

One City, Two Comment Sections: Toxicity and Identity on Sadiq Khan's and Zac Goldsmith's Facebook Pages, 2016

Tristan Hotham*

2026

Preprint, not peer reviewed

Abstract

The 2016 London mayoral election was contested by Sadiq Khan, who became the city's first Muslim mayor, and Zac Goldsmith. Religion and extremism became themes of the campaign. I compare the comment sections of the two candidates' Facebook pages from January to mid-April 2016: 36,087 comments on 701 posts, after removing duplicates. Every comment was scored with an open toxicity classifier. The two comment sections were similarly toxic, at about 8%. Goldsmith's posts drew a median of 75 comments each against 13 on Khan's, from a more concentrated crowd. Comments the classifier labelled as identity attacks were rare on both pages, but more than twice as common on Khan's (0.55% against 0.23%). An AI-assisted reading of all 128 such comments, not yet checked by a second coder, suggests that only about half were hostile attacks on an identity group; many were people objecting to racism or comments not about identity at all. On that reading, comments hostile to Muslims or Islam numbered 25 on Khan's page and 15 on Goldsmith's, a rate of 0.18% against 0.07%. Among comments that mentioned Muslims or Islam, a quarter on Khan's page were toxic, against a sixth on Goldsmith's. Comments mentioning Jewish people showed no clear difference, and if anything the reverse. Mentions of religion rose on both pages as the campaign went on. The 593 people who commented on both pages were similarly toxic on each. Overall toxicity rates can hide differences in the kind of hostility on a page, and automated identity-attack labels need to be read before they are counted.

Keywords: incivility; identity attacks; anti-Muslim hostility; comment sections; Facebook; London mayoral election; toxicity classifiers

1 Introduction

On 5 May 2016 Sadiq Khan was elected Mayor of London, defeating the Conservative candidate Zac Goldsmith. The campaign was mostly about housing, transport and the cost of living. But religion and extremism also became themes, and both candidates addressed them on their Facebook pages. Khan's page carried posts on extremism and hate crime and, on 11 April, a post responding to what he described as his opponent's repeated references to his faith. Goldsmith's page carried a post on 13 April arguing that accusations of Islamophobia were being used to close down legitimate questions. Both pages also

*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.

carried posts condemning antisemitism and recording visits to faith communities. The best-known controversy over Goldsmith's campaign, a newspaper article published on 1 May, falls after the period studied here, and these data say nothing about it.

This paper compares what happened beneath the two candidates' posts. In two earlier preprints I studied the comment sections of Donald Trump's and Hillary Clinton's Facebook pages in 2016 (Hotham 2026b,a). Those studies were limited because comments on Clinton's page were available for a single day. The London data have no such imbalance: I have comments on both candidates' posts across the same three and a half months, collected at the same time.

Research on abuse of politicians online finds that hostility is unevenly spread. In the UK, abuse of MPs and candidates varies with party, prominence and gender (Ward and McLoughlin 2020; Gorrell et al. 2020; Southern and Harmer 2021; Akhtar and Morrison 2019; Collignon and Rüdiger 2020), and Muslims are frequent targets of hostility on social media (Awan 2016; Awan and Zempi 2016; Evolvi 2019). Most of this work studies replies to politicians on Twitter. Less is known about the comment sections of candidates' own Facebook pages, where supporters, opponents and the undecided meet.

I ask three questions. **RQ1.** Who took part in each candidate's comment section? **RQ2.** Were the two comment sections similarly toxic, and did the kind of hostility differ? **RQ3.** How did people who commented on both pages behave on each?

2 Background

Incivility and identity. Incivility is usually defined as communication that breaks norms of respect, such as insults, name-calling and vulgarity (Coe et al. 2014). Rossini (2022) distinguishes incivility from intolerance, which attacks people for who they are. The distinction matters here: two comment sections can be similarly uncivil while differing in intolerance. Studies of abuse of UK politicians show that its content varies with the identity of its target: women MPs, for example, receive more gendered abuse (Southern and Harmer 2021). Separately, studies of social media find that Muslims are often targeted with hostility that draws on their religion (Awan 2016; Awan and Zempi 2016).

