweets where two human annotators agreed. The seven annotation tasks are: toxicity presence, target category, presence of criticism/disagreement, strength of criticism/disagreement, use of sacrasm/irony/humor, rhetorical appeal (cognitive or emotive), and emotion category. The simple unweighted average of F1-Macro scores over all seven tasks is taken for column "Avg.".

Figure C1 presents the accuracy score, or raw agreement rate, calculated on the full validation set for the following pairs: (i) the agreement between the two human annotators (RA1 vs RA2), which serves as the human benchmark; (ii) the agreement between Deepseek and the first human annotator (Deepseek vs RA1); and (iii) the agreement between Deepseek and the second human annotator (Deepseek vs RA2). The figure yields two main insights. First, for every task, the agreement rate between Deepseek and a human annotator is comparable to, and often exceeds, the agreement rate between the two human annotators themselves. For instance, when identifying the presence of political toxicity, the human-human agreement is 89%, while Deepseek's agreement with the humans is 92% and 89%. This suggests that the LLM's annotations achieve a level of consistency and quality that is on par with that of a trained human coder. Second, the results illustrate the inherent subjectivity of certain tasks. For instance, for the classification of sarcasm/irony where inter-human agreement is highest (91%), Deepseek's agreement with humans is also exceptionally high (94% and 89%). Conversely, for more subjective tasks like identifying a specific emotion category where human agreement is lower (62%), Deepseek's performance reflects this difficulty (65% and 75% agreement rates with the two human-coders). This demonstrates that the model's capabilities track human cognitive limits in a predictable and reasonable way, and reinforce the claim that this LLM mimics human judgment on these textual annotation tasks.

Figure C1: Detailed assessment of the best-performing LLM (Deepseekv3)



Notes: The figure displays the accuracy score of the best-performing LLM against each of the two human annotators and compares this to the share of agreement between the two human coders ("RA1 vs RA2"). Scores are computed on the full 300-tweet validation sample.

C.1.3 LLM Annotation Prompts

For transparency and reproducibility, the exact prompts provided to the LLM for each annotation task are detailed below. Each prompt was engineered to be concise and to provide clear, mutually exclusive definitions to guide the model's classification. The prompts for classifying the type of rhetorical appeal used in the tweet (cognitive vs emotional) and the specific emotion used if the tweet is classified as appealing to emotions both build on the prompts used in Algan et al. (2025).

Presence of toxicity

```
You are an expert communication analyst specialized in identifying toxicity in
tweets. Your task is to classify a tweet as: "0" (non-toxic), "1" (toxic), or
"unsure" (ambiguous/unclear). Respond ONLY with the number/word. No explanations.
TOXIC: disrespectful or aggressive language breaking norms of civil political
debate. Includes: insults, degrading comments (to individuals, groups,
institutions, ideas, policies), vilification, hostile tone (demeaning
profanity, extreme sarcasm).
NOT TOXIC: Routine social media communication by politicians. Also, respectful
(even strong) criticism or disagreement, profanity not used to demean, quoting
offensive content without endorsement or escalation.
```

Presence of criticism or disagreement

```
You are an expert communication analyst specialized in identifying disagreement,
criticism, or provocation in tweets. Respond ONLY with "1" (presence), "0"
(absence), or "unsure" (ambiguous/unclear). No explanations.
Mark "1" if the tweet challenges, questions, or expresses an opposing
viewpoint towards: individuals, groups (e.g., social, religious, racial),
parties, institutions, specific policies, issues, or ideologies.
This expression may be aggressive but need NOT to. Polite disagreement,
reasoned criticism, or courteous provocations also count.
Mark "0" for purely informational, supportive, or general neutral statements.
```

Strength of criticism or disagreement

```
You are an expert communication analyst specialized in determining the strength
of criticism, disagreement or provocation expressed in tweets that have already
been identified as expressing such features. Respond ONLY with "1" (Standard),
"2" (Strong), or "NA" (Not Applicable/No Criticism).
```

No explanations.
DEFINITIONS:
"1" (Standard): Clear, direct, and unambiguous criticism or opposition.
The tone is firm but lacks overwhelming intensity.
"2" (Strong): More acute and forceful criticism or opposition. Usually carries pronounced linguistic emphasis (loaded words, amplifying phrases), often expressed aggressively.
Guidance: If a tweet is clearly critical but you are unsure if it is intense enough to be "Strong", lean towards "Standard" ("1").

