

Who Speaks in the Candidate's Comment Section? Participation, Toxicity and Reward on Donald Trump's and Hillary Clinton's Facebook Pages in 2016

Tristan Hotham*

2026

Preprint, not peer reviewed

Abstract

In 2016 the comment sections under Donald Trump's and Hillary Clinton's Facebook posts drew millions of comments. Who took part, how hostile were they, and was hostility rewarded? I study 5.2 million comments: every comment that could be retrieved from 730 posts on Trump's page during the Republican primaries (January to April 2016), and from 16 posts on the two candidates' pages on 1 October 2016. Commenters appear only as anonymised codes. A weighted sample of 111,925 comments was scored with an open toxicity classifier and checked against a dictionary measure applied to every comment. Participation was very unequal: the most active 1% of 1.1 million commenters wrote 31% of the comments. About one comment in ten was toxic, and the rate barely moved across the primaries, falling slightly on voting days. Toxicity was not rewarded: once the timing of comments is taken into account, toxic comments were as likely as others on the same post to receive at least one like. The heaviest commenters were no more toxic than occasional ones. On 1 October the two candidates' comment sections were equally toxic, but the 1,106 people who commented on both pages were more toxic on Clinton's page than on Trump's. Hostility in these spaces was widespread rather than concentrated among the most active.

Keywords: incivility; toxicity; comment sections; Facebook; 2016 US presidential election; Donald Trump; Hillary Clinton; participation inequality

1 Introduction

Candidates have long been wary of talking with the public online. Two decades ago [Stromer-Galley \(2000\)](#) showed that campaigns avoided real interaction on their websites because it cost them control, and later work found the same caution in the social media era ([Stromer-Galley 2014](#); [Kreiss 2016](#)). Facebook changed the terms. A candidate's page comes with a comment section that the campaign does not write and can only partly moderate. By 2016 those comment sections were very large. In the last month of the general election, Trump's posts drew more than eleven million comments and Clinton's nearly four million.

*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.

These spaces matter for two reasons. They are where supporters, opponents and the curious meet in public under the candidate’s name, so they are a visible part of how a campaign looks. And they are a large, naturally occurring record of how ordinary people talk about politics during a campaign. Most research on online incivility studies news comment sections (Coe et al. 2014; Muddiman and Stroud 2017; Kim et al. 2021; Ziegele et al. 2014; Rains et al. 2017), replies to politicians on Twitter (Theocharis et al. 2016, 2020) or discussion forums (Papacharissi 2004; Rossini 2022). The comment sections of presidential candidates, which people choose to visit because of the candidate, have been studied far less at scale.

This paper asks four questions about them. **RQ1.** How unequal was participation? **RQ2.** How toxic were the comments, and did toxicity rise with the campaign’s moments of conflict? **RQ3.** Was toxicity rewarded with likes? **RQ4.** Were the most active commenters the most toxic? A fifth question compares the candidates. **RQ5.** On the same day in the general election, were the two comment sections equally toxic, and how did people who commented on both pages behave on each?

I answer these with comments collected during my doctoral research (Hotham 2021): every comment that could be retrieved from 730 posts on Trump’s page during the Republican primaries, and from 16 posts on the two candidates’ pages on 1 October 2016. Commenters appear only as anonymised codes, but the codes are consistent across posts and pages, so I can follow the same anonymous person through the campaign and across the two pages. I measure toxicity with an open classifier applied to a stratified sample of 111,925 comments, and check it against a dictionary measure applied to every comment. RQ1 to RQ4 concern Trump’s page in the primaries; RQ5 compares the two pages on one day.

The results run against several common expectations. Participation was very unequal, but the regulars who dominated the conversation were no more toxic than everyone else. Toxicity was common but steady, and the campaign’s big moments moved it only a little. Toxic comments were not rewarded with likes. And on the same day the two candidates’ comment sections were equally toxic, although the people who commented on both were more toxic on Clinton’s page than on Trump’s.

2 Comment sections, incivility and reward

Participation. Online political talk has long been dominated by a few. Graham and Wright (2014) called the most active participants in political forums “superparticipants” and found that a small group produced a large share of posts. Whether this matters depends on how the few behave. If they are the most extreme or the most hostile, the conversation that everyone else sees is distorted. Kim et al. (2021) argue that self-selection does this on Facebook. Studying millions of comments on news pages, they found the comments to be much more toxic than comments written by a representative sample of Americans, and attributed the gap to the kind of people who choose to comment. Candidate pages may be even more selective, because people visit them because of the candidate. I expect participation on these pages to have been highly unequal (RQ1) and, following the self-selection account, the most active commenters to have been the most toxic (RQ4).

Incivility and toxicity. Incivility is usually defined as features of discussion that convey an unnecessarily disrespectful tone, such as name-calling, insults, vulgarity and aspersion (Coe et al. 2014; Papacharissi 2004). Rossini (2022) distinguishes incivility, which is about tone, from intolerance, which attacks people for who they are; the two often appear apart. Computational work tends to measure “toxicity”, defined as a rude, disrespectful or unreasonable comment that is likely to make people leave a discussion (Wulczyn et al. 2017; Jigsaw and Conversation AI 2018). Toxicity measures capture mainly the tonal

side of incivility, and they do not map exactly onto incivility as communication scholars define it (Hede et al. 2021). I use toxicity in that narrower sense and report its components.

Incivility in comment sections is common but uneven. Coe et al. (2014) found that about one in five comments on a newspaper’s site was uncivil. Muddiman and Stroud (2017) found that partisan incivility rose with news values such as conflict, and studies of Twitter find that incivility towards politicians rises when campaigns heat up (Theocharis et al. 2016, 2020). A presidential primary offers clear moments of conflict: televised debates, where candidates attack each other, and voting days, when results come in. I expect toxicity to have risen around debates (RQ2). I had no prior expectation for voting days, which bring both conflict and celebration.

Reward. Social media may not only host hostility but reward it. Posts that express moral outrage or attack the out-group are shared and liked more (Brady et al. 2021; Rathje et al. 2021), and people who receive more likes for outrage go on to express more of it (Brady et al. 2021). Uncivil comments can also set norms that others follow (Shmargad et al. 2022; Cheng et al. 2017). In comment sections, likes help decide which comments are shown first, so reward also shapes what others see. Kim et al. (2021) found that more toxic comments on news pages received more likes, up to a point. I expect toxic comments to have received more likes than civil comments on the same post (RQ3).

