Political Science Research and Methods (2026), **14**, 1540–1548
doi:10.1017/psrm.2025.10063

RESEARCH NOTE

The distribution of hate speech and its implications for content moderation

Gloria Gennaro[1], Laura Bronner[2], Laurenz Derksen[2], Maël Kubli[3], Ana Kotarcic[3], Selina Kurer[2], Philip Grech[2], Karsten Donnay[3], Fabrizio Gilardi[3] and Dominik Hangartner[2]

[1]Department of Political Science, University College London, United Kingdom; [2]Public Policy Group, ETH Zurich, Switzerland and [3]Department of Political Science, University of Zurich, Zurich, Switzerland
Corresponding author: Gloria Gennaro; Email: g.gennaro@ucl.ac.uk

(Received 19 July 2024; revised 30 April 2025; accepted 4 July 2025; first published online 22 December 2025)

Abstract

Hate speech is widely seen as a significant obstacle to constructive online discourse, but the most effective strategies to mitigate its effects remain unclear. We claim that understanding its distribution across users is key to developing and evaluating effective content moderation strategies. We address this missing link by first examining the distribution of hate speech in five original datasets that collect user-generated posts across multiple platforms (social media and online newspapers) and countries (Switzerland and the United States). Across these diverse samples, the vast majority of hate speech is produced by a small fraction of users. Second, results from a pre-registered field experiment on Twitter indicate that counterspeech strategies obtain only small reductions of future hate speech, mainly because this approach proves ineffective against the most prolific contributors of hate. These findings suggest that complementary content moderation strategies may be necessary to effectively address the problem.

Keywords: comparative politics; field experiments; mass media and political communication; political behavior; public policy; text and content analysis

Online platforms such as social media, forums, and newspapers comment sections are crucial spaces for democratic debate and engagement. Yet fostering constructive discourse within these digital arenas remains a substantial challenge. Hate speech—derogatory language targeting individuals based on race, religion, gender, or other characteristics (United Nations, 2020)—is a primary obstacle in this context (Siegel, 2020), that fosters hostility and aggression, and inflicts serious psychological harm (e.g. Cao *et al.*, 2023; Müller and Schwarz, 2023).

While there is a broad consensus that online hate speech is a significant problem, solutions remain contested. Large digital platforms employ automated systems and human moderators to remove hate speech. While such measures can be effective, they risk of misclassifying legitimate speech and raise concerns about censorship (Douek, 2021; Pradel *et al.*, 2024). In this context, counterspeech—responding to hate speech messages by encouraging a more constructive and positive discourse—emerges as a promising strategy used by civil society and NGOs (Siegel, 2020; Hangartner *et al.*, 2021; Yildirim *et al.*, 2023; Gennaro *et al.*, 2025). When successful, counterspeech challenges hate without restricting expression, and exposes bystanders to diverse viewpoints. However, this strategy assumes that perpetrators are open to change, which may not always hold true.

© The Author(s), 2025. Published by Cambridge University Press on behalf of EPS Academic Ltd. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psrm.2025.10063 Published online by Cambridge University Press

Regulations against hate speech navigate those trade-off, as they aim to balance the dual imperatives of protecting individuals and communities from harm and preserving the fundamental right to free speech (Douek, 2021).

This paper argues that understanding the distribution of hate speech across users is a crucial element to gauge the magnitude of those trade-offs. We address this missing link in two stages. First, we analyze four original datasets, including over 55 million Swiss tweets, and 5.8 million comments posted on online news media in Switzerland in 2021. Using validated machine learning classifiers for the Swiss context (Kotarcic *et al.*, 2022), we find that a small percentage of users are responsible for the majority of hate speech. We further observe this general pattern in an additional sample of U.S. Twitter users, suggesting that those descriptive findings hold across platforms and linguistic contexts.