Measuring identity attacks. Automated classifiers make measurement at scale possible, but they are known to confuse offensive language with hate speech (Davidson et al. 2017) and to over-flag comments that merely mention identity groups (Dixon et al. 2018; Sap et al. 2019). Toxicity measures also do not map exactly onto the concepts communication scholars use (Hede et al. 2021). A comment objecting to racism can be flagged as an identity attack. For a study in which religion is central, the labels need to be checked by reading.

Comment sections. Candidates' comment sections are dominated by small groups of active participants (Graham and Wright 2014; Hotham 2026b), and hostility often runs across group lines (Rains et al. 2017; Muddiman and Stroud 2017; Theocharis et al. 2016). On Trump's and Clinton's pages, the two comment sections were equally toxic on the same day, and people who commented on both were more toxic on Clinton's (Hotham 2026b).

3 Data and measures

Data. The data were collected with Netvizz (Rieder 2013) on 14 April 2016 as part of my doctoral research (Hotham 2021). They cover every post published by each candidate's page between 1 January

and 13 April 2016, and the comments that could be retrieved beneath them. Khan’s page published 505 posts, 490 of which had comments (484 after removing duplicates, as described below); Goldsmith’s published 217, all with comments. Posts by members of the public on Khan’s page are excluded. The median post was 48 days old on Khan’s page and 49 days old on Goldsmith’s when the data were collected. The tool retrieved 95% of the comments Facebook counted on Khan’s posts and 87% on Goldsmith’s; the difference may reflect comments hidden or deleted by page administrators or their authors, which could affect the comparison in either direction. I removed 1,256 exact duplicates (the same person posting the same text on the same page, whether on the same post or across posts, which removes repeated spam and repeated short replies), leaving 13,906 comments on Khan’s page from 5,983 commenters and 22,181 on Goldsmith’s from 8,436. Each commenter appears as a one-way hash of their Facebook identifier that is the same on both pages, so I can identify, anonymously, the 593 people who commented on both. The data end three weeks before polling day.

Toxicity and identity attacks. I score every comment with Detoxify (original model; [Hanu and Unitary team 2020](#)), an open classifier trained on Wikipedia talk page comments labelled by crowd workers ([Jigsaw and Conversation AI 2018](#)), truncating comments at 96 word-pieces. I count a comment as toxic if its toxicity score is 0.5 or more, and report the classifier’s identity attack, insult and obscenity labels. As a check on general toxicity, I apply a dictionary of profanity and insults, including common British insults; the classifier separates dictionary-flagged comments from others with an area under the curve of 0.93. The dictionary contains no identity terms, so it does not check the identity attack label.

Reading the identity attack labels. Because classifiers can over-flag comments that mention religious groups ([Dixon et al. 2018](#); [Sap et al. 2019](#)), I read all 128 comments labelled as identity attacks and coded each by what it did. The categories were hostility towards Muslims or Islam; antisemitism; hostility towards white or English people; hostility towards other groups (such as other ethnic groups, foreigners in general, or gay or disabled people); counter-speech, meaning comments objecting to racism or defending a group; and comments that were not about identity. This reading was carried out with the assistance of an AI model (Claude, by Anthropic) and has not been checked by a second coder, so it should be treated as a careful first pass rather than a validated coding. No comment is quoted.

Mentions. I flag comments that mention Khan, Goldsmith, the parties, religion or faith groups, Muslims or Islam, Jewish people, and class (such as “posh”, “Eton” or “council estate”).

Analysis. I compare rates between pages with bootstrap confidence intervals that resample posts, and with regressions that control for comment length, replies and week, with standard errors clustered by post. For people who commented on both pages I compare each person with themselves.

4 Findings

4.1 RQ1: Two different crowds

The two pages drew very different crowds (Table 1). Khan’s page posted more than twice as often, but Goldsmith’s posts drew far more comments each: a median of 75, against 13. Goldsmith’s comment section was more concentrated: the most active 1% of commenters wrote 21% of its comments, against 17% on Khan’s. More of its comments were replies to other commenters (38% against 31%). On both pages about two-thirds of commenters wrote only once.