Target of criticism or disagreement

You are an expert communication analyst specialized in determining the target of tweets. Your task is to annotate the target, if any, of a tweet that has already been identified as expressing criticism/disagreement/provocation.
Respond ONLY in this JSON-like format: {
"target_summary": "Y",
"target_category": "Z"
}
NO explanations. Ensure the output is a single, valid JSON object containing EXACTLY 2 fields.
DEFINITIONS:
target_summary : Summarize the specific target in MAX 5 WORDS
(e.g., "President Biden," "The Green New Deal," "Far-right media").
Use "unsure" if difficult to summarize.
target_category : Select ONLY one category number for the identified target.
CATEGORY DEFINITIONS:
"1" = Individual (public figure, Twitter user, ordinary person).
"2" = Collective Political Entity (party, political group, media organization).
"3" = Non-Purely Political Entity/Minority (social, racial, religious, other).
"4" = Issue/Policy/Ideology (e.g., legislation, societal problem, concept).
Use "unsure" if ambiguous or unclear.

Presence of sarcasm, irony or humor

You are an expert communication analyst. Your task is to identify irony, sarcasm, or humor in tweets written by US Members of Congress. Humor includes lighthearted, witty, or subtle tongue-in-cheek remarks.
Respond ONLY with "1" (presence), "0" (absence), or "unsure"

(ambiguous/unclear). No explanations.
CONTEXTUAL NOTE: These tweets are from US politicians. Their communication can be direct and use strong language for emphasis, even when meant literally. Interpret potential sarcasm/irony cautiously, considering this professional and often strategic context. If a statement could be sarcastic in casual conversation but is a plausible (even if strong) literal assertion for a politician, lean "0" unless non-literal intent is very clear.

Rhetorical appeal

You are an expert communication analyst. Classify the tweet's PRIMARY rhetorical appeal: "1" (cognitive), "2" (emotive), "unsure" (ambiguous/unclear.)
Respond ONLY with the number/word. No explanations.
DEFINITIONS:
Cognitive ("1"): Appeals to logic, facts, reason, evidence, or practicalities. Aims to inform or persuade rationally. Includes purely factual/neutral statements presenting information.
Emotive ("2"): Appeals to feelings, sentiments, or triggers emotional responses. Aims to persuade by evoking emotion.
Guidance: Identify the most DOMINANT appeal.

Emotion category

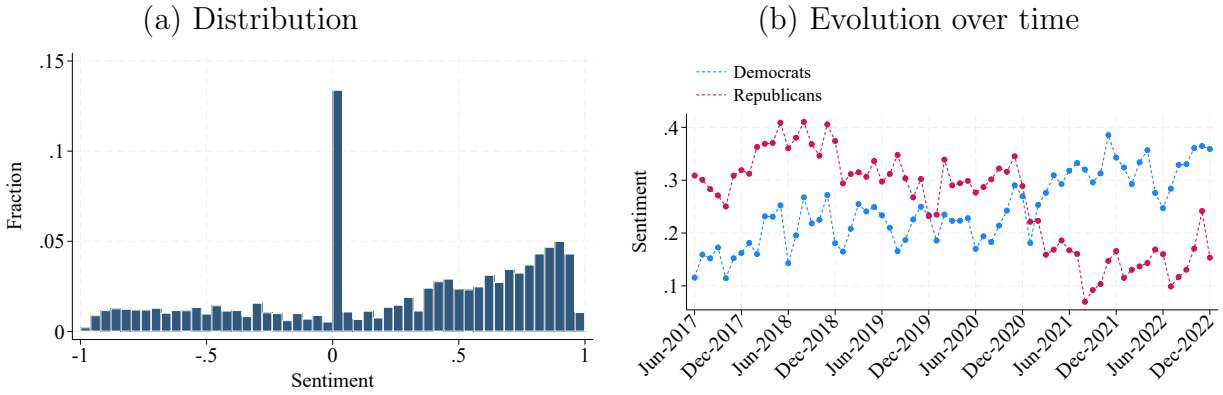
You are an expert communication analyst tasked with determining the emotion expressed in tweets already identified as primarily appealing to emotions. Respond ONLY with ONE of these options: "sadness", "fear", "anger", "disgust", "joy", "hope" (includes motivation to pursue goals), "gratitude", "pride", or "unsure". No explanations.

C.2 Conventionally extracted textual features

A variety of other tweet-level characteristics could make the positive association between toxicity and engagement spurious if left unobserved, e.g. tweet sentiment, topic, lexical diversity. This subsection provides additional details about how these features were extracted from the full text of tweets using more conventional natural language processing techniques.