Second, we report findings from a field experiment conducted on the Swiss Twittersphere, designed to assess the impact of various counterspeech strategies on hate speech production. Preregistered analyses indicate that the interventions mitigated hate speech on average; yet, the effects are small. Additional exploratory analyses suggest that the primary reason for this overall moderate effectiveness is the resilience of the most frequent hate speech contributors to counterspeech efforts. For users who engage in hate speech less frequently than the sample median in the pre-treatment period, we find that counterspeech can decrease the likelihood of future hate speech in the 4-week follow-up period.[1]

The contribution of this study is twofold. First, our results emphasize that a small, concentrated group of determined users contributes the majority of hate speech, and for this group, broad counterspeech strategies might not be sufficient to mitigate the problem. Similar patterns were previously found in Covid-related tweets (He *et al.*, 2021), and our study extends this evidence across different platforms and linguistic contexts, demonstrating that the pattern holds both within and outside the U.S. This finding also parallels research on misinformation, which shows that a small number of accounts drive most conspiracy theories (Dozen, 2021). Evidence suggests that such behaviors are often driven by status-motivated individuals who differ from average social media users and are more visible in online spaces (ElSherief *et al.*, 2018; Bor and Petersen, 2022). Understanding who drives online hate is essential for designing effective interventions.

Second, this paper contributes to the emerging experimental literature on counterspeech strategies. Previous studies show that counterspeech is more effective when delivered by high-status users (Munger, 2017), when emphasizing shared religious identity (Siegel and Badaan, 2020), or when appealing to empathy (Hangartner *et al.*, 2021; Gennaro *et al.*, 2025). Our findings add nuance by showing that counterspeech appears effective primarily among occasional, but not prolific, hate speech users. For this group, complementary content moderation approaches should be devised and may potentially include deplatforming (Thomas and Wahedi, 2023).

1. Distribution of hate speech

1.1. Data and methods

Our study leverages several sources of data. First, we collect tweets from a sample of Swiss Twitter users that can be considered as interested in or attentive to politics and news, similar to Barberá *et al.* (2019), and who posted during 2021. To construct the sample, we began with compiling a list of all national parties and members of parliament, and a separate list of leading Swiss newspapers and journalists. Using each list as a starting point, we sampled users who followed a minimum of three accounts within the list. This resulting set of accounts comprises 96 591 unique users who tweeted in 2021. We collected all 56 026 528 tweets posted from those users in 2021, 76% of which were in German.

[1]The pre-analysis plan is available at https://osf.io/xvwgd/. We discuss minor deviations in the Supplementary material, Appendix section B.

https://doi.org/10.1017/psrm.2025.10063 Published online by Cambridge University Press

Second, we access all comments submitted by users in the comments sections of three major German-language Swiss newspapers in the year 2021.[2] Two of these newspapers, Newspapers 1 and 2, are tabloid-style papers with a large online presence; the third, Newspaper 3, is a broadsheet and has a smaller audience. All three are daily newspapers that serve all of German-speaking Switzerland with frequent online updates. This totals 5.8 million comments, which include both comments that were published eventually, and comments that were not published because they were subject to some form of content moderation. Having access to both published and unpublished comments is a rare feature of this pseudonymized and NDA-protected dataset. The newspaper comments were produced by 155 821 unique registered users.[3] Overall, 51.2% of comments are original comments, while 48.8% are replies to original comments.

We classify German-language tweets and newspaper comments with a BERT-based deep learning classifier tailored to the Swiss context (Kotarcic *et al.*, 2022). This classifier is trained to detect hate speech (intended as identity attacks) as well as toxic messages, and validated on the same type of data used in this paper, namely online newspaper comments (F1 = 0.80) and Tweets (F1 = 0.79). We follow the paper's suggestion to classify tweets as hate when the classifier probability surpasses the 0.85 threshold.

1.2. Results

We start by presenting the distribution of hate speech across users on different platforms. In our main Swiss Twitter sample in Figure 1 (top-left panel), we find that a small minority of users is responsible for the vast majority of hate speech. Using our main classifier, 1% of users are responsible for 46% of the hate speech produced, and 5% of users are responsible for 83% of hate speech in the Swiss sample.[4]

When interpreting those findings, it is important to consider that the Twitter data comes from a selected sample of users (see Section 1.1), and is likely affected by the platform's content moderation policies. To determine if our descriptive results are due to those features rather than genuine posting behavior, we turn to the media data which covers all comments—published and unpublished—submitted to the platform. Results, reported in the top-right and bottom panels in Figure 1, show the distribution of hate speech across users for the three online newspapers. These figures include comments that were intercepted by content moderators and were never published, thus providing a genuine picture of the production of hate speech across the entire population of submitted comments. The distribution of hate speech comments across users reveals a similar pattern to that of the Twitter sample, wherein a small minority of users is responsible for the majority of hate speech. Specifically, in Newspaper 1, 1% of users produce approximately 56% of hate speech, while in Newspaper 2 and Newspaper 3, this figure rises to 70% and 69%, respectively. Moreover, 5% of users are responsible for 87% of hate speech in Newspaper 1, and a complete 100% in both Newspaper 2 and Newspaper 3. The difference across the three newspapers can be due to many factors, including their different audiences: Newspaper 3 is a broadsheet with the smallest audience, while the other two engage in tabloid journalism and have a wider audience.