Table 1. The two comment sections, 1 January to 13 April 2016.

	Khan's page	Goldsmith's page
Candidate's posts (with comments after removing duplicates)	505 (484)	217 (217)
Comments, after removing duplicates	13,906	22,181
Commenters	5,983	8,436
Median comments per post	13	75
Share written by the most active 1% (%)	17	21
Replies to other commenters (%)	31	38
Toxic (%)	7.9	8.3
Insult (%)	3.0	3.0
Obscene (%)	3.2	1.5
Labelled identity attack (%)	0.55	0.23
of which, on reading, hostile to an identity group	40 of 76	23 of 52
hostile to Muslims or Islam	25	15
antisemitic	1	3
hostile to white or English people	4	0
hostile to other groups	10	5
counter-speech	19	15
not about identity	17	14
Mention Muslims or Islam (%)	3.0	2.1
Mention Jewish people (%)	1.0	1.1

Comments on the candidates' own posts. Labels from Detoxify at a threshold of 0.5. The reading of identity attack labels is described in the text.

4.2 RQ2: Similar toxicity, different hostility

Overall, the two comment sections were similarly toxic: 7.9% of comments on Khan's page and 8.3% on Goldsmith's.¹ Controlling for length, replies and week, the difference was -0.4 points (95% interval -1.6 to 0.9). On each page, comments that mentioned the opposing candidate were more often toxic than comments that mentioned the page's own candidate.

Replies. About two in five toxic comments were replies on both pages (43% and 42%). On Khan's page the gap between replies and top-level comments was larger: replies were toxic 11.0% of the time and top-level comments 6.6%, against 9.1% and 7.8% on Goldsmith's (Figure 1).

Identity attacks. Comments the classifier labelled as identity attacks were rare on both pages but more than twice as common on Khan's: 76 comments (0.55%) against 52 (0.23%), a difference of 0.3 points (bootstrap 95% interval 0.1 to 0.5). These comments came from 61 and 48 different people. Reading them, with the caveats described in the measures section, changes the picture in two ways (Table 1). First, only about half were hostile attacks on an identity group: 40 of 76 on Khan's page and 23 of 52 on Goldsmith's. Many of the rest were counter-speech, people objecting to racism or defending a group (19 and 15), or were not about identity at all (17 and 14). The label overstates identity-based hostility on both pages. Second, the hostility that remained took different forms on the two pages. On Khan's page, most of it was directed at Muslims or Islam (25 comments), but it also included hostility towards white or English people (4) and towards other groups (10). On Goldsmith's page, hostility towards Muslims or Islam was also the largest category (15 comments), along with antisemitic comments (3).

¹Without removing duplicates, the figures are 7.8% and 8.3% for toxicity and 76 and 59 comments labelled as identity attacks.

