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]University College London
[2]ETH Zurich
[3]University of Zurich

July 25, 2025

Word count: 4010 (including references)

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.

*Corresponding author:dominik.hangartner@gess.ethz.ch

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). 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. 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. Pre-registered 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 (Bor and Petersen 2022; ElSherief et al. 2018). 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). 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]The pre-analysis plan is available at https://osf.io/xvwgd/. We discuss minor deviations in the Appendix section B.

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.

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

[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.

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,

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

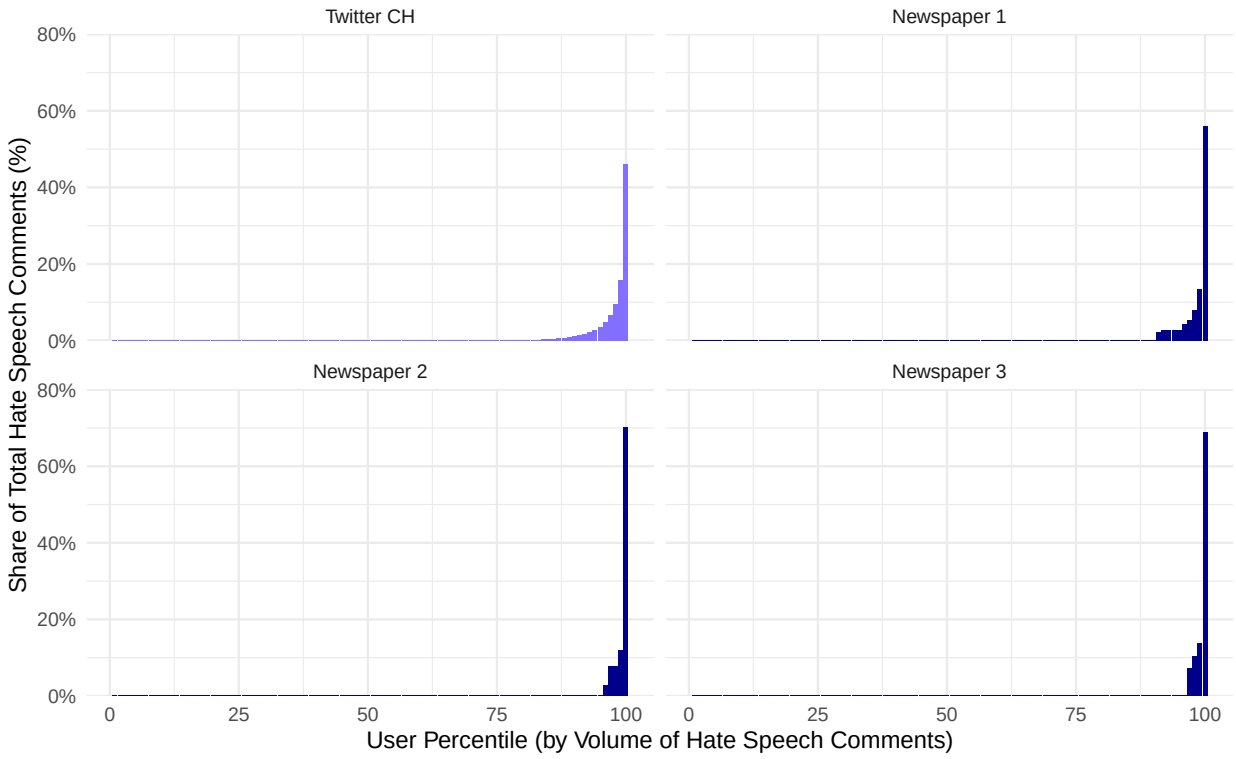
including their different audiences: Newspaper 3 is a broadsheet with the smallest audience, while the other two engage in tabloid journalism and have a wider audience.

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. 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 Appendix C.1.

1.3 Additional results and robustness

Appendix Section D 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. 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.

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*.

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 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 Appendix Section E.1 for the complete list). Upon collection, all tweets received a classifier probability of being hate speech (based on Kotarcic 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 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) and reproduce commonly used counterspeech strategies.

In the *Empathy* condition, the message was designed to prompt subjects to put themselves in the position of the group or person against whom they used hateful language (perspective-taking), or provided with the perspective of a member of the targeted 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

[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.