Despite the many differences between Twitter users and posters on online media, we find strikingly consistent patterns in the distribution of hate speech across samples. To understand whether the results generalize beyond the Swiss context, we further collected tweets from a U.S. Twitter user sample in March 2023, using data collection strategy similar to the one used in Switzerland.

[2]For one of these three papers (Newspaper 2 hereinafter), user registration was required only after May 2021 and we were not able to access all comments before July 2021. For these reasons, for Newspaper 2, we use only comments by registered users starting from July 1, 2021.

[3]More specifically, Newspaper 1 provided 2.9 million comments by 62 870 unique users, Newspaper 2 provided a total of 1.3 million comments by 49 509 registered users, and Newspaper 3 provided 1.6 million comments by 43 442 unique users.

[4]Supplementary material, Appendix Figure C.4 reports similar results for a sample including Swiss French Tweets.

https://doi.org/10.1017/psrm.2025.10063 Published online by Cambridge University Press

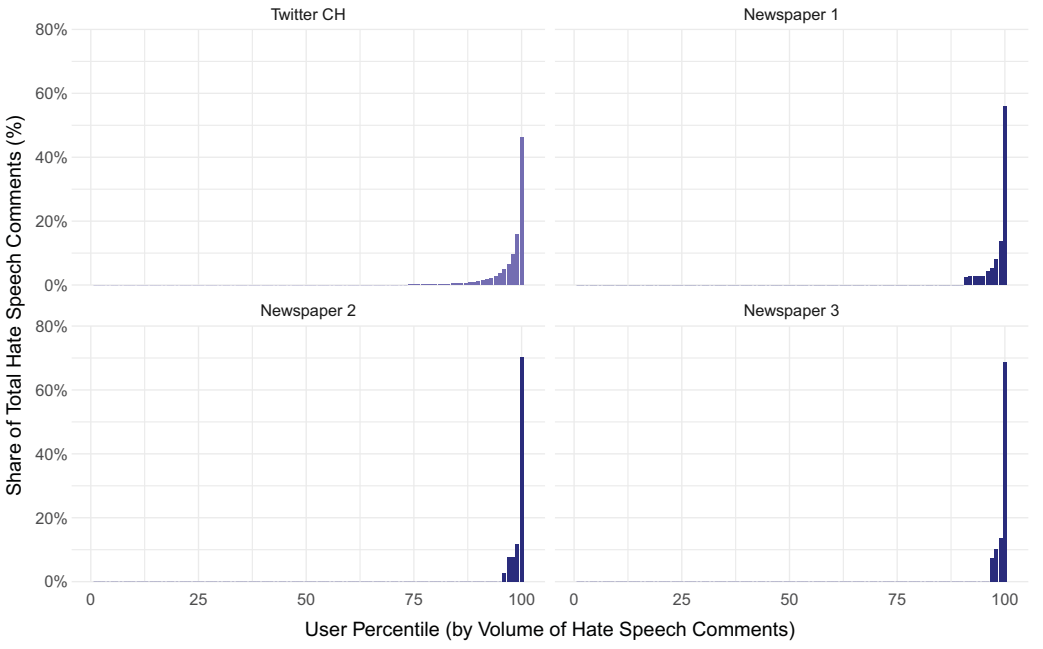


Figure 1. Hate speech on Swiss Twitter and in three online newspapers.
Share of Total Hate Speech Comments indicates the share of hate speech comments produced by each user percentile. *Twitter CH* includes all published tweets by users in the Swiss German Twitter sample ($N = 60\,808$ unique users). The other panels include all comments submitted during 2021 by registered users, published and unpublished. This amounts to $N = 62\,870$ unique users for *Newspaper 1*, $N = 49\,509$ for*Newspaper 2* (which includes only comments submitted from July 1 onwards), and $N = 43\,442$ for *Newspaper 3*.

The procedure resulted in the collection of 23.47 million tweets from 41 656 unique users. This additional robustness check further confirms that hate speech prevalence is relatively low and concentrated among a small set of users. In particular, the distribution of hate speech tweets is again very skewed; 1% of users are responsible for 25% of the hate speech produced, and 5% of users are responsible for 57%. We describe the data collection, analysis, and results in detail in Supplementary material, Appendix Section C.1.