Sentiment analysis: Tweet sentiment is determined using VADER (Valence Aware Dictionary and sEntiment Reasoner) which is a lexicon and rule-based tool commonly used for sentiment analysis on short text, and specifically for text expressed in social media (Hutto and Gilbert, 2014). VADER provides a continuous sentiment score that ranges from -1 (very negative) to +1 (very positive), reflecting the overall sentiment of a tweet. The sentiment score is derived through a lexicon-based approach, where each word in the text is assigned a basic sentiment score. The final sentiment, known as the compound score, is calculated by summing the valence scores of each word in the lexicon, adjusted for rules that consider grammatical and syntactical cues like punctuation, capitalization, and modifier words, which may alter the intensity or polarity of sentiment. This compound score is then normalized to range between -1 and +1. Figure C2 plots the distribution and montly evolution by party of this sentiment score for all the tweets in the dataset.

Figure C2: Distribution and evolution of tweet sentiment



Notes: Panel (a) plots the distribution of tweet sentiment as computed using the VADER sentiment analysis tool. A score of 1 indicates extremely positive sentiment, a score of -1 indicates an extremely negative sentiment. Panel (b) plots the monthly evolution of the average sentiment of tweets posted by Democrat congress members (in blue) and Republicans (in red).

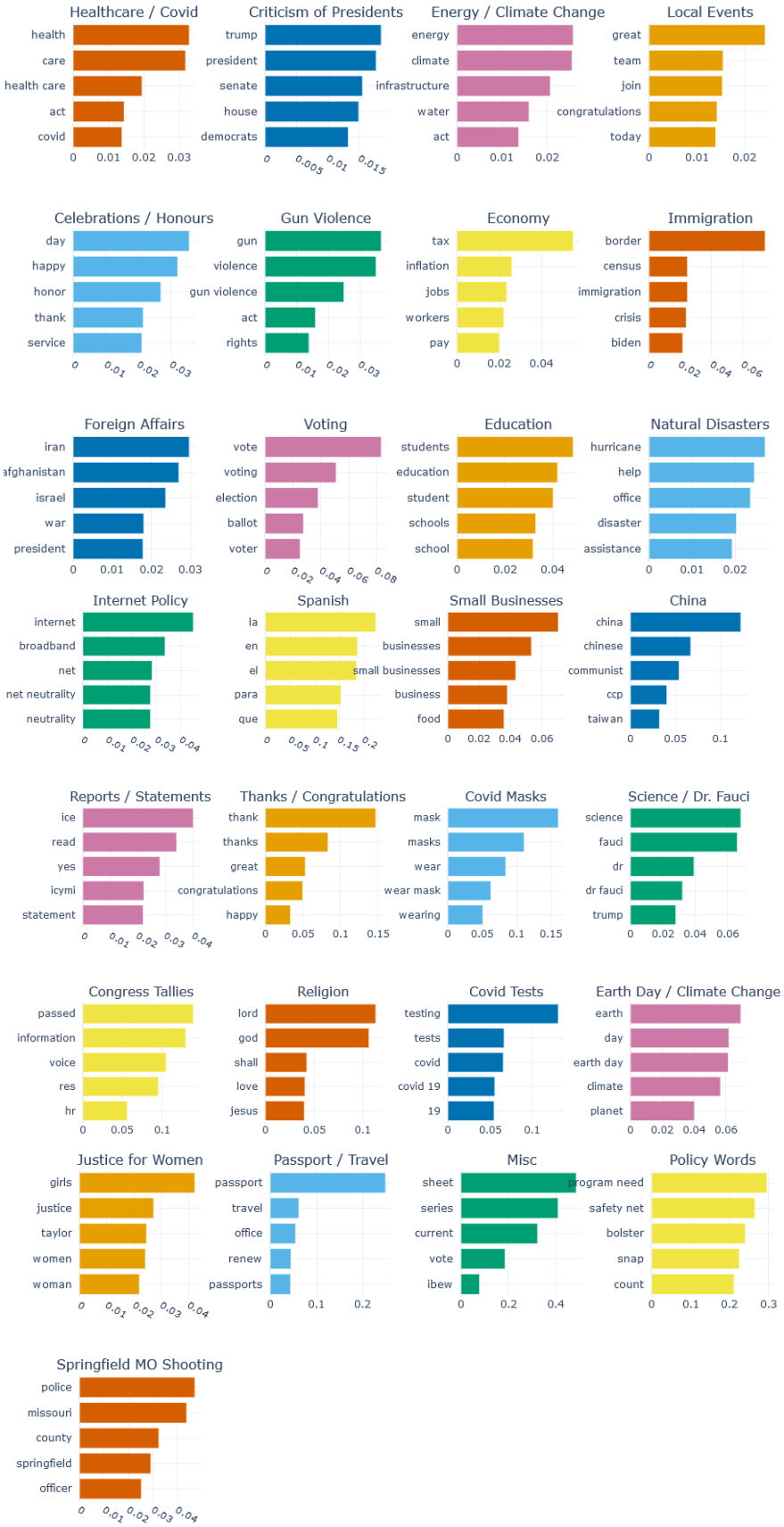
Topic detection: Topic detection was carried out using BERTopic, a state-of-the-art topic modeling technique that leverages transformer-based embeddings to infer topics from large text corpora in an unsupervised way (Grootendorst, 2022). BERTopic operates by first transforming tweets into high-dimensional numerical representations or vectors (known as

