

Setting the Tone or Supplying the Words? Donald Trump’s Facebook Posts and the Hostility of His Comment Section in the 2016 Republican Primaries

Tristan Hotham*

2026

Preprint, not peer reviewed

Abstract

Do politicians set the tone of the conversations beneath their posts? I link 730 posts on Donald Trump’s Facebook page during the 2016 Republican primaries to the 5.1 million comments that could be retrieved beneath them, and measure toxicity with an open classifier on a sample of 87,600 comments and with a dictionary on every comment. Comparing posts published on the same day, comments were more toxic beneath posts that named Hillary Clinton (by 3.5 percentage points, against an average of about 10%) and, on the dictionary, beneath posts that named a Republican rival or used negative words. The 35 critical posts, which named an opponent or the media and used negative words, were followed by comment sections about a quarter more toxic and by about a quarter more comments. Much of that difference is accounted for by comments using the same kind of words the posts used, such as “liar”. Such comments were twice as common beneath critical posts. When those words are removed and the comments re-scored, the remaining difference is between about half a point and one and a half points, and is statistically clear in some specifications but not others. Posts that named Clinton were the exception: they were followed by more toxic comments even with such words removed. The subject of the comments also followed the post: when a post named Clinton, 13% of comments mentioned her, against 3% otherwise. The results are associations, not causal effects. They suggest that a candidate’s posts are linked more clearly to the subject and vocabulary of the conversation beneath them than to its general level of hostility.

Keywords: incivility; elite cues; comment sections; Facebook; Donald Trump; 2016 US presidential election; agenda setting; toxicity

1 Introduction

When a politician posts on Facebook, the post is only part of what people see. Beneath it forms a comment section that the campaign does not write, often far larger than the post itself. In an earlier preprint I showed that about one comment in ten on Donald Trump’s Facebook page during the 2016 Republican primaries was toxic, that this rate was steady over time, and that toxicity was spread across the whole crowd of commenters rather than concentrated among the most active ([Hotham 2026](#)). That

*Independent researcher. ORCID 0000-0003-4931-9349. Correspondence: contact@tristanhotham.com. The data were collected during doctoral research funded by the Economic and Social Research Council. The views expressed are the author’s own and do not represent those of any employer.

paper treated the comment section as a crowd. This one asks how the crowd's comments related to what the candidate posted.

There are good reasons to expect a link. Experiments show that people exposed to uncivil messages from political elites become angrier and more likely to respond uncivilly (Gervais 2015, 2017), and that exposure to toxic comments makes people's own comments more toxic (Kim et al. 2021). Posts that attack the out-group draw more engagement (Rathje et al. 2021; Brady et al. 2017), and negative campaign messages have been followed by more uncivil talk on social media (Hopp and Vargo 2017). Yet most studies of politicians' Facebook posts measure engagement, such as likes, shares and comment counts (Bene 2017; Heiss et al. 2019; Eberl et al. 2020), rather than what the comments say. And where comment tone has been linked to elite messages, it is rarely clear whether audiences become generally more hostile or simply repeat the words they were given. Gervais (2017) asks whether the uncivil response to elite incivility is "more than mimicry". This paper asks the same question of a real campaign.

I link each of 730 posts on Trump's page between January and April 2016 to the comments that could be retrieved beneath it: 5.1 million comments in all. I code each post for whom it named and how it spoke, and I measure the toxicity of its comment section with an open classifier applied to a sample of 120 comments per post and with a dictionary of profanity and insults applied to every comment. Because the campaign posted about eight times a day, I can compare posts published on the same day, which holds constant the news, the stage of the race and the mood of the audience on that day.

I ask four questions. **RQ1.** Which features of a post were associated with more toxic comment sections? **RQ2.** How much of any difference reflects commenters repeating the post's own vocabulary? **RQ3.** Did the people a post named become the subject of the comments? **RQ4.** Was any extra hostility confined to comments about the people named?

The answers are more qualified than the simplest story. Posts that named Clinton were followed by more toxic comments on the classifier, including when the post's kind of vocabulary is removed. Posts that combined an opponent with negative words were followed by more toxic and more numerous comments, but much of the extra toxicity is accounted for by comments using the same kind of negative words as the post. The subject of the comments also followed the subject of the post. The clearest links between the candidate's posts and the comments were in what the comments were about and the words they used.

2 Elite cues and the tone of the crowd

Elite incivility and its echo. Incivility is usually defined as communication that breaks norms of respect, such as name-calling, insults, vulgarity and aspersion (Coe et al. 2014; Papacharissi 2004; Stryker et al. 2016). Political elites model how politics is talked about, and when they are uncivil audiences may follow. In experiments, people who read uncivil posts from political elites felt angrier and were more likely to write uncivil comments (Gervais 2015), and anger appears to carry much of the effect (Gervais 2017). Incivility can also lower approval of the politicians who use it, even among their own supporters (Frimer and Skitka 2018). Outside the laboratory, comments on news stories are more uncivil when the stories feature conflict and partisan figures (Coe et al. 2014; Muddiman and Stroud 2017; Su et al. 2018; Ziegele et al. 2014), incivility towards politicians on Twitter rises when they are in the news (Theocharis et al. 2016, 2020), and negative campaign advertising has been followed by more uncivil talk online (Hopp and Vargo 2017). Users also learn local norms of incivility from what others write (Shmargad et al. 2022; Cheng et al. 2017). Trump's 2016 campaign made unusually heavy use of negative messages

about his opponents (Gross and Johnson 2016) and was widely described as breaking norms of civil political speech (Ott 2017; Enli 2017), which makes his page a natural place to study these processes, though not one from which to generalise about candidates in general. My first question is which features of a post were associated with more toxic comment sections (RQ1).