1.3. Additional results and robustness

Appendix Section D (Supplementary material) presents a validation exercise using manual annotations of classifier results from Kotarcic *et al.* (2022) and the Perspective API on our descriptive samples, as well as details on training and evaluating an alternative classifier. Our findings on the distribution of hate speech across users are robust across classifiers, despite performance variations: all yield similarly skewed distributions, with a small number of users generating most hate speech. However, estimates of overall prevalence vary widely (e.g. 1.2% to 6.9% in the Swiss Twitter sample), making it difficult to draw firm conclusions. These figures suggest that hate speech is relatively uncommon but may be more prevalent than previously reported. Importantly, our analysis focuses on the distribution of hate speech, not its prevalence. Supplementary material, Appendix Section C.3 provides additional descriptive information on hate speech users, showing that they have larger networks and engage in more intense activity compared to users who have never used hate speech.

2. Counterspeech for frequent and occasional hate speech users

The descriptive findings challenge the assumption that hate speech is a widespread behavior online, showing instead that it is concentrated among a small subset of users. Next, we examine how this

https://doi.org/10.1017/psrm.2025.10063 Published online by Cambridge University Press

concentration affects the effectiveness of counterspeech, distinguishing between frequent and occasional hate speech users.

2.1. Data and methods

We conducted a field experiment to explore the effects of counterspeech interventions from November 14th, 2021, to February 28th, 2022, following our preregistered design. For a subset of users from our Swiss Twitter sample, we collected tweets from the last 24 hours daily.[5] We complemented this sample with tweets mentioning political hashtags and keywords commonly used in the Swiss context (see Supplementary material, Appendix Section E.1 for the complete list). Upon collection, all tweets received a classifier probability of being hate speech (based onKotarcic *et al.*, 2022) and were ranked by descending probability. Research assistants manually checked all tweets to ensure that the experimental sample included only tweets containing hate speech. They also excluded tweets originating from minors, organizations, or bots, and manually coded the targets of hate (see the complete workflow in Supplementary material, Appendix Section E.2). The $N = 2\,387$ users who used hate speech constitute our experimental sample. We randomly assigned study subjects to one of five treatment variations, each with a 15% probability, or the control group with a 25% probability. The main analysis groups those treatment variations into three main categories that build on Hangartner *et al.* (2021); Gennaro *et al.* (2025) and reproduce commonly used counterspeech strategies.

In the *Empathy* condition, the message encouraged subjects to put themselves in the position of the group or person against whom they used hateful language (perspective-taking), or reported the experience of a member of that outgroup (perspective-getting). An example message would be: "When [Muslim] friends of mine see tweets like this, it depresses them every time." In the *Warning of Consequences* conditions, subjects were reminded of the possible online and real-life consequences of using hate speech online, including from their employers and/or legal consequences. For example: "You should be aware that your colleagues, including your work environment, could also read this." The *Alerting of Hate Speech* condition made clear that the message has crossed the line into hate speech. For instance: "Are you aware that this comment is hate speech?" (alert), or "Thank you for the nice hate speech comment. I will embroider it onto a pillow" (humor). The Supplementary material, Online Appendix Sections E.3 and E.4 report additional details on treatment variations, as well as results for their separate effects.

The intervention consisted of issuing a direct, publicly visible counterspeech reply to the hateful tweet within 24 hours of the subject's original post. The treatments were administered through five researcher-controlled sockpuppet accounts that had been created at least four weeks prior to the start of the field phase. These accounts were created to look like real Twitter users, but did not convey any identifying demographic information.

We report results for two main preregistered outcomes. *Original Hate Tweet Deleted* measures whether the sender deleted their hate tweet within 12 hours after treatment. *Probability of Hate Tweets* measures the average classifier probability of hate speech in tweets posted up to four weeks after the intervention. Results for two other preregistered outcomes, i.e. the absolute and relative number of hate tweets, are reported in Supplementary material, Appendix Section E.4. We also collect users' complete timeline, which we use to differentiate among frequent and occasional hate speech users.

2.2. Results

A few users deleted their accounts ($N = 122$), set them to private ($N = 22$), or were suspended by Twitter ($N = 67$) between the treatment application and the end of the study period. We were unable to retrieve outcome data for one additional account. We exclude these accounts from the analysis.