Two candidates. Trump’s rhetoric was widely described as breaking norms of civil political speech (Ott 2017; Enli 2017), and exposure to uncivil political talk can make audiences angrier and more uncivil themselves (Gervais 2015; Mutz 2015). On this view, the comment section of a candidate with a more combative style should be more uncivil. Against this, incivility in online political talk often runs across group lines: people are more uncivil towards those they see as opponents (Rains et al. 2017; Muddiman and Stroud 2017; Rossini 2022), in line with the wider growth of affective polarisation (Iyengar et al. 2019), and people who are hostile online tend to be hostile offline too (Bor and Petersen 2022). On this view, what matters is less the host’s style than who takes part and where. Comparing the two pages on the same day, and following people who commented on both, lets me look at both (RQ5).

3 Data

Collection. I collected the comments with Netvizz, an academic tool for extracting public Facebook page data that was available until 2019 (Rieder 2013). For each post, Netvizz returned the comments it could retrieve, with when each was posted, how many likes it had at the time of collection, whether it was a reply to another comment, and an anonymised code for the person who wrote it. The code is a one-way hash of the commenter’s Facebook identifier, so I cannot identify anyone, but the same person receives the same code on every post and on both pages; codes match across extractions made months apart (13,778 people appear in both the February and the autumn extractions of Trump’s page). Facebook counted more comments than the tool could retrieve, so the corpus is the retrievable set rather than a census. Where a comment appeared in more than one extraction, I kept it once.

Timing of collection. The extractions were made at different distances from the posts. The posts of 1 October were extracted six to nine days after publication. Most posts from the primaries were extracted months later, a median of 176 days after publication (interquartile range 71 to 221 days). This has two consequences. Likes on primary comments accumulated over months, so likes cannot be compared across the two parts of the corpus. And comments removed by the campaign, by Facebook or by their authors before extraction are missing, so all estimates describe the comment sections as they survived, a point I return to in the limitations. In the models of likes I control for each comment’s age at extraction.

Table 1. The corpus.

	Comments	Posts	Commenters	Scored sample
Trump, primaries (1 Jan to 4 Apr 2016)	5,140,922	730	1,110,866	87,600
Trump, 1 Oct 2016	32,301	5	18,429	8,819
Clinton, 1 Oct 2016	49,421	11	17,677	13,826
Trump, 29 Oct to 7 Nov 2016	85,163	14		1,680
All, after removing duplicates	5,307,807	760	1,161,753	111,925

Commenters are distinct anonymised codes. The scored sample was drawn by post; rates are weighted back to all comments.

Corpus. The corpus has two parts (Table 1). Dates are in US Eastern time unless stated. The primary corpus covers Trump’s page from 1 January to 4 April 2016, from four weeks before the Iowa caucuses to the day before the Wisconsin primary, with posts on every day. It contains 5,140,922 comments on 730 posts by 1,110,866 distinct commenters. The general-election corpus covers the posts the two candidates published on 1 October 2016 by Coordinated Universal Time, that is, between 8pm on 30 September and 8pm on 1 October Eastern time, five days after the first presidential debate: 32,301 comments on five Trump posts and 49,421 comments on eleven Clinton posts. These were the posts collected for that day. Trump’s page published about eleven posts a day in October, so the five Trump posts are not necessarily every post of that day, and Clinton’s page contributes more posts; I therefore compare the pages per comment rather than in total. A further 85,163 comments on 14 Trump posts from the last ten days of the campaign are used only for a descriptive check. For context, I also use post-level counts for every post on both pages between 10 October and 6 November 2016.

Ethics. The data are public comments collected under the platform’s terms at the time, as part of doctoral research at the University of Bath, and are analysed only in aggregate. Commenters are represented by one-way hashes, I do not quote any comment, and the data released with this paper contain only aggregate counts.

4 Measures and methods

Toxicity. I score comments with Detoxify (Hanu and Unitary team 2020), an open BERT-based classifier trained on the Jigsaw toxic comment data, which consist of Wikipedia talk page comments labelled by crowd workers (Jigsaw and Conversation AI 2018). I loaded the published model weights directly. For each comment it returns probabilities for toxicity, severe toxicity, obscenity, threat, insult and identity attack. I treat a comment as toxic if its toxicity score is 0.5 or more, the midpoint of the probability scale, and I repeat the main analyses at thresholds of 0.3 and 0.7. Comments were truncated at 96 word-pieces; 92% of comments fit in full.

Scoring every comment was not practical, so I scored a stratified sample. From the primary corpus I drew 120 comments at random from every post (87,600 comments). From the 1 October posts I drew up to 1,200 comments from each Trump post and up to 550 from each Clinton post, and added every comment written that day by people who commented on both pages (1,106 people, 12,418 comments). From the late posts I drew 120 per post. Rates and shares are weighted: each comment is weighted by the number of comments in its post (and, on 1 October, in its group of cross-page or other commenters) divided by the number scored, so that they describe all comments rather than the average post. The regression models are not weighted; because the primary sample has the same number of comments

from every post, they give each post equal weight. Confidence intervals for rates come from a bootstrap that resamples posts.

Validation. Classifiers trained on one kind of text can misread another, and toxicity classifiers are known to over-flag comments that mention some identity groups (Dixon et al. 2018; Sap et al. 2019). I had no resources for human coding at this scale, so I rely on a second, independent measure. I built a dictionary of profanity and insult words and applied it to every comment in the corpus. The two measures agree well: the dictionary flags 6.4% of primary comments; 66% of dictionary-flagged comments are toxic by Detoxify against 6% of other comments; and the Detoxify score separates dictionary-flagged from other comments with an area under the curve of 0.91. The two measures overlap in what they capture, so this agreement shows that the classifier behaves sensibly on these comments rather than that it measures incivility in full. I use the dictionary as a full-corpus check on the main results and say where the two disagree. I treat the identity attack label with particular caution.

Mentions. I flag whether each comment mentions Clinton, Sanders, Obama, each of the main Republican rivals, the media, Muslims or related terms, immigration and the border, God, voting, Trump and Trump’s slogans, using lists of names, nicknames and related words. The lists are broad (Table 2 gives what each covers), and a mention is not the same as a target.

Participation. For each anonymised commenter in the primary corpus I count their comments and the number of weeks in which they commented. I call the most active 1%, those who wrote 56 or more comments in the three months, superparticipants, in the spirit of Graham and Wright (2014).