Tone or words? An association between a critical post and a hostile comment section could arise in two ways. Commenters might become generally more hostile, in a change of mood that spreads beyond the post’s target. Or they might take up the post’s language, repeating its nicknames and insults. The second is mimicry rather than a change in mood, and it matters for how we read toxicity measures: a supporter repeating a candidate’s nickname for a rival will be scored as toxic by most classifiers and dictionaries. I therefore ask how much of any difference reflects the post’s own vocabulary (RQ2).

Agenda setting in miniature. The news media influence what audiences think about by choosing what to cover (McCombs and Shaw 1972), and in 2016 Trump’s own messages shaped the agenda of news coverage (Wells et al. 2016). A candidate’s post may do the same for the comment section beneath it. This matters for tone because comments about opponents are more often uncivil than comments about one’s own side (Rains et al. 2017; Rossini 2022). I ask whether the people named in a post became the subject of the comments (RQ3), and whether any extra hostility was confined to comments about them (RQ4).

3 Data and measures

Posts and comments. The data were collected with Netvizz (Rieder 2013) during my doctoral research (Hotham 2021). The posts are those published on Trump’s page between 1 January and 4 April 2016 (Eastern time) for which comments were collected: 730 of the 760 posts published in that period, leaving aside a batch of 133 photo-album uploads on New Year’s Day. The 30 missing posts include 11 from 22 March, when no comments were collected. For each post I have its text, its type (photo, video, link or text only) and Facebook’s counts of comments, reactions and shares as recorded in October 2016. Beneath these posts were 5,140,922 comments that could be retrieved, from 1.1 million commenters identified by consistent anonymised codes. The corpus and its collection are described in my earlier preprint (Hotham 2026), including its main limitation: most posts were collected months after publication, so the comments are those that survived moderation and deletion. The share of comments that could be retrieved was about the same for critical posts and other posts (76% and 77% of Facebook’s count), which suggests that removal was not concentrated beneath critical posts.

Post features. I coded each post with transparent word lists (Appendix Table 3). The features capture whom the post named (Clinton; a Republican rival; Obama, Sanders or the Democrats; the media), whether it used negative words (such as “liar”, “weak”, “disgrace”, “loser” or “dishonest”), and what it asked or reported (asking people to vote, thanking supporters, mentioning polls, rallies and appearances, issues such as immigration, security and the economy, endorsements, and the campaign’s slogans). I also identify posts published on the evenings of the six Republican debates in which Trump took part. I call a post a *critical post* if it both named an opponent or the media and used negative words: 35 posts, published on 22 days. I scored each post’s own text with the same toxicity classifier used for the comments.

To check the word lists, a random sample of 100 posts was read and coded for each main feature (Appendix Table 4). This reading was carried out with the assistance of an AI model (Claude, by Anthropic) and was not blind to the word lists; it has not been checked by a second human coder, so it is

a check on obvious errors rather than a formal reliability test. The lists for naming Clinton, rivals, the Democrats and the media, and for thanking supporters, agreed with the reading in 99 or 100 of the 100 posts. The list for asking people to vote missed four of 20 such posts. The critical-post definition made no false positives, and all 35 critical posts were also read to confirm that each criticised a named person or institution, but it found only four of the nine posts in the sample that criticised a named opponent. The definition is therefore precise but conservative: many critical posts fall into the comparison group, which, if anything, makes differences between the groups smaller. With so few positive cases in the sample, these figures say little about precision and recall for the rarer features, and the lists for negative words, polls and debate nights were not checked in this way.

Comment toxicity. I measure the tone of each post’s comment section in two ways. The first uses Detoxify (original model; Hanu and Unitary team 2020), an open classifier trained on crowd-labelled comments (Jigsaw and Conversation AI 2018), applied to a random sample of 120 comments from each post (87,600 in all). A comment counts as toxic if its toxicity score is 0.5 or more; comments were truncated at 96 word-pieces. Rates are weighted to all of each post’s comments. The second applies a dictionary of profanity and insults to every comment beneath each post. The two measures are related (the classifier separates dictionary-flagged comments from others with an area under the curve of 0.91; Hotham 2026), but they are not independent: both respond to insult words, and they make some of the same errors (Dixon et al. 2018; Sap et al. 2019; Hede et al. 2021). The average post’s comment section was 9.6% toxic on the classifier and 6.2% flagged on the dictionary.

Echoed words. To address RQ2, I flag comments that use any of the negative words in the post word list, or the campaign’s nicknames for its rivals (such as “lyin’”, “crooked”, “choker” and “low energy”). I call these *echoed words*; they appear in 7.5% of comments. I then recompute both measures excluding these comments, and for the dictionary also excluding these words from the dictionary itself. These echo-free measures average 7.4% on the classifier and 2.8% on the dictionary. Because removing whole comments also removes comments that are hostile for other reasons, I add a third, masked measure: I delete the echoed words from the text of each comment that contains them, keep every comment, and score the masked text again with the classifier (average 7.9%). Of the comments containing echoed words, 36% were toxic as written and 13% once the words were deleted.