[5]Instead of using the same list of users, we use a more condensed list of 125,690 users following at least five accounts in the political or news-aligned list.

https://doi.org/10.1017/psrm.2025.10063 Published online by Cambridge University Press

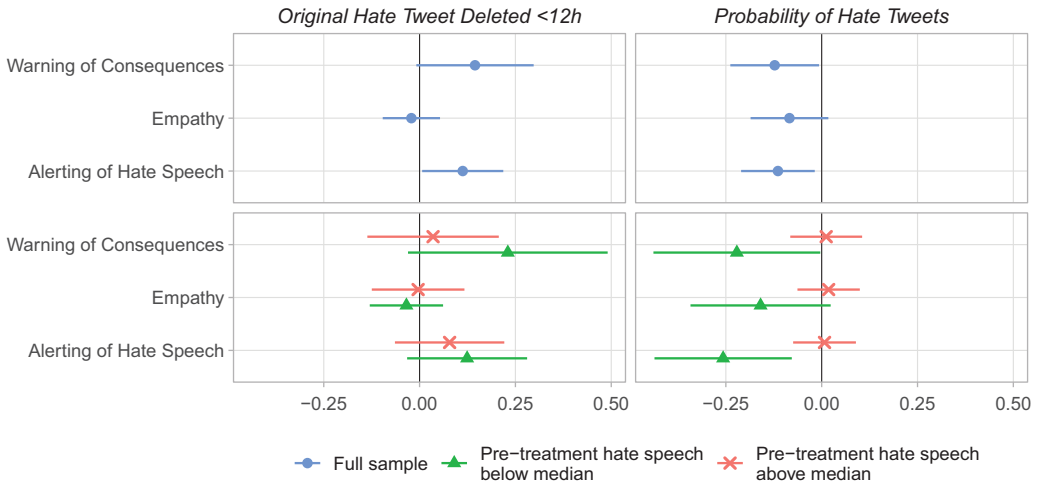


Figure 2. Experimental results.
Point estimates with 95% confidence intervals from OLS regressions. Outcomes are standardized (mean = 0, SD = 1) and include the probability of original hate tweet deletion within 12 hours and the classifier-predicted probability of hate tweets over four weeks. Full-sample regressions were preregistered; sub-sample regressions are exploratory. Full results are reported in Supplementary material, Appendix Tables E.5 and E.11.

Supplementary material, Appendix Table E.9 shows that attrition occurs with the same likelihood across all treatment arms and the control group. The remaining analysis sample contains $N = 2,175$ users, of which $N = 627$ are assigned to empathy, $N = 680$ to alerting of hate speech, $N = 325$ to warning of consequences, and $N = 543$ to the control group.[6] Supplementary material, Appendix Table E.10 shows that randomization generally produced comparable experimental groups.

For each experimental condition, we estimate the treatment effect by regressing the outcome on an indicator variable that takes the value of 1 for users assigned to that condition and 0 for users assigned to the control group. In line with our preregistered analysis, we use Lasso-based post-double selection (Belloni *et al.*, 2014) to select the predictive covariates (and first-order interactions) from Twitter account features. We report results from a second preregistered specification without covariates in Supplementary material, Appendix section E.4. We find mostly small effects across all treatments and outcomes. The upper-left panel of Figure 2 reports some evidence in support for a positive effect of counterspeech messages that warn of offline consequences ($\beta = 0.143$, $SE = 0.078$, $p = 0.068$, $p_{BH} = 0.091$) and alert of hate speech ($\beta = 0.110$, $SE = 0.054$, $p = 0.044$, $p_{BH} = 0.087$) on the probability of deleting the original hate tweet. The same treatments also reduce the probability of hate speech in the following four weeks (Alerting of Hate Speech: $\beta = -0.116$, $SE = 0.049$, $p = 0.018$, $p_{BH} = 0.071$; Warning of consequences: $\beta = -0.123$, $SE = 0.059$, $p = 0.037$, $p_{BH} = 0.075$), as shown in the upper-right panel. Those results remain statistically significant at the 10% level when accounting for multiple comparison adjustment. In contrast, we find no statistically significant effect of empathy-based treatments on these two outcomes, or any effect of any treatment on two other preregistered outcomes, i.e. the number of hate tweets and their share over the total number of tweets. Supplementary material, Appendix Table E.5 reports the complete regression estimates; Appendix Sections E.4 and E.7 report the results of additional pre-registered specifications and heterogeneity analyses.