"embeddings"). These tweet embeddings are then reduced in dimensionality before being clustered into groups.[29] Tweets belonging to the same group or cluster form a topic. Importantly, relying on an embedding-based approach for topic modeling is particularly suitable for short text like tweets. Indeed, other popular topic modeling approaches such as LDA perform poorly on this type of data as they need to be trained on long text documents to learn their underlying topic representation.

An important parameter the researcher controls is the number of topics to be produced through BERTopic. This was set to 30 topics. The labeling of each group is then manually determined by examining the words and the tweets that are most representative of the group. Figure C3 gives an overview of the top 5 words associated to each topic, along with the ex-post assigned topic label. In addition, Figure C4 plots the topic distribution of the tweets. This allows to observe that some of most common topics expressed in politicians' tweets are "Criticize presidents", "Celebration/Honors", "Voting", "Health/Covid" and "Economy".

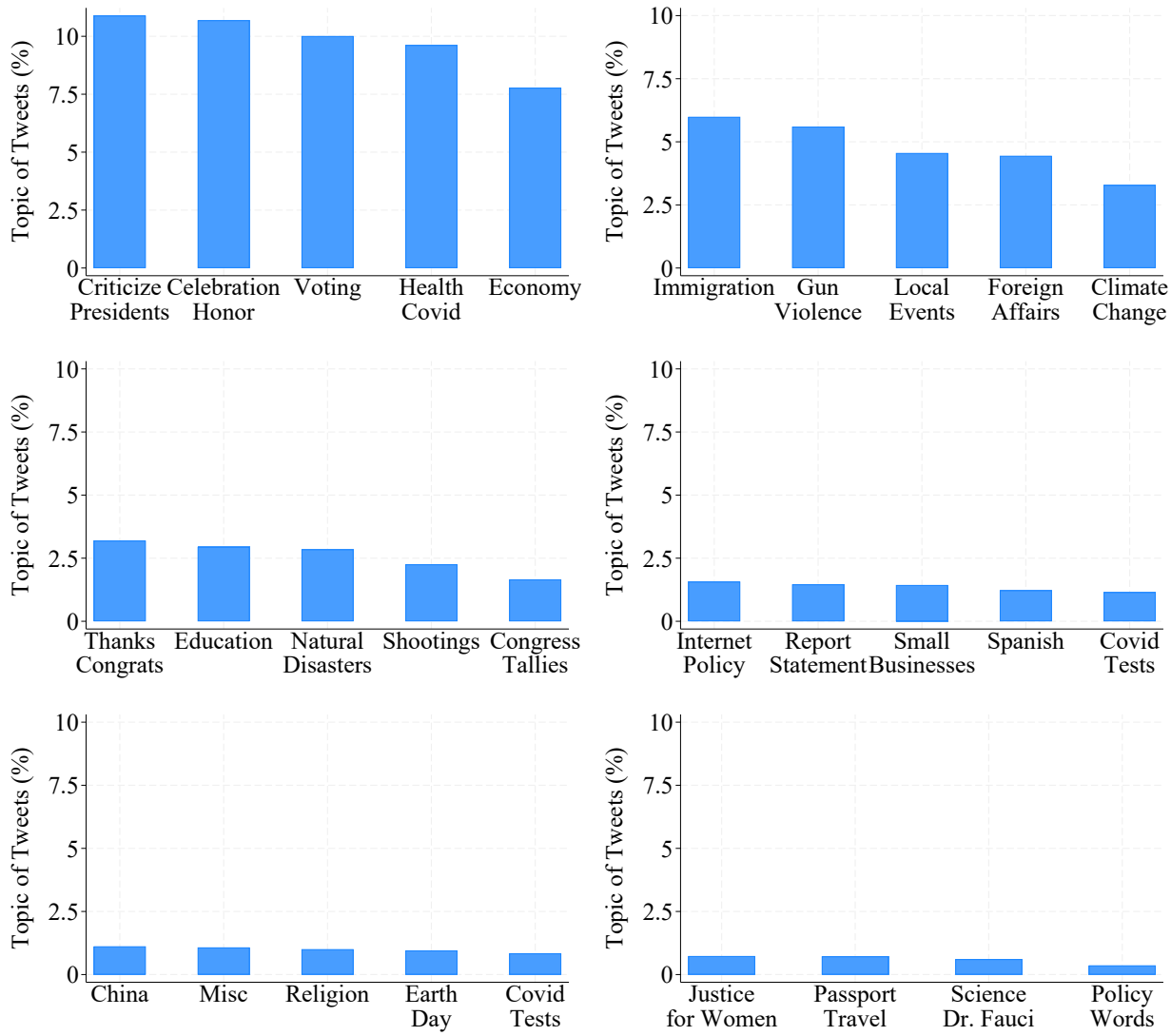
[29]The UMAP algorithm (Uniform Manifold Approximation and Projection) was used to reduce the dimensionality of embeddings and clustering was performed using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise).