Comment subjects and volume. I measure the share of each post’s comments that mentioned Clinton, a Republican rival or the media, using the same name lists as for the posts. For RQ4, I flag comments that refer to any opponent or the media by name or nickname, using an expanded list (Appendix). I measure volume with Facebook’s own count of comments on each post.

Models. I regress each outcome on all the post features at once, with a fixed effect for each day, so that each post is compared only with the other posts the campaign published that day. Controls are the post’s type, its length, the share of capital letters, the time of day in four bands and the time since the campaign’s previous post. Standard errors are clustered by day. With about eight posts a day, the day fixed effects absorb the news, the stage of the race, the day of the week and the mood of the audience. Because I test many features, I also report false discovery rate adjusted q -values (Benjamini and Hochberg 1995). For critical posts, which appear on only 22 days, I add randomisation inference, reshuffling which posts on each day are critical 1,000 times, and re-estimate the model leaving out each of the 22 days in turn. Comment-level models on the scored sample also use day fixed effects and standard errors clustered by day; they are unweighted and so give each sampled comment equal weight.

Table 1. Post features and the comment section, Trump’s page, January to April 2016. Same-day comparisons.

	All comments		Excluding echoed words		Comments
	Classifier	Dictionary	Classifier	Dictionary	
Names Clinton	+3.49***† (0.78)	+1.45***† (0.42)	+2.91***† (0.70)	+0.30 (0.20)	+6.45 (7.82)
Names a Republican rival	+1.75* (0.77)	+1.47***† (0.38)	+0.77 (0.61)	+0.04 (0.15)	+12.60 (7.13)
Names Obama, Sanders or Democrats	+0.42 (0.76)	+0.17 (0.36)	+0.67 (0.67)	+0.42 (0.21)	−6.28 (9.33)
Names the media	+0.40 (0.68)	+0.49 (0.34)	+0.28 (0.67)	+0.43***† (0.15)	+7.14 (10.54)
Negative words	+1.74* (0.66)	+1.99***† (0.39)	+0.77 (0.69)	+0.55***† (0.14)	+15.85 (10.15)
Asks people to vote	−0.46 (0.44)	+0.24 (0.26)	−0.74* (0.35)	−0.11 (0.07)	−22.00*** (5.60)
Says thank you	−0.63 (0.36)	−0.46***† (0.16)	−0.34 (0.32)	−0.16 (0.08)	−13.11** (4.69)
Mentions polls	−1.21 (0.70)	−0.51 (0.31)	−0.60 (0.62)	−0.21* (0.10)	+7.18 (8.47)
Debate-night post	−1.46 (0.88)	−0.84 (0.59)	−0.94 (0.79)	−0.12 (0.19)	−60.60*** (14.02)
Rally or appearance	−0.24 (0.63)	−0.15 (0.29)	−0.03 (0.59)	+0.02 (0.18)	−12.42 (10.50)
Immigration	+1.06 (0.68)	+0.44 (0.33)	+0.76 (0.62)	+0.40* (0.18)	+14.44 (9.07)
Security	+0.28 (0.65)	−0.05 (0.31)	+0.38 (0.65)	+0.17 (0.15)	+11.65 (9.48)
Economy	−0.47 (0.55)	+0.15 (0.26)	−0.51 (0.44)	−0.09 (0.13)	+1.97 (9.04)
Endorsement	+0.02 (0.91)	+0.30 (0.44)	+0.08 (0.69)	−0.06 (0.20)	−17.22* (8.50)
Slogan	−0.38 (0.41)	−0.36 (0.25)	−0.23 (0.32)	−0.12 (0.08)	+10.84 (5.57)
Photo	−0.72 (0.42)	−0.38 (0.20)	−0.84* (0.40)	−0.36*** (0.10)	+10.88 (5.83)
Video	+0.18 (0.51)	−0.08 (0.23)	+0.05 (0.41)	−0.08 (0.10)	+22.71*** (6.63)
Text only	+0.64 (0.50)	+0.07 (0.27)	+0.50 (0.41)	+0.09 (0.13)	+30.78*** (7.50)
Mean of outcome	9.6	6.2	7.4	2.8	

Linear models with day fixed effects; 730 posts. Coefficients are differences in percentage points of comments flagged (log points $\times 100$ for the number of comments, roughly a percentage difference). Standard errors clustered by day in parentheses. Controls: post length, share of capitals, time-of-day band, time since previous post. Type is compared with link posts.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; † false discovery rate $q < 0.05$ across the 15 content features.

4 Findings

4.1 RQ1: Which posts were followed by more toxic comments

Table 1 and Figure 1 compare each post with the other posts published the same day. The clearest result concerns Clinton. A post that named her was followed by a comment section 3.5 percentage points more toxic on the classifier, against an average of about 10%, and 1.5 points more on the dictionary. Both survive the adjustment for multiple tests. Posts that named a Republican rival (1.8 points on the classifier, 1.5 on the dictionary) or used negative words (1.7 and 2.0 points) were also followed by more toxic comments; on the dictionary these survive adjustment, on the classifier they are weaker ($q = 0.13$ and $q = 0.08$). Naming Obama, Sanders or the Democrats made no clear difference, and naming the media made a difference only on the echo-free dictionary (0.4 points). Posts that thanked supporters were followed by slightly less profane comment sections on the dictionary. Most other features made little difference to tone.