Models. To test whether toxic comments were rewarded, I model each comment’s likes with a Poisson regression with a fixed effect for each post (Cameron and Trivedi 2013), so that toxic comments are compared only with other comments on the same post. I control for comment length, whether it was a reply, how long after the post it was written and its age at extraction, because later comments and replies are shown less and liked less. Because likes are highly skewed, I also report a linear model for whether a comment received any like. Standard errors are clustered by post. To test whether superparticipants were more toxic, I use linear probability and logit models with post fixed effects and standard errors clustered by post and by commenter. To test campaign events, I regress each post’s toxicity rate on indicators for posts published on the days of Republican debates in which Trump took part (14 January, 6 February, 13 February, 25 February, 3 March and 10 March), on voting days (1, 9, 20 and 23 February, 1, 5, 8, 15 and 22 March), on the days after each, and on 11 and 12 March, after Trump’s Chicago rally was cancelled, with week fixed effects, a control for the number of comments on the post and standard errors clustered by day. This is a comparison between posts. For the comparison of the two pages I compare weighted toxicity rates and use a permutation test over the 16 posts, because standard errors clustered on so few posts are unreliable. For people who commented on both pages I use a model with a fixed effect for each person, so that each person is compared only with themselves.

5 Findings

5.1 RQ1: A crowd with a core

Participation in Trump’s comment section was very unequal (Figure 1). Of the 1.1 million people who commented during the primaries, 56% commented only once. The most active 1%, about 11,000 people, wrote 31% of all comments and the most active 10% wrote 65%. The Gini coefficient of comments per person was 0.69. The most prolific single commenter wrote 7,451 comments in three months, about 80 a

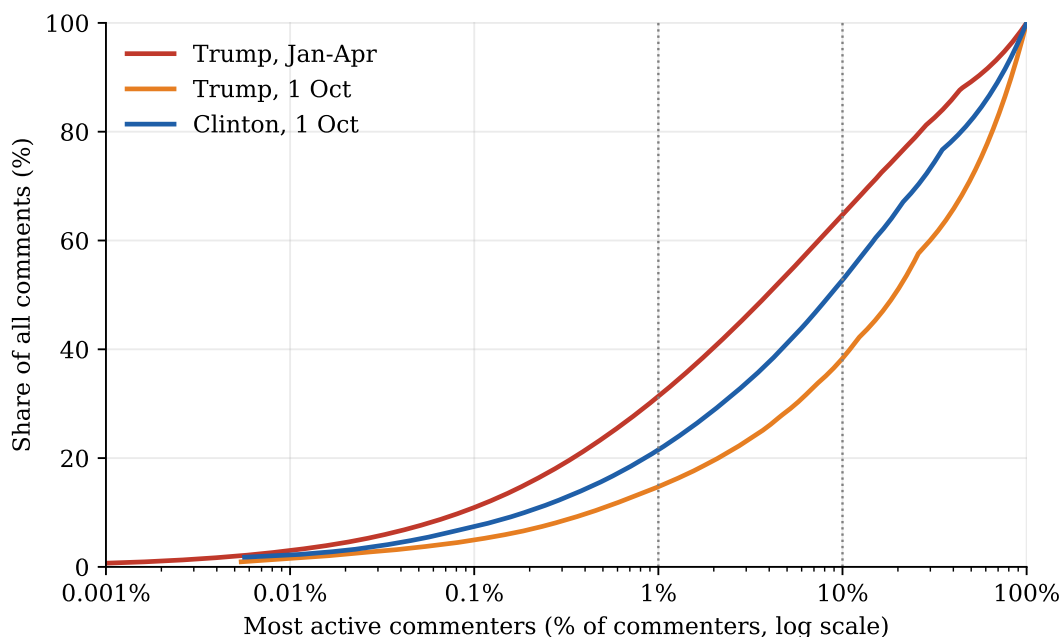


Figure 1. Concentration of comments. Each line shows the share of all comments written by the most active commenters. Dotted lines mark the top 1% and top 10%. The primary line covers three months and the 1 October lines cover a single day, so the levels are not directly comparable across them.

day.

Most people came once and left. Two-thirds of commenters (66%) commented in only one week of the fifteen, and only 8.5% commented in five weeks or more. The superparticipants, unsurprisingly given how much they wrote, commented in 11.3 weeks on average. The audience grew as the race went on (Figure 2), peaking at 624,000 comments in the week of Super Tuesday. The share of each week's commenters who had not commented before in the data fell from 85% in the first full week of January to under half by the second week of February and to 31% by the end of March. Part of this decline is mechanical, because people who commented before January cannot be seen. But the pattern is clear: a growing crowd formed around a stable core.

On 1 October both pages showed the same structure, though over a single day. The average commenter on Clinton's page wrote 2.8 comments against 1.8 on Trump's (the median was one on both), and Clinton's comment section was more concentrated: the top 1% wrote 21% of its comments against 15% on Trump's, and one person wrote 880 comments on her posts that day. Clinton's page was also more conversational. Replies to other commenters made up 40% of its comments against 25% on Trump's page that day and 21% during the primaries.

The post-level data for the last month of the campaign point the same way. Between 10 October and 6 November the median post on each page drew almost exactly the same number of comments (7,698 for Clinton and 7,708 for Trump). Trump's posts drew more reactions (a median of 68,397 against 27,304), so comments made up a larger share of the response on Clinton's page: 0.29 comments per reaction against 0.12.