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 Online Appendix Sections E.3 and E.4 report additional details on treatment variations, as well as results for their separate effects.

The intervention involved delivering a direct and publicly visible counterspeech response to the hate tweet, within 24 hours of the subject's initial posting. To administer the treatments, we utilized five human-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 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. 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] Appendix Table E.10 shows that randomization generally produced comparable

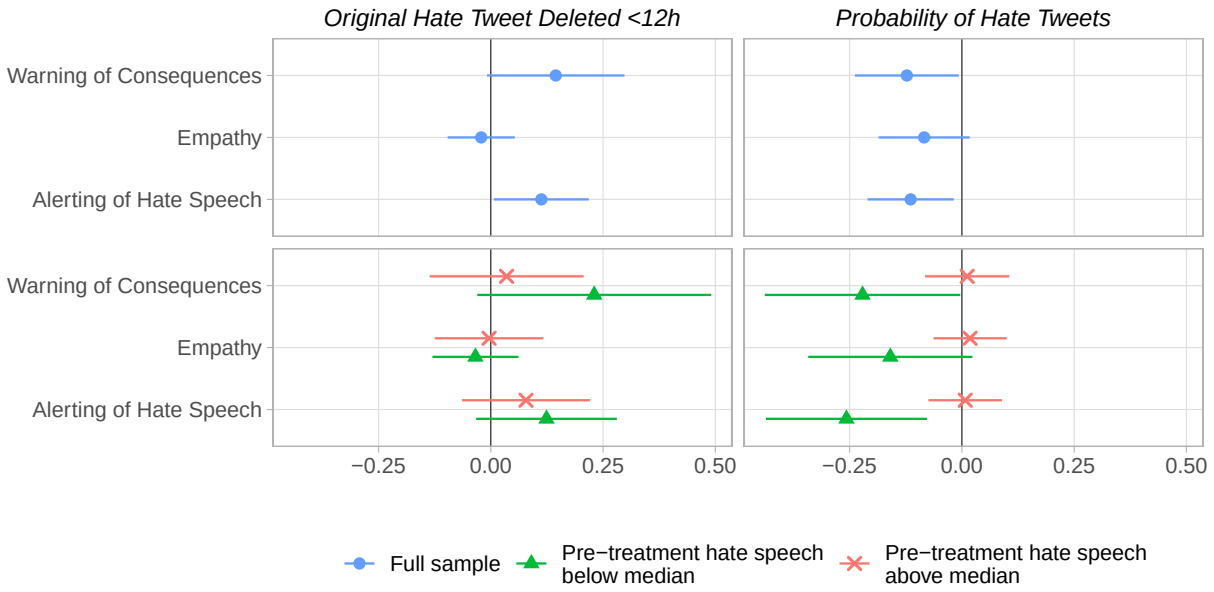
[6]Within Empathy: $N = 318$ to perspective-taking, $N = 309$ to perspective-getting. Within alerting of hate speech: $N = 350$ to alerting of hate speech, $N = 330$ to humor.

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 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. 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.

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.

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 Appendix Tables E.5 and E.11.

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. 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. 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 Appendix Section F.

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-driven interventions remain relevant at a time when platform moderation is being scaled back and online hate is rising.

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.

Data Availability: Replication materials, including anonymized data and code, are available at the Harvard Dataverse [*DOI to be added after acceptance*].

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

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

References

Barberá, P., A. Casas, J. Nagler, P. J. Egan, R. Bonneau, J. T. Jost, and J. A. Tucker (2019). "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data". In: *American Political Science Review* 113.4, pp. 883–901.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). "Inference on treatment effects after selection among high-dimensional controls". In: *Review of Economic Studies* 81.2, pp. 608–650.

Bor, A. and M. B. Petersen (2022). "The Psychology of Online Political Hostility: A Comprehensive, Cross-National Test of the Mismatch Hypothesis". In: *American Political Science Review* 116.1, pp. 1–18.

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

Douek, E. (2021). "Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability". In: *Columbia Law Review* 121.3, pp. 759–833.

Dozen, C. D. (2021). "Why Platforms Must Act on Twelve Leading Online Anti-Vaxxers". In: *Online: https://www. counterhate. com/disinformationdozen (accessed 3 May 2021)*.

ElSherief, M., S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding (2018). "Peer to Peer Hate: Hate Speech Instigators and Their Targets". In: *Proceedings of the International AAAI Conference on Web and Social Media* 12.1.