The 35 critical posts, which named an opponent or the media and used negative words, stand out. Their comment sections averaged 13.0% toxic on the classifier, against 9.4% for other posts, and 9.9% against 6.1% on the dictionary. Comparing posts published the same day, critical posts were followed by comment sections 2.3 percentage points more toxic on the classifier (standard error 0.8, $p = 0.009$; randomisation inference $p = 0.008$) and 3.7 points more on the dictionary ($p < 0.001$; none of 1,000 reshuffles produced so large an estimate), and by 1.1 points more insults ($p = 0.007$). Severe toxicity, obscenity, threats and identity attacks did not differ. Leaving out each of the 22 days in turn, the classifier estimate ranges from 1.7 to 2.6 points and the dictionary estimate from 3.4 to 3.9. The estimates on all

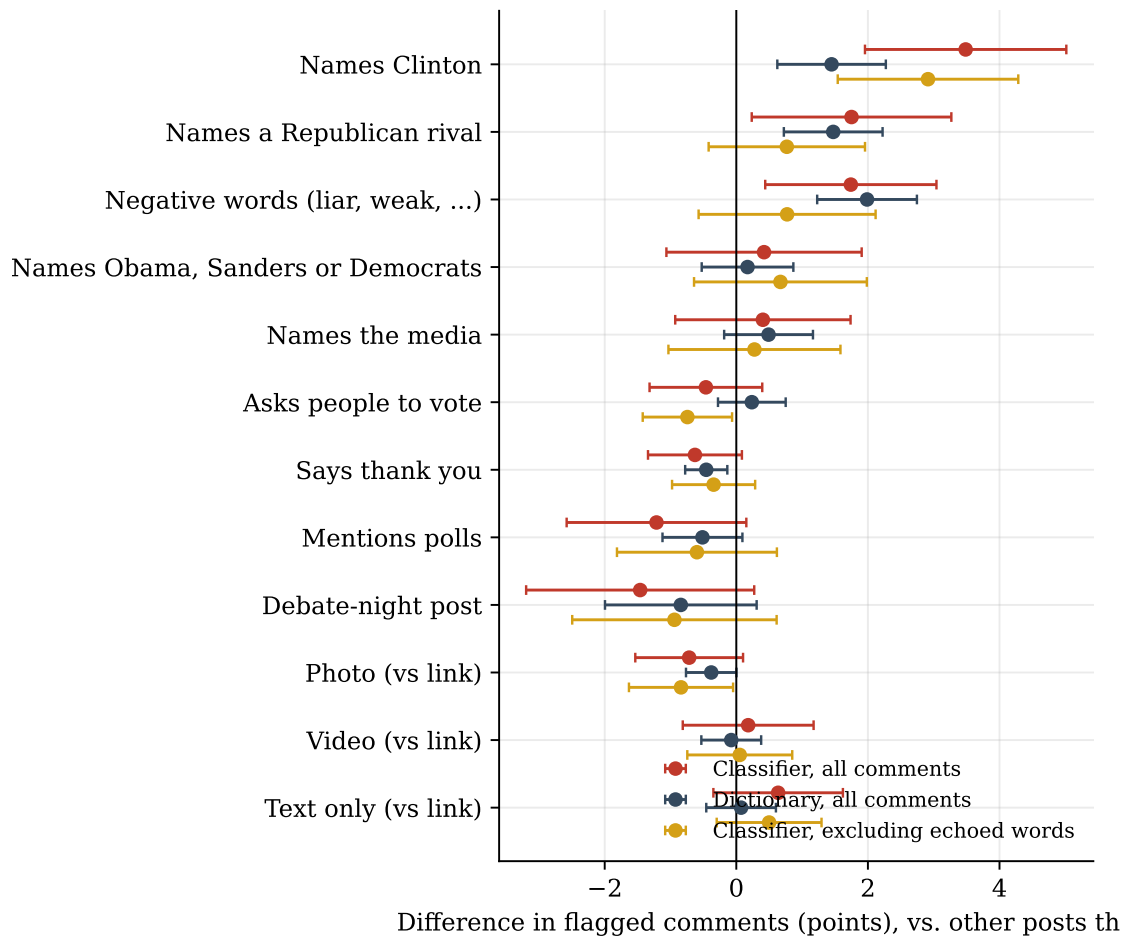


Figure 1. Differences in the share of flagged comments by post feature, compared with other posts published the same day (percentage points, 95% confidence intervals). The classifier is applied to a sample of 120 comments per post; the dictionary to every comment. The third series excludes comments using the negative words and nicknames found in the posts.

comments are similar with week instead of day fixed effects, when the number of reactions to each post is controlled as a rough measure of reach, when debate-night posts are dropped, and when critical posts that qualify only by naming the media are excluded.

Critical posts also had more comments. The median critical post had 11,557 comments on Facebook’s count, against 7,717 for other posts, and comparing posts on the same day, critical posts had about a quarter more (24 log points, $p = 0.009$). Posts asking people to vote, thanking supporters or announcing endorsements had fewer comments. Posts published on debate nights had far fewer, probably because the campaign posted many times during each debate and attention was split between them. Videos and text-only posts had more comments than links.