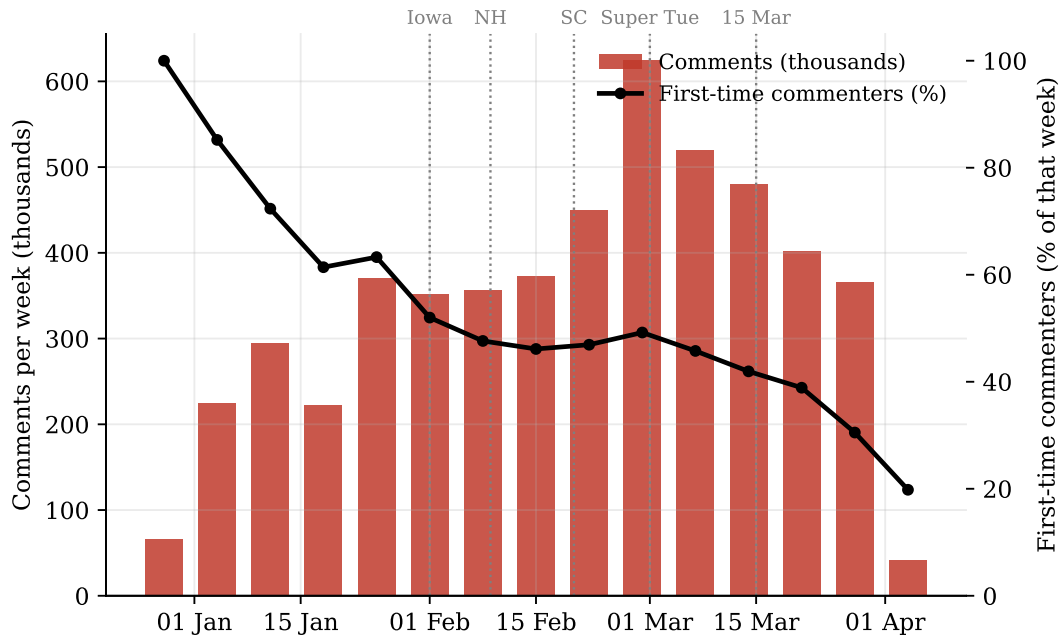


Figure 2. Comments per week on Trump’s page (bars) and the share of each week’s commenters appearing for the first time in the data (line), January to April 2016.

5.2 RQ2: Toxicity was common and steady

About one comment in ten on Trump’s page during the primaries was toxic: 10.0% (95% confidence interval 9.6% to 10.5%), weighted to all comments. The level depends on the threshold, from 13.0% at 0.3 to 7.2% at 0.7; Appendix Table 5 shows which results hold at each. Insults were the most common form (3.5% of comments), followed by obscenity (2.0%). Identity attacks (0.3%) and threats (0.2%) were rare, and severe toxicity was almost absent. With roughly 5.1 million comments, that still means about half a million toxic comments in three months.

Toxicity moved little over the primaries (Figure 3). The monthly rate was 10.1% in January, 9.3% in February, 10.6% in March and 9.2% in early April, and there was no trend (an estimated change of 0.1 percentage points per month, $p = 0.54$). Posts published on debate days had toxicity rates 1.6 percentage points higher than other posts in the same week (standard error 0.6, $p = 0.009$), and the difference held at both other thresholds of the classifier. The dictionary measure, however, shows no debate-day difference (0.1 points, not significant), so this result depends on the measure. Posts published on voting days had lower toxicity, by 1.1 points on the classifier ($p = 0.03$; not significant at the 0.7 threshold) and by 1.2 points on the dictionary ($p = 0.001$). Neither effect carried over to the next day, and the cancellation of the Chicago rally left no measurable trace.

Who the comments were about. What changed over the primaries was less the level of hostility than its subject. The Republican rivals were mentioned most in February and March, when Cruz, Rubio or Bush appeared in up to 14% of comments on some days, while Clinton came back into view from mid-March (Figure 4). Most comments, though, were about Trump or about no one on the list. Trump was mentioned in 43% of comments, his slogans in 8% and voting in 13%, against 6% for Cruz, 5% for the media, 4% for Clinton and 3% for Obama.

Toxicity varied by subject (Table 2). Comments mentioning Muslims, Islam, refugees or terrorism were the most often toxic (17%), and 4% were flagged as identity attacks, though this label is the one

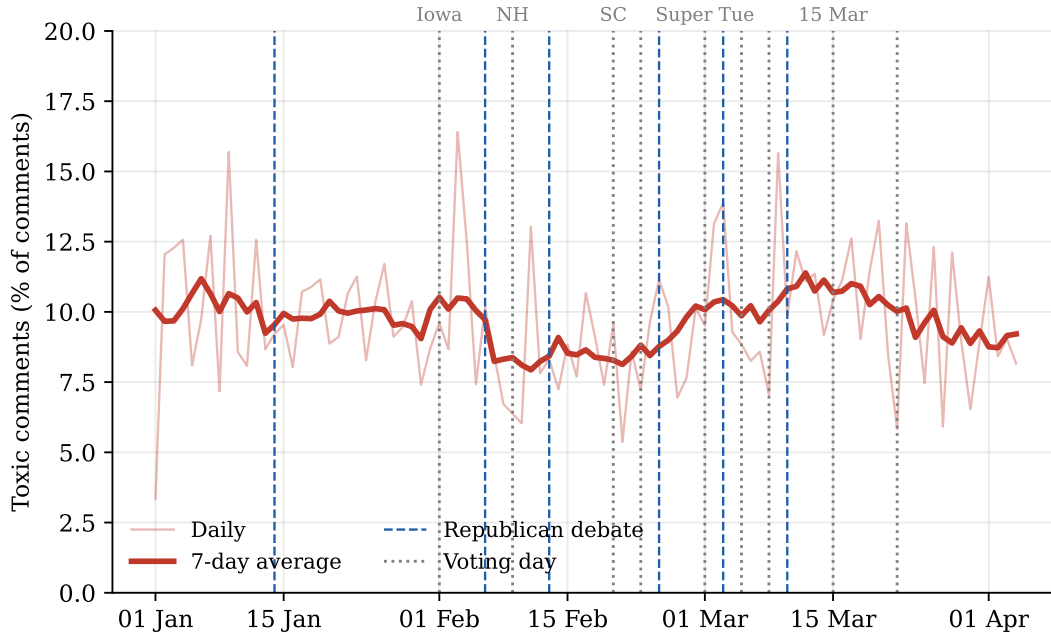


Figure 3. Share of toxic comments on Trump’s page by day of post, January to April 2016, with Republican debates in which Trump took part (dashed) and voting days (dotted). Each day’s rate is weighted to all comments on that day’s posts.

most likely to over-flag mentions of a group (Dixon et al. 2018). Comments mentioning Obama, Cruz, Clinton or Rubio were toxic about 12% of the time, close to comments that mentioned no one on the list. Comments using Trump’s slogans or mentioning God were the least often toxic.

Between commenters. Some of the hostility was aimed at other members of the audience rather than at politicians. Replies to other comments were more likely to be toxic than top-level comments (12.8% against 9.2%), and just over half (52%) of toxic comments named none of the politicians, groups or institutions on the list. Most toxic comments (72%) were nonetheless top-level comments, so the comment section’s hostility was not confined to arguments in reply threads.