[6]Within Empathy: $N = 318$ to perspective-taking, $N = 309$ to perspective-getting. Within alerting of hate speech: $N = 350$ to alerting of hate speech, $N = 330$ to humor.

https://doi.org/10.1017/psrm.2025.10063 Published online by Cambridge University Press

Instructed by the descriptive results, we conducted an exploratory analysis to understand if the limited overall effectiveness of the counterspeech interventions may be due to underlying heterogeneity among users. In particular, users who frequently produce hate speech may be less responsive to counterspeech compared to those who use it only occasionally. As this analysis was not pre-registered, the results should be interpreted with caution; the limited sample size in the subgroup analysis may affect the robustness of the findings.

To investigate this pattern, we classify each tweet in users' pre-treatment timelines and split the users in two groups, based on the median number of pre-treatment hate tweets. Then, we run the analysis separately on the two samples. While this reduces the power of the analysis, it allows us to reveal meaningful differences across groups. The bottom-left panel of Figure 2 reports the estimated treatment effects on the probability of deleting the original hate tweet. For the sample of users with low pre-treatment hate speech use (green triangles), the effect estimates of Alerting of Hate Speech ($\beta = 0.122$, $SE = 0.080$, $p = 0.127$) and Warning of consequences ($\beta = 0.232$, $SE = 0.133$, $p = 0.079$) are positive and sizeable. The same treatments also significantly reduce the probability of hate speech (Alerting of Hate Speech: $\beta = -0.257$, $SE = 0.091$, $p = 0.005$; Warning of consequences: $\beta = -0.222$, $SE = 0.111$, $p = 0.045$), as shown in the bottom-right panel. Again, empathy-based treatments do not have significant effects on those outcomes. In the same panels, the red crosses represent the effect estimates for the sample of users with above-median usage of hate speech pre-treatment. For these users, all point estimates are close to zero and not statistically significant. Supplementary material, Appendix Table E.11 reports the complete regression estimates.

The results suggest that while the intervention had weak effects on changing users' behavior, the reason behind this may be that the most prolific hate speech users are strongly entrenched in their posting habits. The skewness of the Swiss hate speech distribution suggests that users who employ hate speech are few but do so massively. Exploratory analyses of the experimental results indicate that this type of user will not be moved by the intervention.

2.3. Additional results and robustness

The wide confidence intervals observed in the experiment may raise concerns about statistical power. Supplementary material, Appendix Section F shows that, based on realized sample sizes and R-squared values, the minimum detectable effect is smaller than the estimated treatment effects. The study is well-powered to detect effects of 0.1–0.2 standard deviations, consistent with the pre-analysis plan. Equivalence confidence intervals further rule out true effects larger than approximately ± 0.2 standard deviations, confirming that the main effects are small.

The subgroup analyses are exploratory and rely on small sample sizes, limiting statistical power. Among infrequent hate users, estimated effects exceed the minimum detectable size; conversely, among frequent users, estimated effect sizes fall below detection thresholds, and null results may reflect true absence of an effect or insufficient power. An interaction analysis shows that pre-treatment hatefulness significantly moderates future hate tweeting (but not deletion), though these results should be interpreted with caution given limited power and p-values larger than 0.05 for two treatment arms (Alerting of Hate Speech and Empathy). Full details are in Supplementary material, Appendix Section F.

Supplementary material, Appendix Section G provides additional information on hate speech targets, based on manual annotations by research assistants. Most detected hate speech appears politically motivated. We find no consistent evidence of heterogeneous treatment effects across different target groups.

3. Conclusion

Our study captures a specific moment in the evolution of Twitter (now X). Although the platform has changed since our data collection, the insights on the limitations and potential of user-initiated

https://doi.org/10.1017/psrm.2025.10063 Published online by Cambridge University Press

interventions still apply in an era of declining platform commitment to moderation of harmful content, and rising online hate.

This study contributes to the understanding of online hate speech in two key ways. First, our descriptive analysis shows that a small minority of users accounts for the vast majority of hate speech. This pattern is consistent across platforms, linguistic contexts, and classification methods. Second, our experimental results indicate that the most prolific hate speech producers are largely resistant to counterspeech interventions, limiting counterspeech effectiveness to occasional offenders.