A post’s own toxicity, as scored by the same classifier, pointed the same way: each 0.1 increase in the post’s score was associated with comment sections 0.8 points more toxic (standard error 0.2). Most posts scored close to zero, however (median 0.001), and only ten reached the 0.5 threshold, so this estimate rests on a small number of posts.

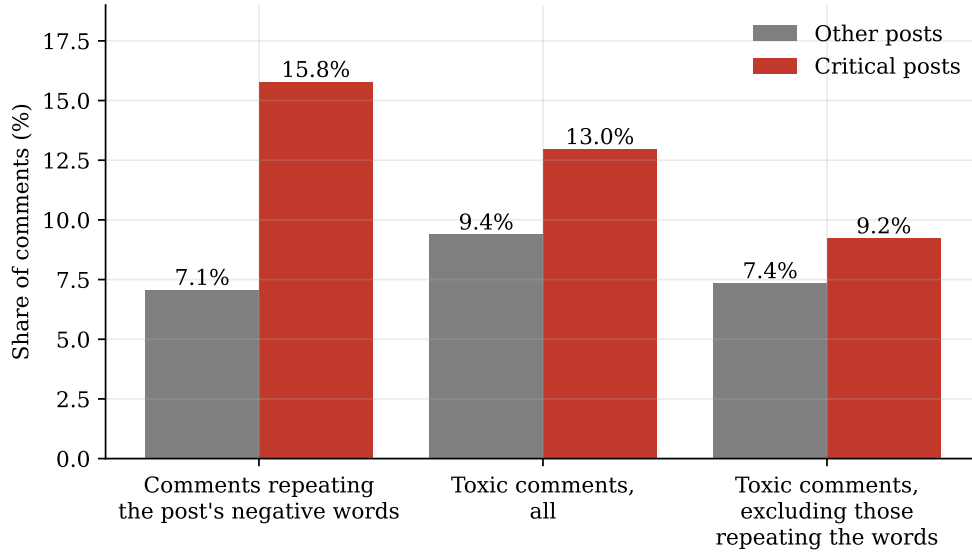


Figure 2. Comments beneath critical posts and other posts: the share using the negative words and nicknames found in the posts, the share toxic on the classifier, and the share toxic once those comments are excluded. Raw weighted averages across posts. Most of the remaining raw gap (9.2 against 7.4%) disappears when posts are compared within the same day.

4.2 RQ2: Tone or words?

Commenters often used the same kind of words the posts used (Figure 2). Beneath critical posts, 15.8% of comments contained one of the posts’ negative words or nicknames, against 7.1% beneath other posts (8.2 points more, comparing posts on the same day). These comments are scored as toxic far more often than others, 37% against 8% on the classifier, partly because the words themselves are insults.

Once the echoed words are taken out, much of the difference goes. At the level of posts, the comment sections of critical posts were 0.4 points more toxic on the classifier than those of other posts the same day when comments with echoed words are excluded (95% interval -1.1 to 1.9 , $p = 0.60$), and 0.6 points more when the words are masked and the comments re-scored (-0.7 to 1.9 , $p = 0.40$). On the dictionary, with the echoed words removed, they were 0.8 points more profane ($p < 0.001$). Comment-level models, which give every sampled comment equal weight rather than weighting each post’s comments to its size, find somewhat larger remaining differences: 1.5 points excluding echo comments ($p = 0.04$) and 1.6 points with the words masked ($p = 0.008$), against 3.4 points before either step. The difference between the two approaches arises because the post-level estimates give more weight to the comments of large posts.

Taken together, between about half and four-fifths of the extra toxicity beneath critical posts is accounted for by the post’s kind of vocabulary appearing in the comments. What remains is between about half a point and one and a half points, on a base of about 7 to 8%. In relative terms that could still be a rise of up to a fifth, so a modest general rise in hostility cannot be ruled out. The dictionary points the same way: its echo-free difference is small in absolute terms but about a quarter of its low base.

Clinton was the exception. Posts that named her were followed by comment sections more toxic on the classifier whichever way the echoed words are handled: 2.9 points excluding echo comments ($q = 0.001$) and 3.1 points with the words masked ($p < 0.001$). Because classifiers can over-flag gendered language (Dixon et al. 2018), and comments beneath posts about Clinton used female pronouns and gendered