5.3 RQ3: Toxicity was not rewarded

Toxic comments were not rewarded with likes (Table 3). Within the same post, and once the timing of comments is taken into account, a toxic comment was exactly as likely to receive at least one like as a non-toxic one (a difference of 0.0 percentage points, standard error 0.5, against 29% of primary comments that received any like). No band of the toxicity score differed from the least toxic comments (Figure 5). In the Poisson model of the number of likes, toxic comments received about half as many likes as others (rate ratio 0.53). That gap comes from the small number of comments with very many likes: when the most-liked 1% of comments are excluded, the estimate with the same controls is close to zero (rate ratio 0.99, standard error of the coefficient 0.03). The full corpus, using the dictionary, gives no sign that likes favoured or disfavoured hostile comments in a consistent way. The single most-liked comment on each post was profane or insulting in 5.6% of posts, against 6.4% of all comments, a difference within sampling error; the ten most-liked comments on each post were flagged 7.1% of the time and the hundred most-liked 8.1%; and comments with at least one like were flagged slightly more often than comments with none (7.4% against 6.0%), before any controls. In the Poisson model the insult,

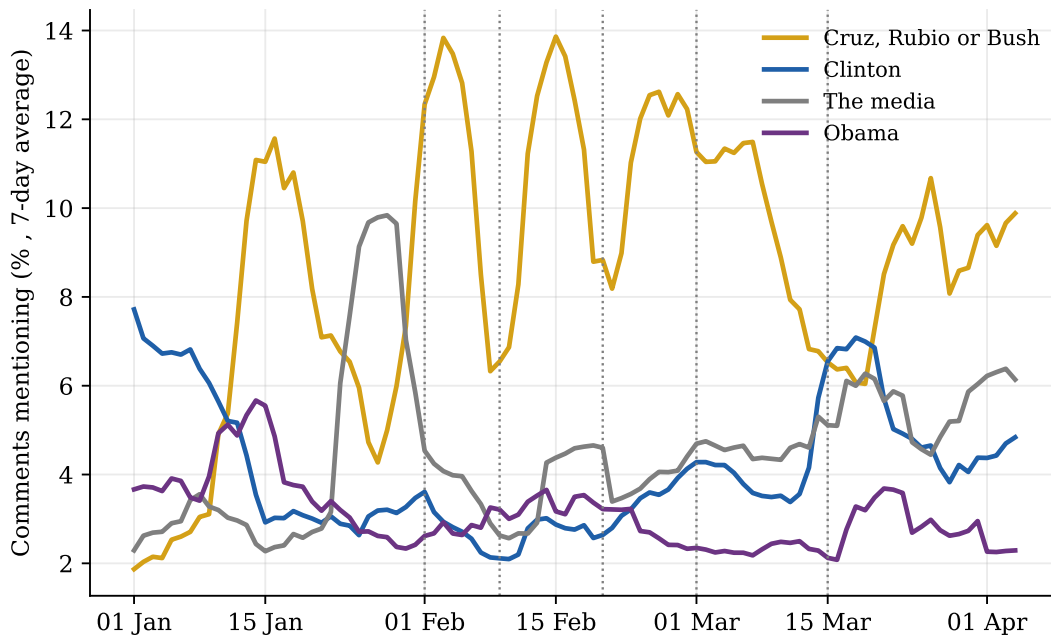


Figure 4. Share of comments on Trump’s page mentioning his Republican rivals, Clinton, the media and Obama, January to April 2016 (seven-day averages). Dotted lines mark the main voting days.

threat and identity attack labels and the dictionary measure point the same way as the main estimate; obscenity on its own made no difference. Whether a toxic comment mentioned Clinton, a Republican rival or the media made no clear difference either.

On 1 October toxic comments on Trump’s page again received fewer likes (rate ratio 0.42), while on Clinton’s page they received slightly more (1.32), a difference that was not significant. With five and eleven posts these estimates are imprecise, and I do not read much into the contrast.

5.4 RQ4: The regulars were not the most toxic

The self-selection account leads us to expect the most active commenters to be the most toxic. They were not (Figure 6). The share of toxic comments fell steadily with activity, from 12.0% among people who commented once to 8.6% among the superparticipants. Within the same post and controlling for length and replies, a superparticipant’s comment was 2.4 percentage points less likely to be toxic than anyone else’s (standard error 0.3, clustered by post and commenter). The result holds in a logit model, for the continuous score and at both other thresholds. On the full corpus, however, the dictionary shows no difference between superparticipants and others (6.4% each). The classifier and the dictionary may be picking up a difference in style, such as cheering and slogans, rather than in hostility. The safe conclusion is that the regulars were no more hostile than other commenters.

Because the regulars wrote so much, they still produced a large share of the hostility. Superparticipants wrote 27% of toxic comments (against 31% of all comments), and people with five or more comments wrote 69%. People who commented only once wrote 15%. Toxicity was spread across the whole crowd rather than concentrated at its core or at its edges.

Table 2. Toxicity by subject of comment, Trump’s page, primaries.

Comment mentions	Share of comments (%)	Toxic (%)	Insult (%)
Muslims, Islam, refugees or terrorism	1.6	17.1	3.6
Obama	3.0	12.5	3.2
Cruz (including “Ted”)	5.7	11.8	4.1
Clinton (including nicknames)	4.2	11.8	3.1
Rubio	2.3	11.6	3.2
Border, wall, immigration or Mexico	3.9	11.2	3.0
Sanders	1.5	10.6	3.0
Bush	1.2	9.6	2.5
Media, news, press or named outlets	4.8	9.0	2.1
Trump	42.8	7.6	2.3
Voting, polls or primaries	12.7	7.2	2.2
God or prayer	3.7	5.9	1.2
Trump’s slogans	8.3	4.5	1.0
None of the above	41.0	11.8	4.5
All comments	100	10.0	3.5

Weighted to all comments. A comment can mention more than one subject. Kasich and Carson, each mentioned in under 1% of comments, are omitted.