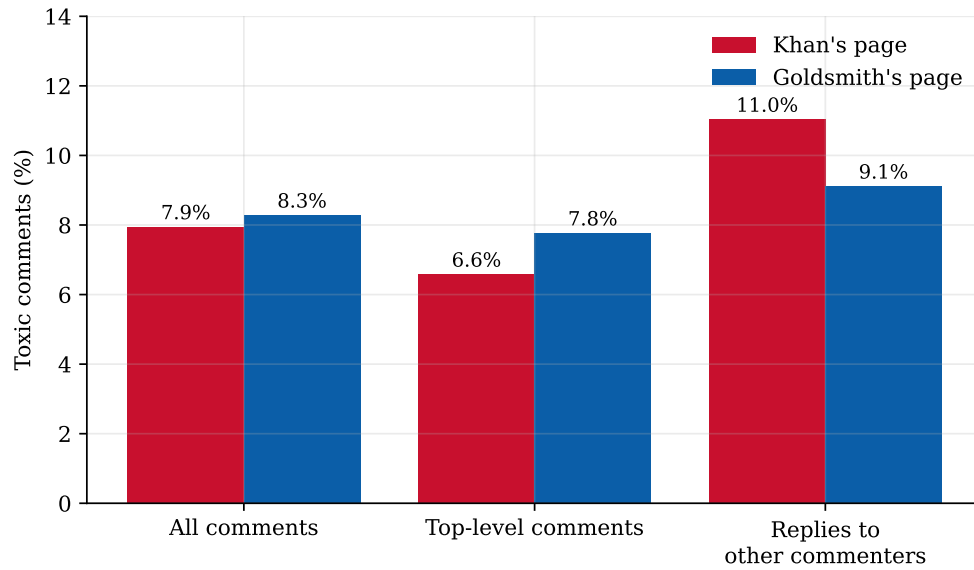


Figure 1. Share of toxic comments on each candidate’s page, for all comments, top-level comments and replies, January to April 2016.

Hostility towards Muslims or Islam was therefore present on both pages, and more frequent on Khan’s: 0.18% of comments against 0.07% (bootstrap 95% interval for the difference 0.02 to 0.21 points). These are small numbers. Few comments on either page were hostile on grounds of religion, and some of what the classifier flagged was the opposite: people defending Muslims or Jewish people.

When religion was mentioned. The difference is clearer among comments that mentioned Muslims or Islam, which made up 3.0% of comments on Khan’s page and 2.1% on Goldsmith’s (Figure 2). On Khan’s page, 26% of these were toxic, against 17% on Goldsmith’s (a difference of 9.3 points, $p = 0.002$). Comparing comments that mention the same groups holds constant some, though not all, of the classifier’s tendency to over-flag religious terms. Comments that mentioned Jewish people pointed the other way, though the difference was not significant: 15% were toxic on Khan’s page and 20% on Goldsmith’s, $p = 0.35$), and identity attack labels were similar (3.5% and 4.1%). Hostility in comments about religion did not run in one direction only.

Over time. Religion became a more frequent subject on both pages as the campaign went on (Figure 3). On Khan’s page, the share of comments mentioning religion rose from 1.8% in January to 5.7% in the first half of April; on Goldsmith’s page, from 2.6% to 4.1%. Part of the April figure on each page comes from a single post: Khan’s 11 April post and Goldsmith’s 13 April post each drew comments of which 15% mentioned religion. Without them, the April shares were 5.0% and 3.0%.

Who wrote the toxic comments. On Khan’s page there was no clear relationship between how often someone commented and how toxic their comments were. On Goldsmith’s page, more frequent commenters were more toxic: the most active 1% were toxic 10.5% of the time against 8.2% for people who commented once, and the relationship was clear when comparing comments on the same post ($p = 0.002$).

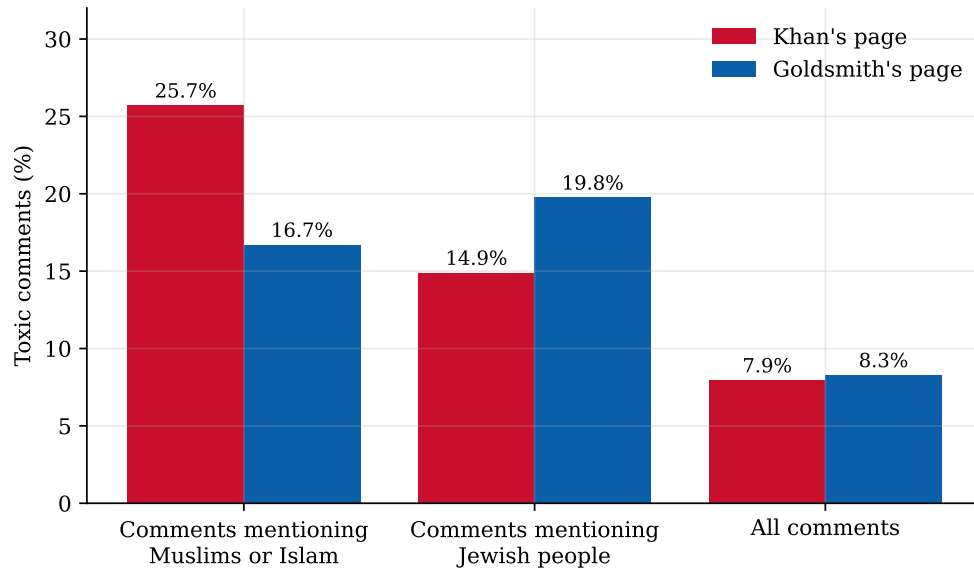


Figure 2. Share of toxic comments among comments mentioning Muslims or Islam, among comments mentioning Jewish people, and among all comments, on each candidate's page.