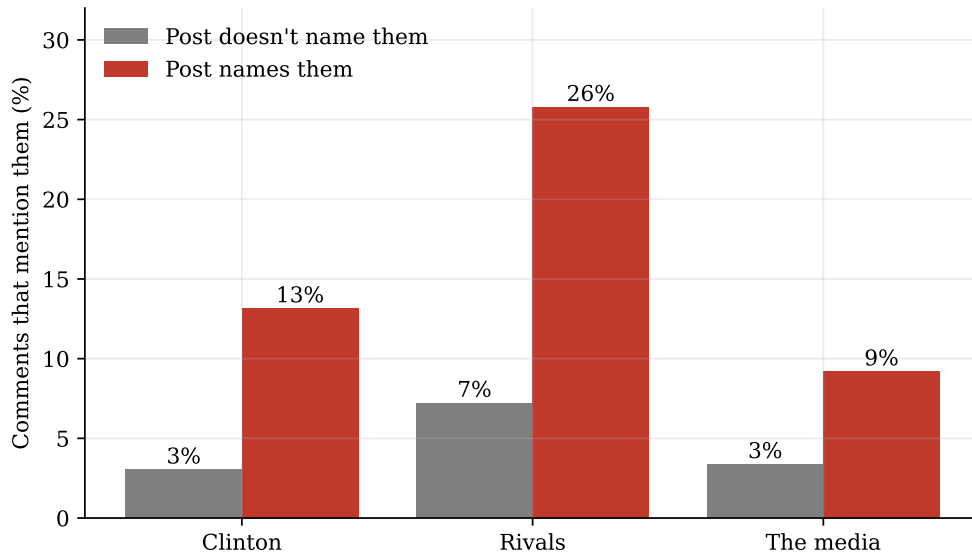


Figure 3. Share of comments mentioning Clinton, a Republican rival or the media, by whether the post named them. Every retrieved comment, averaged across posts.

words far more often (15% against 4%), I re-estimated this at the comment level excluding comments with such words as well as echo comments; the difference remains 2.2 points ($p < 0.001$). On the dictionary, however, the echo-free difference for Clinton is small and not significant (0.3 points), so this result rests on the classifier.

4.3 RQ3: What the comments talked about

The subject of the comments followed the subject of the post (Figure 3, Table 2). When a post named Clinton, 13% of its comments mentioned her, against 3% beneath other posts. When a post named a Republican rival, 26% of comments mentioned a rival, against 7%. When a post named the media, 9% of comments mentioned the media, against 3%. Comparing posts on the same day with all features entered together, naming Clinton was associated with 9.7 points more comments mentioning her, naming a rival with 17.2 points more comments mentioning rivals, and naming the media with 4.9 points more comments mentioning the media. Each association was largely specific to its subject. Debate-night posts were followed by more talk about the rivals on stage (5.5 points). Because posts and comments are measured with the same name lists, these associations partly reflect the obvious fact that comments respond to the post they sit beneath; their size, rather than their existence, is the finding.

4.4 RQ4: Was the extra hostility confined to the target?

The extra toxicity beneath critical posts was not confined to comments about the people the posts named. Using the expanded list of names and nicknames, 21% of comments referred to an opponent or the media. Among the other 79%, 12.3% were toxic beneath critical posts against 9.3% beneath other posts, a difference of 3.4 points comparing posts on the same day ($p < 0.001$). Among comments that did refer to an opponent, the difference was 2.0 points and not significant. A model that tests whether the difference itself differs between the two kinds of comment finds no evidence that it does (-0.5 points, $p = 0.62$). Adding pronouns to the list, which on this page often refer to Trump himself, gives the same

Table 2. What the comments talked about. Share of each post’s comments mentioning each subject (%).

Post feature	Clinton	Republican rival	The media
Names Clinton	+9.68*** (0.79)	−1.06 (0.68)	−0.56 (0.54)
Names a Republican rival	+0.71** (0.24)	+17.22*** (1.15)	−0.58 (0.35)
Names the media	−0.25 (0.39)	+0.40 (0.61)	+4.85*** (1.03)
Names Obama, Sanders or Democrats	−0.14 (0.52)	−0.17 (0.74)	−0.32 (0.38)
Negative words	+0.51 (0.28)	+0.45 (0.67)	+0.50 (0.50)
Debate-night post	−0.19 (0.34)	+5.47*** (1.26)	−0.02 (0.55)
Mean of outcome	3.6	8.8	3.7

Linear models with day fixed effects and the full set of post features and controls; 730 posts; every retrieved comment; same name lists as for posts. Standard errors clustered by day in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

picture. Comments mentioning Trump himself were also more toxic beneath critical posts (3.5 points, $p = 0.004$), which may reflect criticism of Trump from opponents as well as hostility from supporters. The data cannot say at whom these comments were aimed.

The difference was concentrated in top-level comments written in response to the post (4.5 points, $p < 0.001$) rather than in replies between commenters (−1.0 points, not significant). It appeared among both the most active 1% of commenters (2.3 points, $p = 0.04$) and everyone else (3.8 points, $p < 0.001$). The mix of commenters changed only a little. The share of comments written by the most active 1% did not differ beneath critical posts. The share written by people who commented only once was 2.4 points lower, on an average of 10% ($p = 0.04$). Because one-time commenters were the most toxic group in the earlier analysis (Hotham 2026), this shift would, if anything, lower toxicity beneath critical posts rather than raise it. (The most active 1% wrote 31% of all comments but 36% of the average post’s comments, because they commented more on posts with fewer comments.)

5 Discussion

This paper asked whether a candidate’s posts set the tone of the conversation beneath them. The answer depends on what is meant by tone.

First, what the comments were about followed the post. Naming someone was followed by a conversation about them, with the share of comments mentioning the person several times higher. This is agenda setting (McCombs and Shaw 1972) on a small scale and in real time, consistent with evidence that Trump’s messages shaped the agenda of news coverage in 2016 (Wells et al. 2016).

Second, the words commenters used followed the post. Beneath critical posts, commenters were twice as likely to use the posts’ kind of negative words and nicknames, and this accounts for much of the higher toxicity of those comment sections. What remains is modest and its size depends on how comments are weighted. In Gervais (2017)’s terms, much of the response to elite incivility on Trump’s page looks like mimicry, though some rise in other hostility cannot be ruled out. This matters for how toxicity measures are read. A supporter repeating a candidate’s nickname for a rival is scored as toxic, and from the rival’s point of view it is hostile. But it is different from a general rise in anger, and studies that link elite messages to audience toxicity should distinguish the two.