5.5 RQ5: Same toxicity, different behaviour by the people who crossed

On 1 October 2016 the two candidates’ comment sections were equally toxic (Table 4). Of comments on Trump’s posts, 13.3% were toxic (95% interval 11.3% to 15.0%), and of comments on Clinton’s, 13.4% (11.4% to 15.8%). A permutation test across the 16 posts finds no difference ($p = 0.97$), and the rates were also equal at the other thresholds. Insults were somewhat more common on Clinton’s page (5.1% against 3.8%). Both pages were more toxic that day than Trump’s page had been during the primaries. Trump’s page in the last ten days of the campaign was closer to its primary level (8.7%, from a sample of 1,680 comments).

The equal levels hide a difference among the people who took part in both conversations. On Clinton’s page, comments that mentioned Trump were more often toxic (15.0%) than comments that mentioned Clinton (10.8%). And 1,106 people commented on both candidates’ pages that day. They wrote 18% of the comments on Clinton’s page but 11% on Trump’s; the average person among them wrote 8.0 comments on Clinton’s posts against 3.2 on Trump’s. Comparing each of these people with themselves, their comments on Clinton’s page were 3.8 percentage points more likely to be toxic than their comments on Trump’s (standard error 0.8, $p < 0.001$). Averaging each person’s own difference, so that heavy commenters do not dominate, gives a similar 3.5 points (95% interval 1.5 to 5.4), and the difference holds at both other thresholds and when people with more than 30 comments are excluded (Figure 7). People who commented on only one page were about equally toxic on the two pages (13.7% on Trump’s and 13.3% on Clinton’s). Set against them, the people who crossed were less toxic than other commenters on Trump’s page (10.4% against 13.7%) and close to other commenters on Clinton’s (14.0% against 13.3%). The finding can therefore be described in two ways: people who crossed were calmer on Trump’s page, or sharper on Clinton’s. These comparisons are between different people, so they cannot separate the effect of the page from the effect of the person.

I cannot see whom these people supported. One independent piece of information is their history. Of the 1,106, 608 had commented on Trump’s page during the primaries. They were more toxic on

Table 3. Toxicity and likes, Trump’s page, primaries. Models with post fixed effects.

Model	Outcome	Estimate	SE	Interpretation
<i>With controls for comment timing and age at extraction</i>				
Linear	Any like (0/1)	0.000	0.005	No difference
Poisson	Likes	−0.64	0.20	Rate ratio 0.53
<i>Without timing controls</i>				
Linear	Any like (0/1)	−0.014	0.005	−1.4 points
Poisson, all comments	Likes	−0.76	0.26	Rate ratio 0.47
Poisson, top-level comments	Likes	−1.07	0.40	Rate ratio 0.34
Poisson, replies	Likes	0.03	0.04	Rate ratio 1.04
Poisson, excluding top 1% of likes	Likes	−0.05	0.03	Rate ratio 0.95
Poisson, insult label	Likes	−1.05	0.15	Rate ratio 0.35
Poisson, identity attack label	Likes	−0.65	0.27	Rate ratio 0.52
Poisson, threat label	Likes	−1.82	0.34	Rate ratio 0.16
Poisson, obscenity label	Likes	−0.05	0.66	Rate ratio 0.95
Poisson, dictionary measure	Likes	−0.71	0.27	Rate ratio 0.49

$n = 87,600$ comments on 730 posts (66,524 top-level; 21,076 replies). All models control for log comment length and, where both are included, whether the comment was a reply. Timing controls are the log of minutes between post and comment and the log of days between comment and extraction. Standard errors clustered by post.

Clinton’s page than on Trump’s (15.3% against 11.5%; within-person difference 4.7 points, standard error 1.2). The 498 without that history showed a difference in the same direction (11.8% against 8.7%; 2.7 points, standard error 1.2). The two groups do not differ significantly, and averaging each person’s own difference makes them nearly identical (3.8 and 3.1 points). Having commented on Trump’s page earlier does not in any case show that someone supported him. The dictionary measure shows the same direction for the people with a primary history (9.0% against 8.0%) and the reverse for the others (6.4% against 6.9%), so the size of these differences depends on the measure. What the evidence supports is that the people who crossed between the pages were, on balance, more toxic on Clinton’s page than on Trump’s, whatever their history.

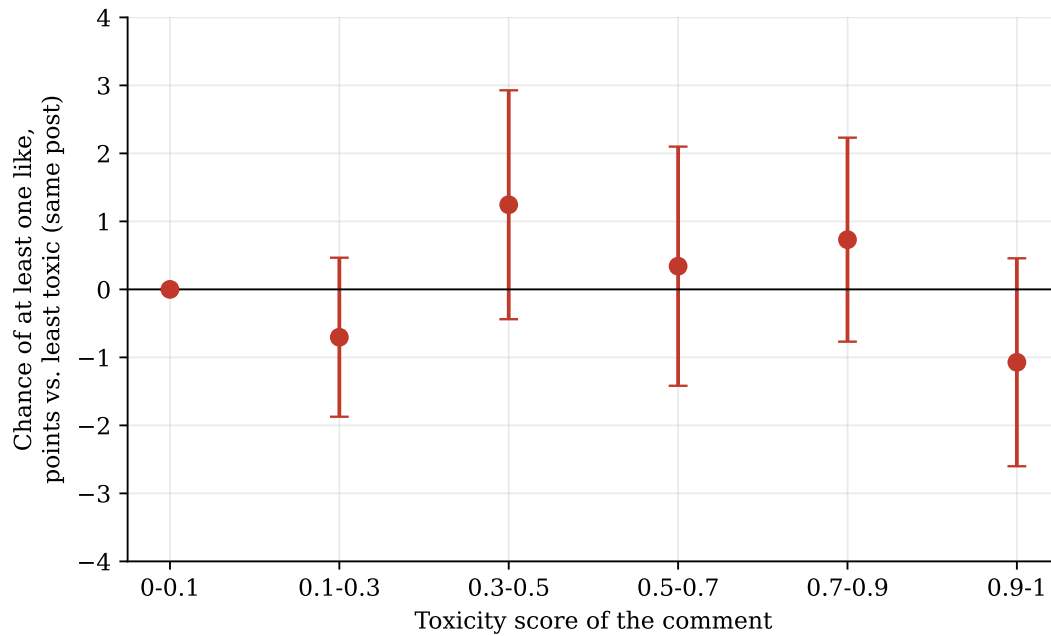


Figure 5. Chance that a comment received at least one like, by toxicity score, compared with the least toxic comments on the same post (percentage points, 95% confidence intervals), controlling for length, replies and comment timing. Trump’s page, primaries.