Our findings suggest that broad-based strategies, including counterspeech, face significant challenges. Uniform interventions risk expending resources on harmless content and unnecessarily burdening users who do not engage in hate speech. At the same time, these approaches may fail to reach persistent offenders, whose motivations may differ from occasional users.

A promising direction for policy is to move toward more targeted—and potentially personalized—interventions focused on the small minority of users responsible for most harmful content. Future research should explore how different types of users respond to counterspeech, and whether certain counterspeech strategies are more effective for specific audiences or types of hate speech. These heterogeneity analyses should be pre-registered to ensure adequate statistical power for detecting small effects within experimental subgroups.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2025.10063. This study has been approved by the ETH ethics committee with protocol EK 2021-N-178.

Funding statement. We are grateful to InnoSuisse (Grant 46165.1 IP-SBM), and the Swiss Federal Office of Communications for funding.

Data availability statement. Replication materials, including anonymized data and code, are available at https://doi.org/10.7910/DVN/CFN6X9.

Competing interests. Authors declare that they have no competing interests.

References

Barberá P, Casas A, Nagler J, Egan PJ, Bonneau R, Jost JT and Tucker JA (2019) Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review* **113**, 883–901.

Belloni A, Chernozhukov V and Hansen C (2014) Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* **81**, 608–650.

Bor A and Petersen MB (2022) The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review* **116**, 1–18.

Cao A, Lindo JM and Zhong J (2023) Can social media rhetoric incite hate incidents? Evidence from trump's "chinese virus" tweets. *Journal of Urban Economics* **137**, 103590.

Douek E (2021) Governing online speech: From 'posts-as-trumps' to proportionality and probability. *Columbia Law Review* **121**, 759–833.

Dozen CD (2021) Why platforms must act on twelve leading online anti-vaxxers. https://www.counterhate.com/disinformationdozen (accessed 3 May 2021).

ElSherief M, Nilizadeh S, Nguyen D, Vigna G and Belding E (2018) Peer to peer hate: Hate speech instigators and their targets. *Proceedings of the International AAAI Conference on Web and Social Media* **12**, 52–61.

Gennaro G, Derksen L, Abdelrahman A, Broggini E, Green MA, Haerter VA, Heer E, Heidler I, Kauer F and Kim H-N *et al.* (2025) Counterspeech encouraging users to adopt the perspective of minority groups reduces hate speech and its amplification on social media. *Scientific Reports* **15**, 22018.

Hangartner D, Gennaro G, Alasiri S, Bahrich N, Bornhoft A, Boucher J, Demirci BB, Derksen L, Hall A, Jochum M, Munoz MM, Richter M, Vogel F, Wittwer S, Wüthrich F, Gilardi F and Donnay K (2021) Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences* **118**, e2116310118.

He B, Ziems C, Soni S, Ramakrishnan N, Yang D and Kumar S (2021) Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis **137**, 90–94.

Kotarcic A, Hangartner D, Gilardi F, Kurer S and Donnay K (2022) Human-in-the-loop hate speech classification in a multilingual context *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 7414–7442.

https://doi.org/10.1017/psrm.2025.10063 Published online by Cambridge University Press

Müller K and Schwarz C (2023) From Hashtag to Hate Crime: Twitter and Antiminority Sentiment. *Applied Economics* **15**, 270–312.

Munger K (2017) Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior* **39**, 629–649.

Pradel F, Zilinsky J, Kosmidis S and Theocharis Y (2024) Toxic speech and limited demand for content moderation on social media. *American Political Science Review* **118**, 1895–1912.

Siegel AA and Badaan V (2020) #No 2Sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review* **114**, 837–855.

Siegel AA (2020) Online hate speech *Social media and democracy: The state of the field, prospects for reform*, 56–88.

Thomas DR and Wahedi LA (2023) Disrupting hate: The effect of deplatforming hate organizations on their online audience. *Proceedings of the National Academy of Sciences* **120**, e2214080120.

United Nations (2020) *United Nations Strategy and Plan of Action on Hate speech*. New York: United Nations.

Yildirim MM, Nagler J, Bonneau R and Tucker JA (2023) Short of suspension: How suspension warnings can reduce hate speech on twitter. *Perspectives on Politics* **21**, 651–663.

Cite this article: Gennaro G *et al*. (2026) The distribution of hate speech and its implications for content moderation. *Political Science Research and Methods* **14**, 1540–1548. https://doi.org/10.1017/psrm.2025.10063

https://doi.org/10.1017/psrm.2025.10063 Published online by Cambridge University Press