Third, posts that named Clinton were the exception. On the classifier they were followed by more toxic comments however the echoed words are handled, and the difference is not explained by gendered words;

on the echo-free dictionary the difference is small. This fits work showing that incivility is concentrated in talk across group lines (Rains et al. 2017; Rossini 2022). One possible explanation, which these data cannot test, is that in the Republican primaries Clinton was the shared opponent of almost everyone on Trump’s page, while comments about rival Republicans may have been divided between their supporters and Trump’s.

Fourth, critical posts had more comments as well as more toxic ones. Studies of politicians’ Facebook pages usually ask whether negativity drives engagement (Bene 2017; Heiss et al. 2019; Eberl et al. 2020; Hopp and Vargo 2017). These results suggest that some of the engagement it drives consists of the audience repeating the negativity back.

Whether these patterns are specific to Trump cannot be answered with one candidate. Comparable posts by other candidates might be followed by similar responses, and some of the hostility beneath Trump’s critical posts may have been directed at Trump himself.

6 Limitations

The results are associations, not causal effects. Comparing posts on the same day removes much of what might drive both a post’s content and its comment section, such as the news and the state of the race, but the campaign may have chosen to criticise in response to events within a day that also angered its audience, and critical posts may have reached different people through Facebook’s ranking. The two toxicity measures are not independent, and neither is a substitute for human coding. The post coding uses word lists that are precise but miss some critical posts. Only 35 posts met the critical-post definition, on 22 days, and only ten posts were themselves scored as toxic, so these estimates are imprecise despite the randomisation checks. The echo-free and masked measures depend on a fixed list of words, and the size of what remains after removing them depends on how comments are weighted. The check of the word lists was AI-assisted and not independently coded. The volume measure is Facebook’s count of comments as recorded in October 2016, months after the posts, and reflects reach and ranking as well as response. The comments are those that survived until collection, which for most posts was months later. Finally, the study covers a single candidate in a single primary campaign.

7 Conclusion

On Donald Trump’s Facebook page in the 2016 primaries, the candidate’s posts were closely linked to what his comment section talked about and the words it used. Posts that named someone were followed by talk about that person. Posts that criticised opponents with negative words were followed by more comments and more toxic ones, much of it accounted for by commenters using the same kind of words. Posts about Clinton were followed by more toxic comments on the classifier even with those words removed. The clearest links were between the candidate’s posts and the subject and vocabulary of the conversation beneath them. Any general rise in hostility was smaller and harder to pin down.

Provenance and AI use

The data were collected with Netvizz for my doctoral research, before Netvizz closed in 2019. This paper builds on my doctoral thesis (Hotham 2021). The analysis and writing were completed in 2026.

The analysis code and drafts of the text were prepared with the assistance of an AI system (Claude, by Anthropic), which was also used to review drafts. I checked the analysis and the text and take responsibility for both.

Data availability

Aggregate post-level data supporting the tables and figures (post features, comment-section rates and comment counts, without post or comment text) are available with this preprint. Comment text and commenter codes are not shared, in line with the platform's terms and to protect commenters.

Appendix

Table 3. Post features: word lists and number of posts.

Feature	Words and phrases (case-insensitive, whole words)	Posts
Names Clinton	hillary, clinton	40
Names a Republican rival	cruz, ted, rubio, marco, jeb, bush (not George W. Bush), kasich, carson, fiorina, christie, rand paul	62
Names Obama, Sanders or Democrats	bernie, sanders, obama, democrat(s)	23
Names the media	media, press, cnn, fox, megyn, nbc, abc, cbs, msnbc, new york times, nytimes, washington post, reporter(s), journalist(s)	40
Negative words	lie(s), liar, lying, lyin, dishonest, disgrace, weak, failed, loser, pathetic, incompetent, stupid, corrupt, crooked, nasty, terrible, disaster, sad, dumb, clown, phony, fraud	58
Critical post	names any of the four above and uses negative words	35
Asks people to vote	vote, voting, caucus, polling location or place, polls open or close	134
Says thank you	thank, thanks	242
Mentions polls	poll, polls	39
Debate-night post	published from 8pm Eastern on the day of a Republican debate in which Trump took part	66
Rally or appearance	rally, rallies, tonight, join me, tickets, livestream, tune in, interview	43
Immigration	wall (not Wall Street), border, illegal, immigration	39
Security	isis, terror, military, veteran(s), vets, police, islam, muslim(s)	46
Economy	job(s), trade, china, economy, debt, tax, deficit, worker(s)	57
Endorsement	endorse, endorsement	28
Slogans	make america great, #maga, trump train, #trump2016	243
Echoed words (comments)	the negative words above, plus choker, lightweight, low energy, little marco	
Refers to an opponent (comments)	the names above, plus paul, dems, liberal(s), kelly, establishment, rino, gop and the nicknames above	

Post types: 256 photos, 183 text only, 176 links, 115 videos.

Table 4. Check of the word lists against a reading of 100 randomly selected posts.