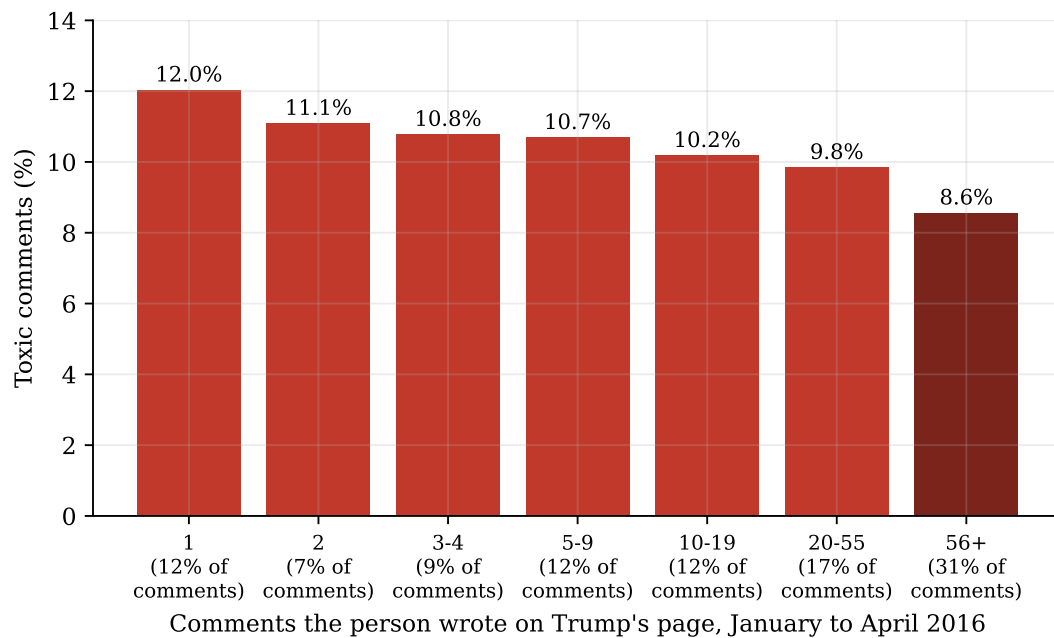


Figure 6. Share of toxic comments by how many comments the person wrote on Trump’s page during the primaries. Labels show each group’s share of all comments.

Table 4. Trump's and Clinton's comment sections on 1 October 2016.

	Trump's page	Clinton's page
Comments (posts)	32,301 (5)	49,421 (11)
Scored sample	8,819	13,826
Toxic (%)	13.3	13.4
Insult (%)	3.8	5.1
Replies (% of comments)	25.2	39.7
Mention Clinton (%)	24.5	34.0
of which toxic (%)	12.3	10.8
Mention Trump (%)	33.4	20.9
of which toxic (%)	12.9	15.0
Toxic, people who commented only on this page (%)	13.7	13.3
Comments by people who commented on both pages (%)	10.9	18.0

Weighted to all comments on that day's posts. Replies are from the full corpus.

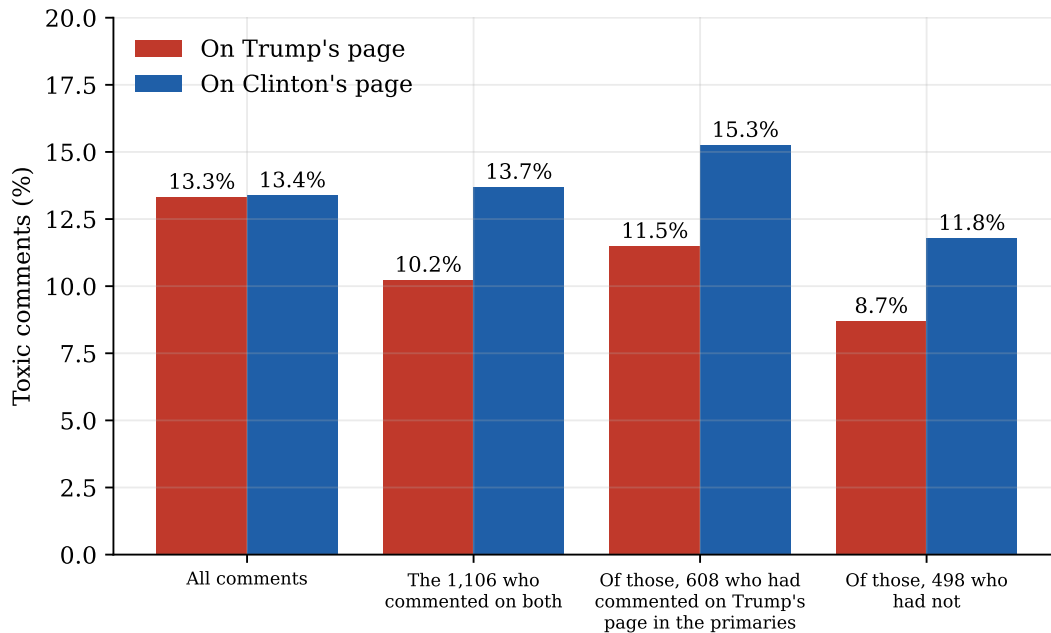


Figure 7. Toxic comments on each candidate's page on 1 October 2016: all comments (weighted), the 1,106 people who commented on both pages, and those people split by whether they had commented on Trump's page during the primaries. The cross-page figures average each person's own rate. The two subgroups do not differ significantly from each other.

6 Discussion

This paper gives a large-scale picture of who spoke, and how, in the comment sections of the two 2016 presidential candidates. Four findings stand out.

First, a candidate's comment section is a crowd with a core. A small group of regulars wrote a third of everything and kept coming back, while most people commented once. This is the superparticipant pattern found in forums ([Graham and Wright 2014](#)), at a far larger scale. Readers did not necessarily see all comments equally, since the order in which comments were displayed could vary, but the regulars were a large part of what there was to see.

Second, those regulars were not more hostile than anyone else. The self-selection account holds that commenters as a whole are more hostile than the public ([Kim et al. 2021](#)), and that may well be true. Within the comment section, though, more commenting did not mean more toxicity. One explanation is that regulars are invested in the space and their standing in it, while occasional commenters arrive in reaction to something and leave. Another is that regulars on a candidate's page are mostly supporters, and supportive talk is mostly civil. Work on antisocial behaviour online suggests a third: users who behave worst are often banned or leave ([Cheng et al. 2015](#)), so they appear in the data as occasional commenters. Whatever the reason, hostility was spread across the whole crowd.