4.3 RQ3: People who commented on both pages

The 593 people who commented on both pages wrote 19% of the comments on Khan's page and 18% on Goldsmith's. Averaging each person's own rate, they were toxic 8.2% of the time on Khan's page and 9.3% on Goldsmith's, a difference that is not statistically significant (-1.1 points, 95% interval -3.4 to 1.2). They wrote few of the comments read as hostile to an identity group (4 on Khan's page and 2 on Goldsmith's). This differs from the pattern on Trump's and Clinton's pages, where people who commented on both were more toxic on Clinton's ([Hotham 2026b](#)), though a non-significant difference in London is not evidence that the two campaigns differed.

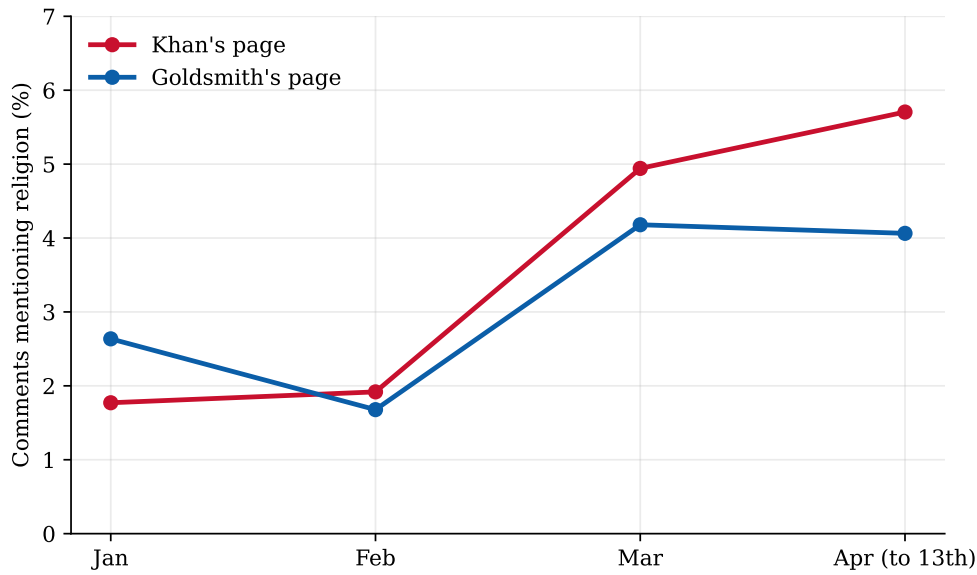


Figure 3. Share of comments mentioning religion or faith groups on each candidate’s page, by month in which the comment was written. April covers 1 to 13 April.

5 Discussion

The comment sections of the 2016 London mayoral race were, by the most common measure, similarly toxic. About one comment in twelve was toxic on each page, and people who commented on both behaved similarly on each.

Beneath that similarity there were differences in the kind of hostility, though they rest on small numbers. Comments hostile to Muslims or Islam appeared on both pages and were more frequent on Khan’s, and comments that mentioned Muslims or Islam were more often toxic there. Comments that mentioned Jewish people showed no such difference, and antisemitic comments were few on both pages. Khan’s page also hosted some hostility towards white or English people. This is consistent with research showing that the content of hostility follows the identity of those involved ([Southern and Harmer 2021](#); [Awan 2016](#)) and with the distinction between incivility and intolerance ([Rossini 2022](#)). The data cannot show at whom each comment was aimed, or whether it was written by a supporter or an opponent of either candidate.