Figure C3: Top 5 words associated to each topic (BERTopic)



Notes: This figure plots the top 5 words associated to each cluster of tweets (topic) as determined by BERTopic (Grootendorst, 2022). Model training was specified to produce 30 distinct clusters of tweets. Topic labeling is performed ex-post by the author by examining the top words and top documents representing each topic.

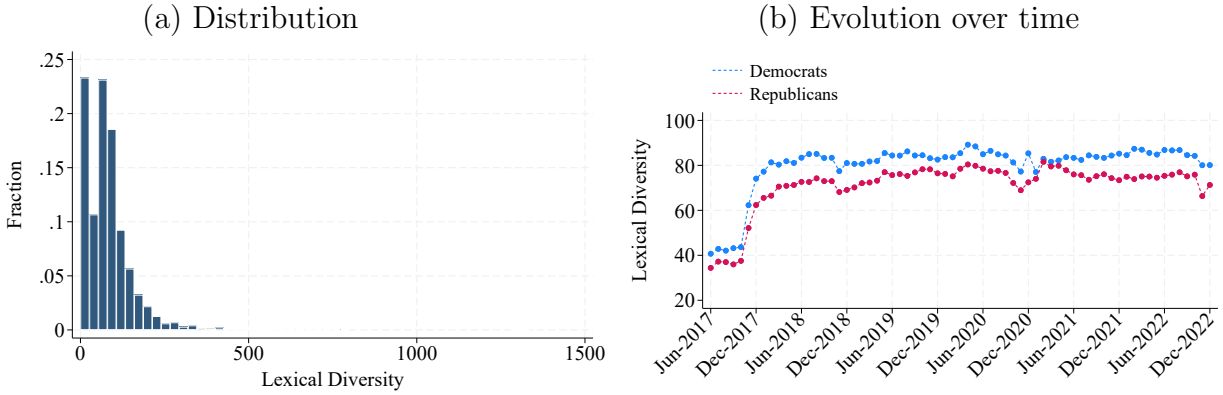
Figure C4: Tweet topic distribution



Notes: This figure presents the distribution of tweet topics in the sample of U.S. congress members' tweets (2017-2022). BERTopic is used to detect topics in an unsupervised way (Grootendorst, 2022). Model training was specified to produce 30 distinct clusters of tweets. Topic labeling is performed ex-post by the author by examining the top words and top documents representing each topic.

Lexical diversity: Lexical diversity is proxied using the Measure of Textual Lexical Diversity (MTLD). MTLD has been widely adopted in linguistic studies as a measure of text's lexical richness and complexity (McCarthy and Jarvis, 2010). It is a continuous score, where higher values indicate greater lexical diversity, suggesting a richer and more varied vocabulary usage within a given text. Figure C5 plots the distribution and monthly evolution by party of tweet irony score.