Third, toxicity was not rewarded with likes on Trump's page. This differs from evidence that outrage and out-group animosity attract engagement ([Brady et al. 2021](#); [Rathje et al. 2021](#)) and from the finding that more toxic comments on news pages attract more likes ([Kim et al. 2021](#)). The difference may lie in who does the liking. On a candidate's own page, the people liking comments are mostly the candidate's supporters, who may have had less appetite for hostility than readers of news pages. Toxic comments were as likely as others to receive a like; the very few comments that collected large numbers of likes were less often toxic; and the full-corpus dictionary measure shows no consistent pattern either way. On this evidence, likes neither clearly rewarded nor clearly punished hostility.

Fourth, the two candidates' comment sections were equally toxic on the same day. This cuts against the idea that a candidate's own style sets the level of hostility in their comment section, at least on this evidence. What differed was the behaviour of the people who took part in both conversations: they were more toxic on Clinton's page than on Trump's. This fits work showing that incivility rises in exchanges across group lines ([Rains et al. 2017](#); [Muddiman and Stroud 2017](#); [Rossini 2022](#)). It also fits work showing that the same people behave differently in different settings ([Cheng et al. 2017](#)). The page, and who one meets there, matters as well as who one is.

For campaigns, the comment section is a public face they do not fully control ([Stromer-Galley 2000](#)). These findings suggest that its tone depends less on the candidate's style or on the most loyal followers than on the wider crowd, including people who arrive from elsewhere. For platforms, they suggest that likes did not, on these pages, consistently favour hostile comments (on moderation and ranking more broadly, see [Gillespie 2018](#)).

7 Limitations

The study has several limits. The most important is survival. Netvizz could not retrieve every comment Facebook counted, and the comments that were removed by the campaign, by Facebook or by their authors before extraction are missing. Most primary posts were extracted months after publication, so

there was plenty of time for removal. If the most hostile comments were removed, these estimates understate what was written. If hostile users were banned, they appear as occasional commenters, which could contribute to the finding that regulars were less toxic. If the two campaigns moderated differently, the equal toxicity on 1 October describes what survived on each page, not what was written. Likes accumulated over different periods in the two parts of the corpus and cannot be compared between them; within the primary corpus I control for each comment’s age at extraction. The primary corpus covers only Trump’s page, so RQ1 to RQ4 describe one candidate’s comment section, and the comparison of the two candidates covers a single day, chosen because it was collected for both pages. That day fell five days after the first debate, and the window included the evening a major news story about Trump’s taxes was published, so its levels may not be typical; the five Trump posts are not every post of that day. Detoxify was trained on Wikipedia comments, toxicity is narrower than incivility (Hede et al. 2021), and the classifier is known to over-flag some identity terms (Dixon et al. 2018; Sap et al. 2019). Agreement with the dictionary is high, but neither substitutes for human coding, and several results are weaker or absent on the dictionary: the debate-day rise, the lower toxicity of superparticipants, the size of the cross-page difference (which reverses for people without a primary history on Trump’s page) and the pattern of likes for the most-liked comments. Mentions are not targets, and the mention lists are broad. Finally, all comparisons are observational. Superparticipants differ from occasional commenters in many ways, and toxic comments differ from civil ones in more than toxicity.

8 Conclusion

The comment sections of the 2016 candidates were not the uniformly hostile spaces they are often imagined to be, nor were they civil. About one comment in ten was toxic, which on Trump’s page in the primaries meant about half a million comments. That hostility was steady rather than escalating, it was spread across the whole crowd rather than concentrated among the most active, and likes on Trump’s page did not consistently reward it. On the same day the two candidates’ comment sections were equally toxic, but the people who took part in both were more toxic on Clinton’s page than on Trump’s. Explaining hostility in political comment sections means looking beyond the candidate and the loudest regulars, at the whole crowd, at what survives moderation, and at where people choose to argue.

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 data supporting the figures and tables (counts and rates by week, day, subject, activity group and page) 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 5. Main results at three toxicity thresholds.

	0.3	0.5	0.7
Toxic, primaries (%)	13.0	10.0	7.2
Debate days (points)	+1.8*	+1.6**	+1.3**
Voting days (points)	−1.9**	−1.1*	−0.6
Any like, toxic vs. other (points)	+0.3	0.0	−0.1
Likes, Poisson coefficient	−0.30	−0.64**	−0.47*
Superparticipant vs. other (points)	−2.8***	−2.4***	−2.0***
Toxic on 1 Oct, Trump’s page (%)	17.6	13.3	9.0
Toxic on 1 Oct, Clinton’s page (%)	17.9	13.4	9.5
Cross-page, Clinton’s vs. Trump’s page (points)	+5.0***	+3.8***	+1.9**

Models as in the text. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

References

- Bor, A. and Petersen, M. B. (2022). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review*, 116(1):1–18.
- Brady, W. J., McLoughlin, K., Doan, T. N., and Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7(33):eabe5641.
- Cameron, A. C. and Trivedi, P. K. (2013). *Regression Analysis of Count Data*. Cambridge University Press, Cambridge, 2nd edition.
- 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*.
- Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2015). Antisocial behavior in online discussion communities. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*.
- 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.
- 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.
- 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.

- Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, New Haven.
- Graham, T. and Wright, S. (2014). Discursive equality and everyday talk online: The impact of “superparticipants”. *Journal of Computer-Mediated Communication*, 19(3):625–642.
- 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*.
- Hotham, T. (2021). *How Do Political Parties Use Facebook and What Does It Offer to Their Campaigns*. PhD thesis, University of Bath.
- Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22:129–146.
- 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.
- Kreiss, D. (2016). *Prototype Politics: Technology-Intensive Campaigning and the Data of Democracy*. Oxford University Press, New York.
- Muddiman, A. and Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, 67(4):586–609.
- Mutz, D. C. (2015). *In-Your-Face Politics: The Consequences of Uncivil Media*. Princeton University Press, Princeton.
- 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.
- Stromer-Galley, J. (2000). On-line interaction and why candidates avoid it. *Journal of Communication*, 50(4):111–132.
- Stromer-Galley, J. (2014). *Presidential Campaigning in the Internet Age*. Oxford University Press, New York.
- 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.
- Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.
- 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.