Feature	In reading	By word list	Both	Word list only	Reading only
Names Clinton	5	6	5	1	0
Names a Republican rival	9	9	9	0	0
Names Obama, Sanders or Democrats	3	3	3	0	0
Names the media	4	4	4	0	0
Critical post	9	4	4	0	5
Asks people to vote	20	16	16	0	4
Says thank you	34	34	34	0	0

The one false positive for Clinton is a reference to the town of Clinton, Iowa. In the reading, a critical post is one that criticises a named opponent or the media.

References

- Bene, M. (2017). Go viral on the Facebook! Interactions between candidates and followers on Facebook during the Hungarian general election campaign of 2014. *Information, Communication & Society*, 20(4):513–529.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300.
- Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., and Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.
- Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*.
- Coe, K., Kenski, K., and Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64(4):658–679.
- Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.
- Eberl, J.-M., Tolochko, P., Jost, P., Heidenreich, T., and Boomgaarden, H. G. (2020). What’s in a post? How sentiment and issue salience affect users’ emotional reactions on Facebook. *Journal of Information Technology & Politics*, 17(1):48–65.
- Enli, G. (2017). Twitter as arena for the authentic outsider: Exploring the social media campaigns of Trump and Clinton in the 2016 US presidential election. *European Journal of Communication*, 32(1):50–61.
- Frimer, J. A. and Skitka, L. J. (2018). The Montagu principle: Incivility decreases politicians’ public approval, even with their political base. *Journal of Personality and Social Psychology*, 115(5):845–866.
- Gervais, B. T. (2015). Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics*, 12(2):167–185.

- Gervais, B. T. (2017). More than mimicry? The role of anger in uncivil reactions to elite political incivility. *International Journal of Public Opinion Research*, 29(3):384–405.
- Gross, J. H. and Johnson, K. T. (2016). Twitter taunts and tirades: Negative campaigning in the age of Trump. *PS: Political Science & Politics*, 49(4):748–754.
- Hanu, L. and Unitary team (2020). Detoxify. GitHub repository, <https://github.com/unitaryai/detoxify>.
- Hede, A., Agarwal, O., Lu, L., Mutz, D. C., and Nenkova, A. (2021). From toxicity in online comments to incivility in American news: Proceed with caution. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.
- Heiss, R., Schmuck, D., and Matthes, J. (2019). What drives interaction in political actors' Facebook posts? Profile and content predictors of user engagement and political actors' reactions. *Information, Communication & Society*, 22(10):1497–1513.
- Hopp, T. and Vargo, C. J. (2017). Does negative campaign advertising stimulate uncivil communication on social media? Measuring audience response using big data. *Computers in Human Behavior*, 68:368–377.
- Hotham, T. (2021). *How Do Political Parties Use Facebook and What Does It Offer to Their Campaigns*. PhD thesis, University of Bath.
- Hotham, T. (2026). Who speaks in the candidate's comment section? Participation, toxicity and reward on Donald Trump's and Hillary Clinton's Facebook pages in 2016. Preprint, Zenodo, October 2026, <https://doi.org/10.5281/zenodo.23287107>.
- Jigsaw and Conversation AI (2018). Toxic comment classification challenge. Kaggle, <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>.
- Kim, J. W., Guess, A., Nyhan, B., and Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, 71(6):922–946.
- McCombs, M. E. and Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2):176–187.
- Muddiman, A. and Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, 67(4):586–609.
- Ott, B. L. (2017). The age of Twitter: Donald J. Trump and the politics of debasement. *Critical Studies in Media Communication*, 34(1):59–68.
- Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, 6(2):259–283.
- Rains, S. A., Kenski, K., Coe, K., and Harwood, J. (2017). Incivility and political identity on the Internet: Intergroup factors as predictors of incivility in discussions of news online. *Journal of Computer-Mediated Communication*, 22(4):163–178.
- Rathje, S., Van Bavel, J. J., and van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26):e2024292118.

- Rieder, B. (2013). Studying Facebook via data extraction: The Netvizz application. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 346–355.
- Rossini, P. (2022). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49(3):399–425.
- Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.
- Shmargad, Y., Coe, K., Kenski, K., and Rains, S. A. (2022). Social norms and the dynamics of online incivility. *Social Science Computer Review*, 40(3):717–735.
- Stryker, R., Conway, B. A., and Danielson, J. T. (2016). What is political incivility? *Communication Monographs*, 83(4):535–556.
- Su, L. Y.-F., Xenos, M. A., Rose, K. M., Wirz, C., Scheufele, D. A., and Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. *New Media & Society*, 20(10):3678–3699.
- Theocharis, Y., Barberá, P., Fazekas, Z., and Popa, S. A. (2020). The dynamics of political incivility on Twitter. *SAGE Open*, 10(2). <https://doi.org/10.1177/2158244020919447>.
- Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., and Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens’ uncivil Twitter use when interacting with party candidates. *Journal of Communication*, 66(6):1007–1031.
- Wells, C., Shah, D. V., Pevehouse, J. C., Yang, J., Pelled, A., Boehm, F., Lukito, J., Ghosh, S., and Schmidt, J. L. (2016). How Trump drove coverage to the nomination: Hybrid media campaigning. *Political Communication*, 33(4):669–676.
- Ziegele, M., Breiner, T., and Quiring, O. (2014). What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user comments on news items. *Journal of Communication*, 64(6):1111–1138.