Figure C5: Distribution and evolution of tweet lexical diversity (MTLD)



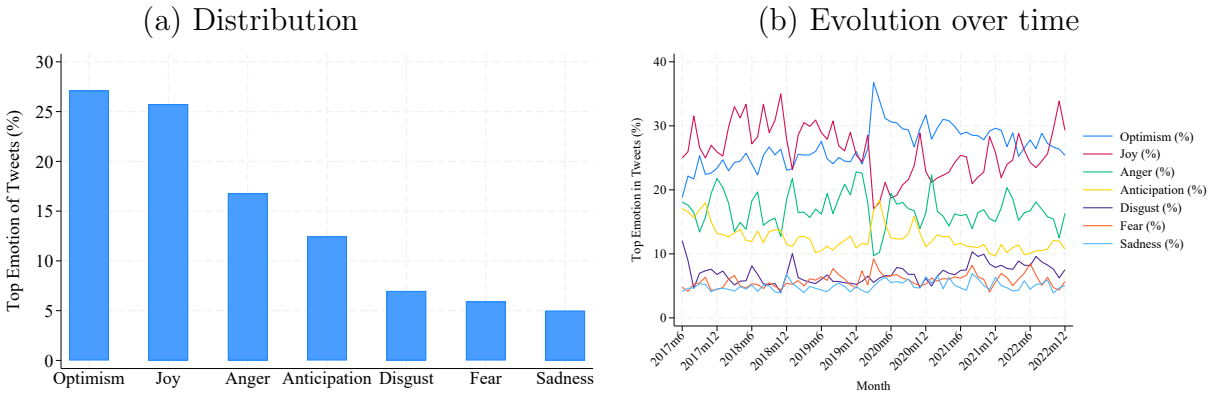
Notes: Panel (a) plots the distribution of tweets' lexical diversity as proxied by the Measure of Textual Lexical Diversity (McCarthy and Jarvis, 2010). Panel (b) plots the monthly evolution of the average lexical diversity of tweets posted by Democrat congress members (in blue) and Republicans (in red).

Emotion recognition: Extraction of continuous emotional scores was conducted using a pre-trained deep learning model from TweetNLP (Camacho-Collados et al., 2022). This procedure is specifically suited for analyzing emotions expressed in tweets as it relies on a version of the RoBERTa transformer model that is fine-tuned on a dataset comprising 60 million tweets, including 3,300 tweets that were manually annotated with emotional labels. As such, the procedure enables to predict the dominant emotion class in a tweet out of 11 emotion classes: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust. Figure C6 plots the distribution and monthly evolution of the top emotion class.

I build on these continuous emotion probabilities to construct a continuous measure of emotional intensity which is equal to the maximal probability across all eleven emotion categories. The continuous emotional intensity measure is used in section 4 to identify spikes in the emotional language used in politicians' tweets. This is done to test for whether changes in voting intentions following unusual political communication is specific to the use of toxicity or expands to other dimensions of political communication.

Irony detection: TweetNLP's irony detection model was used to determine an irony score

Figure C6: Distribution and evolution of tweet top emotion class (TweetNLP)

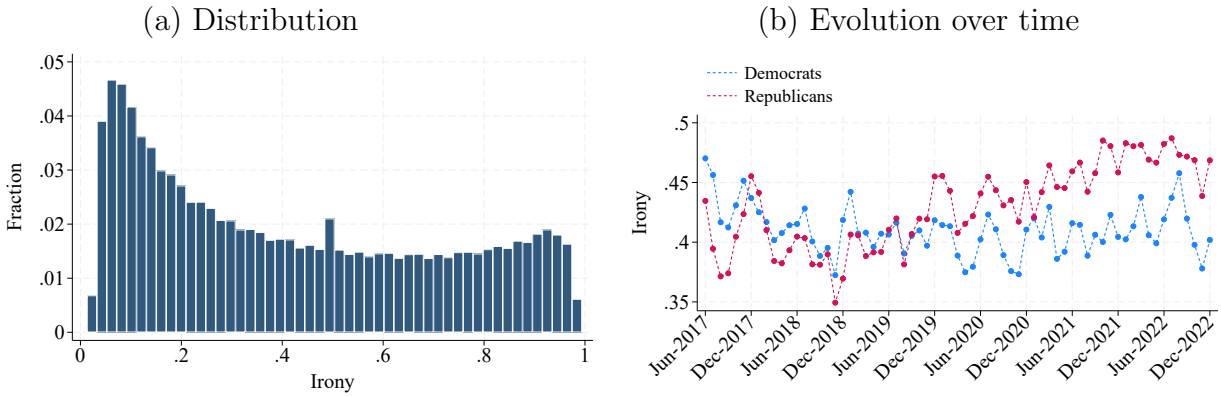


Notes: Panel (a) plots the distribution of tweets' top emotion as detected by TweetNLP's emotion recognition deep-learning model. Panel (b) plots the monthly evolution of tweets' top emotion class.

ranging from 0 to 1 which represents the probability that a tweet uses irony. The original model is a fine-tuned version of the RoBERTa base model on 60 million tweets, including 2,900 tweets that were manually labelled for expressing irony (binary label) as detailed in Camacho-Collados et al. (2022). Figure C7 plots the distribution and monthly evolution by party of tweet irony score.