The second lesson is about measurement. Half of the comments the classifier labelled as identity attacks were not hostile attacks on an identity group. Many were people objecting to racism. A study that counted the labels without reading them would have overstated identity-based hostility on both pages and mistaken some counter-speech for abuse. This is consistent with known weaknesses of these classifiers ([Davidson et al. 2017](#); [Dixon et al. 2018](#); [Sap et al. 2019](#)), and it suggests that studies of identity-based abuse of politicians should read, or at least sample, what their classifiers flag.

Finally, religion became a more frequent subject in both comment sections as the campaign went on, and the rise in April owed something to single posts on each page in which the candidates addressed religion directly. The data are consistent with comment sections following the topics the candidates raised, but they cannot show that the candidates’ posts caused the hostility that followed.

6 Limitations

The data end on 13 April 2016, three weeks before polling day, and miss the final stage of the campaign. The classifier was trained on Wikipedia comments and will miss abuse in coded language; comparing comments that mention the same groups reduces but does not remove its bias. The reading of identity attack labels was AI-assisted, was done by a single coder, and has not yet been checked by a human coder, which matters because borderline cases can change these small counts; it also covers only flagged comments, so hostile comments the classifier missed are not counted. Identity-based hostility is rare, so the differences rest on tens of comments, some of them from the same people. Retrieval differed between the pages (95% and 87%), and comments hidden or deleted before collection are missing, which could bias the comparison in either direction. The analysis is observational, and differences between the pages may reflect differences in who visited each page as well as in what the candidates posted.

7 Conclusion

Sadiq Khan’s and Zac Goldsmith’s Facebook comment sections were similarly toxic overall. Hostility towards Muslims or Islam appeared on both pages and was more frequent on Khan’s, while comments about Jewish people showed no such difference. Half of what an automated classifier labelled as identity attacks turned out, on reading, to be something else, often people objecting to prejudice. Overall toxicity rates can hide differences in the kind of hostility on a page, and automated identity labels need to be read before they are counted.

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.

Ethics and data availability

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 and no comment is quoted. The reading of flagged comments used an AI model within the author’s own analysis environment, and comment text is not shared beyond that. Aggregate data supporting the tables and figures are available with this preprint. Comment text and commenter codes are not shared.

References

Akhtar, S. and Morrison, C. M. (2019). The prevalence and impact of online trolling of UK members of parliament. *Computers in Human Behavior*, 99:322–327.

- Awan, I. (2016). Islamophobia on social media: A qualitative analysis of the Facebook's walls of hate. *International Journal of Cyber Criminology*, 10(1):1–20.
- Awan, I. and Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and Violent Behavior*, 27:1–8.
- 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.
- Collignon, S. and Rüdiger, W. (2020). Harassment and intimidation of parliamentary candidates in the United Kingdom. *The Political Quarterly*, 91(2):422–429.
- Davidson, T., Warmley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*.
- 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.
- Evolvi, G. (2019). #Islamexit: Inter-group antagonism on Twitter. *Information, Communication & Society*, 22(3):386–401.
- Gorrell, G., Bakir, M. E., Roberts, I., Greenwood, M. A., and Bontcheva, K. (2020). Which politicians receive abuse? Four factors illuminated in the UK general election 2019. *EPJ Data Science*, 9:18.
- 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.
- Hotham, T. (2026a). Setting the tone or supplying the words? Donald Trump's Facebook posts and the hostility of his comment section in the 2016 Republican primaries. Preprint, Zenodo, October 2026, <https://doi.org/10.5281/zenodo.23288125>.
- Hotham, T. (2026b). 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>.
- Muddiman, A. and Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, 67(4):586–609.

- 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.
- 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.
- Southern, R. and Harmer, E. (2021). Twitter, incivility and “everyday” gendered othering: An analysis of tweets sent to UK members of parliament. *Social Science Computer Review*, 39(2):259–275.
- 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.
- Ward, S. and McLoughlin, L. (2020). Turds, traitors and tossers: The abuse of UK MPs via Twitter. *The Journal of Legislative Studies*, 26(1):47–73.