Hangartner, D., G. Gennaro, et al. (2021). "Empathy-based counterspeech can reduce racist hate speech in a social media field experiment". In: *Proceedings of the National Academy of Sciences* 118.50, e2116310118.

He, B., C. Ziems, S. Soni, N. Ramakrishnan, D. Yang, and S. Kumar (2021). "Racism is a virus: Anti-Asian hate and counterspeech in social media during the COVID-19 crisis".

In: *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining*, pp. 90–94.

Kotarcic, A., D. Hangartner, F. Gilardi, S. Kurer, and K. Donnay (2022). "Human-in-the-Loop Hate Speech Classification in a Multilingual Context". In: *Findings of the Association for Computational Linguistics: EMNLP*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 7414–7442.

Müller, K. and C. Schwarz (2023). "From Hashtag to Hate Crime: Twitter and Antiminority Sentiment". In: *Applied Economics* 15.3, pp. 270–312.

Munger, K. (2017). "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment". In: *Political Behavior* 39.3, pp. 629–649.

Pradel, F., J. Zilinsky, S. Kosmidis, and Y. Theocharis (2024). "Toxic Speech and Limited Demand for Content Moderation on Social Media". In: *American Political Science Review*, pp. 1–18.

Siegel, A. A. (2020). "Online Hate Speech". In: *Social Media and Democracy: The State of the Field, Prospects for Reform*. Ed. by N. Persily and J. A. Tucker. Cambridge University Press, pp. 56–88.

Siegel, A. A. and V. Badaan (2020). "#No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online". In: *American Political Science Review* 114.3, pp. 837–855.

Thomas, D. R. and L. A. Wahedi (2023). "Disrupting hate: The effect of deplatforming hate organizations on their online audience". In: *Proceedings of the National Academy of Sciences* 120.24, e2214080120.

United Nations (2020). *United Nations strategy and plan of action on hate speech*.

Yildirim, M. M., J. Nagler, R. Bonneau, and J. A. Tucker (2023). "Short of suspension: How suspension warnings can reduce hate speech on twitter". In: *Perspectives on Politics* 21.2, pp. 651–663.

Supplementary Information

Table of Contents

A Ethical considerations **2**

B Relation to the Pre-analysis Plan **3**

C Distribution of Hate Speech **5**

C.1 Collection of the U.S. Twitter sample 5

C.2 Swiss sample, including French Tweets 6

C.3 Profiling users who share hate speech 7

D Robustness of the descriptive results **11**

D.1 Validation of Hate Speech Classifiers 11

D.2 Alternative Fine-Tuned LLM Classifiers 13

D.3 Distributional Changes 15

E Experiment **17**

E.1 Data Collection Query 17

E.2 Flowchart of Counter-speech Application 18

E.3 Treatments 19

E.4 Tables 19

E.5 Attrition 20

E.6 Balance 21

E.7 Heterogeneous Treatment Effects 22

F Power and equivalence bounds **31**

G Targets of hate speech **33**

A Ethical considerations

This study has been approved without reservation by the ETH ethics committee [*protocol reference to be added upon acceptance*]. In this section, we report the main ethical considerations discussed in the application and subsequent approval.

Participants' consent. The subjects of this study do not give their informed consent to participate in the experiment. Two main reasons guide this decision. First, getting their informed consent could affect their behavior and introduce experimenter effects, whereby the participants alter their natural behavior because they know they are being observed. Second, this would also expose researchers' identities to, mostly anonymous, online users. As some of the participants may be active members of online hate groups, we are concerned about potential online and offline consequences for researchers' safety, including bullying and harassment. Our approach is in line with similar studies conducting digital field experiments on Twitter (Hangartner et al. 2021; Munger 2017) and respects contemporary Twitter's policy regarding academic research. Moreover, Twitter users are aware that their posts are public, and receiving replies, including counterspeech, pertains to the usual experience on the platform. Users have the option of restricting the public availability of their posts if they wish to (by making their accounts private).

Backlash. Another concern is that counterspeech messages may backlash, and produce an increase in the use of xenophobic hate speech, rather than a reduction. We have selected non-hostile strategies theorized to be effective, therefore we do not expect a significant backlash effect. Previous studies have documented only shortlived (Munger 2017) or no backlash (Hangartner et al. 2021). Our expectation was that the minimal risk of a backlash would be offset by the reduction in hate speech in the medium and long term.