This continuous measure of irony is used in section 4 to identify spikes in politicians' use of irony. This is done to test for whether changes in voting intentions following unusual political communication is specific to the use of toxicity or expands to other dimensions of political communication.

Figure C7: Distribution and evolution of tweet irony score (TweetNLP)

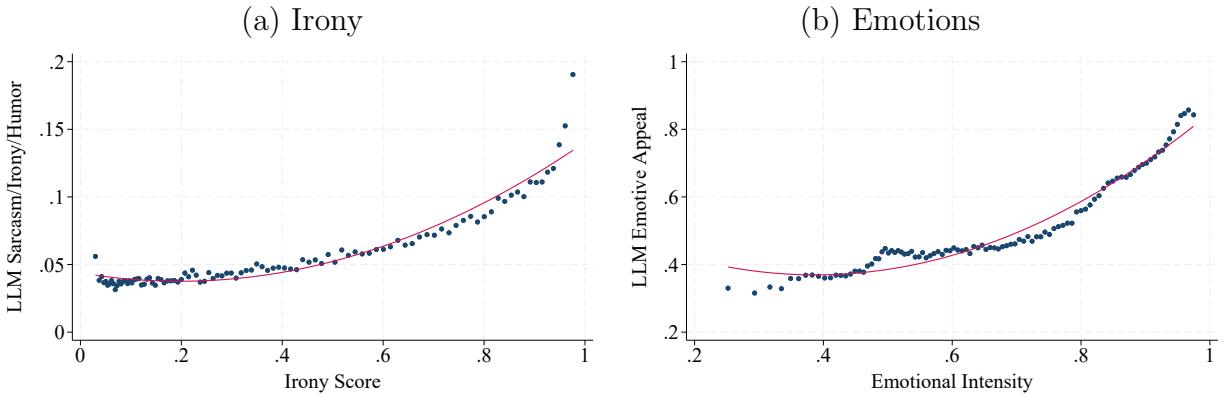


Notes: Panel (a) plots the distribution of tweets' irony score as predicted by TweetNLP's irony detection deep-learning model. Panel (b) plots the monthly evolution of the average irony score of tweets posted by Democrat congress members (in blue) and Republicans (in red).

Correlation between continuous and binary LLM-annotated tweet features: For the same reasons that the continuous PAI toxicity measure is preferred over the LLM-annotated measure of toxicity to identify spikes in politicians' rhetoric, I rely on continu-

ous measures of emotional intensity and irony to identify spikes in alternative dimensions of communication. While these two measures are not exactly identical to their LLM-annotated counterparts – respectively an indicator for emotional appeals, and the presence of sarcasm/irony/humor – I provide evidence that there is strong overlap within each pair of variables. Figure C8 shows that both continuous scores are positively correlated with the respective LLM-annotated indicator, building confidence that they proxy for the intensity of emotional appeals and the use of sarcasm/irony/humor.

Figure C8: Correlation between continuous and binary LLM-annotated tweet features



Notes: The figure presents two binscatter plots validating continuous scores of political communication dimensions against their LLM-annotated counterparts. Panel (a) plots the probability that a tweet is classified as containing sarcasm, irony, or humor by the LLM (Y-axis) against a continuous irony score predicted by a deep learning model (X-axis). Panel (b) plots the probability that a tweet is classified as having an emotive appeal by the LLM (Y-axis) against a continuous emotional intensity score (X-axis). Emotional intensity is defined as the maximal probability associated to eleven emotions predicted by a deep learning model (Camacho-Collados et al., 2022). Both plots are constructed over 100 equal-sized bins of their respective continuous scores. The solid red line in each panel shows the quadratic fit.

C.3 Choice of threshold to binarize PAI toxicity

To allow for a fair comparison between toxicity and other stylistic drivers of online engagement in section 3, I transform the continuous Perspective API (PAI) measure into an indicator variable, since most other textual features are measured as binary or categorical variables. This requires defining a threshold. Tweets with a continuous PAI toxicity score above that threshold will be marked as toxic, and non-toxic if the score is below the threshold.

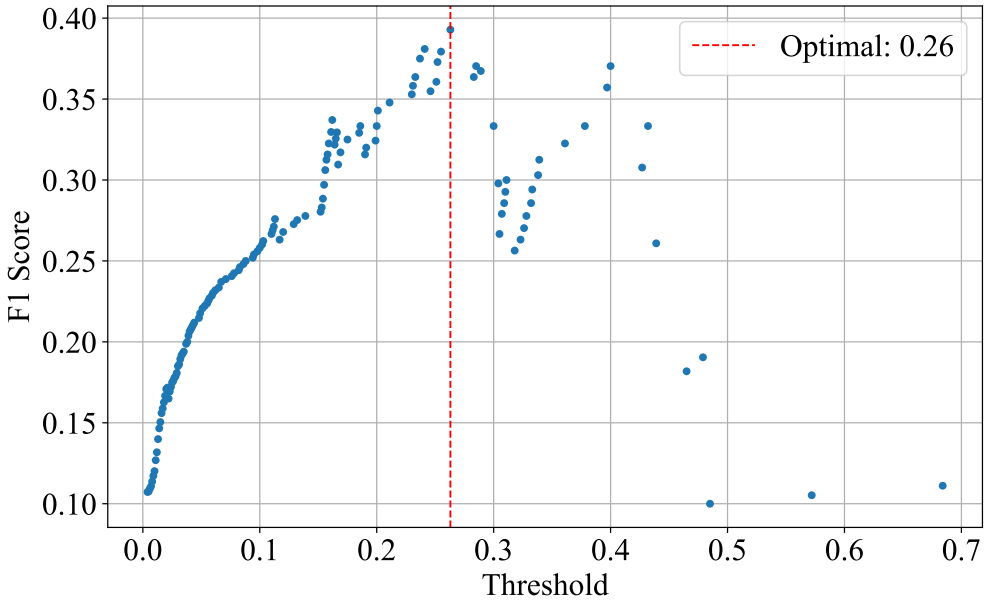
I select this threshold using the following data-driven procedure. First, a human annotator classifies the 300 tweets in the validation sample described in Appendix C.1, marking tweets as toxic according to PAI's definition of toxicity (i.e., "a rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion"). This ensures that the comparison between the continuous PAI score and human annotations are based on the same conceptual ground. Indeed, the idea underlying this procedure is to identify the threshold that best allows the PAI measure to replicate human judgments of toxicity.

Second, a grid search is performed over all unique continuous PAI score values in the validation sample. Each value is then used as a candidate threshold. Specifically, for each of these candidate values, I create an indicator variable and evaluate its performance against the human-annotated ground truth. The primary metric used for evaluation is the F1-score for the toxic class, a performance metric which trades off type 1 and type 2 errors in binary classification tasks. It is particularly well-suited in cases where there is substantial class imbalance in the data, as is the case of toxicity.[30] In such instances, simpler alternatives like accuracy (i.e. the raw rate of agreement between the predicted classifications and the ground-truth) can be misleading, as a model that naively classifies all tweets as non-toxic would achieve high accuracy but be practically useless. The F1-score is also preferred over the F1-Macro score as the latter is better-suited in multi-class (i.e. categorical) classification tasks.

Figure C9 plots the F1-score as a function of the candidate PAI threshold values. The score is maximized at a threshold of 0.26, marked by the dashed vertical line in red. I rely on this value to define the binary toxicity indicator in the main analysis of online engagement. Finally, note that the results exposed in section 3 are robust to threshold choice, as shown in Figure B2 and would still hold had the optimal threshold of 0.40 been selected under an alternative evaluation metric (F1-Macro score).

[30] As noted in Appendix C.1, the F1 score is the harmonic mean between precision and recall. Precision measures the share of predicted positive cases that are truly positive (a low precision implies a high rate of Type I errors, or false positives). Recall measures the share of true positive cases that are correctly identified (a low recall implies a high rate of Type II errors, or false negatives). The F1-score provides a single metric that balances this trade-off by penalizing models with low precision *or* low recall.

Figure C9: Optimal Threshold Selection for PAI Toxicity Indicator



Notes: The figure displays the performance of a simple threshold classifier in replicating human judgments of toxicity. Each point represents the F1-score (Y-axis) for the positive ("toxic") class achieved when using a given PAI score as the threshold to transform the continuous PAI measure into an indicator variable (X-axis). The analysis is performed on the 300-tweet validation sample, where human annotations were made according to the official PAI definition of toxicity. The vertical red dashed line indicates the threshold of 0.26, which maximizes the F1-score. This data-driven procedure is used to select the threshold for the main analysis.