Participants' identity. Subjects could be exposed to a data protection risk, if their Twitter handle, User ID, or the content of their tweets were revealed. To mitigate those risks

during the project phase, we stored the data in three tiers. First, each subject received a randomly generated pseudonym number. Personal identifiers were stored alongside the corresponding pseudonym numbers in an identification key data set, separate from other data. Second, the tweets were stored independently and identified only by the pseudonym numbers. In a third dataset, their post-treatment behavior was sorted using the pseudonym numbers. Throughout the project, data access was limited to the research team, and safeguarded by the ETH IT infrastructure under an identity management system. All members of the research team were informed of the ETH research code of conduct. Upon publication, the text of the tweets and all identifying information will be deleted. The pseudo-anonymized data will be published in a public repository, for the purpose of scientific replication.

B Relation to the Pre-analysis Plan

The online field experiment was pre-registered before the start of the data collection (available at https://osf.io/xvwgd/). In this section, we illustrate and justify instances where the reported results deviate from the original pre-analysis plan.

Workflow, sample selection, treatment administration. The implemented workflow, sample selection, and treatment administration fully comply with the pre-registered plan. We collected 2 387 study participants, which is close to the 2 400 pre-registered sample.

Outcomes. We report results for the four pre-registered "primary outcomes of interest" in the appendix Tables E.5 and E.6. Two of these are also reported in the main text, namely the probability of deleting the original hate tweet, and the classifier probability of hate tweets (Figure 2).

Statistical analysis. The implemented statistical analysis fully complies with the pre-registered plan. We pre-registered two sets of analyses to estimate the effect of the individual

treatments, and of the grouped treatments, on the outcomes. For each set of these analyses, we pre-registered two different statistical models: an unconditional OLS (Ordinary Least Squares) regression of the outcomes on treatment indicators, and a Lasso-based regression. Results for individual treatments are reported in Tables E.8 and E.7. The results for grouped treatments are reported in Tables E.5 and E.6. For the main analyses in Tables E.5 and E.7, we also report p-values adjusted for multiple comparisons.

Heterogeneous treatment effects. We pre-registered three heterogeneity analyses: by account status, by account age, and by anonymity. These are reported in Tables E.13, E.14, and E.15. In the main text, we report an exploratory analysis that shows heterogeneous treatment effects by pre-treatment hate speech.

Hypothesis. We do not find systematic support for our four primary hypotheses, that stated that receiving any treatment would (i) increase the likelihood of deleting the original hate tweets, (ii) reduce the number of hate tweets post-treatment, (iii) reduce the share of hate to non-hate tweets, (iv) reduced the classifier hate probability. Those null results are reported in Tables E.5 and E.6. As we do not find support for our primary hypotheses, we do not report formal estimates for the second set of hypotheses, that stated that *empathy*-based treatment should display the largest effects. Nonetheless, it is clear from the main effect estimates that those secondary hypotheses are also not supported. Finally, we report the test of the subgroup effects in Tables E.13, E.14, and E.15. We do not find support for systematic differences in the treatment effects across our three pre-registered moderators: account status, age, and anonymity.

Analyses not reported. We do not report results on the pre-registered exploratory analyses. This includes analyses on secondary outcomes (e.g., number of likes of the hate tweet, number of tweets), and on secondary treatment variations (e.g., question vs. statement). Those analyses are not informative when the main effect is not supported. We also do not

included analyses that were meant to test the robustness of a possibly significant result: the seemingly unrelated regression model, and the outcome imputation for accounts subject to attrition.

Not pre-registered analysis. The analysis of the heterogeneous treatment effects of our intervention, comparing effects across users posting high and low volumes of hate speech, was not pre-registered. We added this in light of the experiment's results, to understand if the weakness of the main results might be due to underlying heterogeneity among users that we did not anticipate in the PAP.

C Distribution of Hate Speech

C.1 Collection of the U.S. Twitter sample

To be able to generalize more on our Swiss results, we also collected Tweets over one month in 2023 (March) from a sample of U.S. Twitter users. The data collection procedure followed the same steps as that used for primary data collection in Switzerland. We used the RTwitterV2[7] package in R to gather the data. Due to the larger scale of the U.S., we encountered a significantly higher number of users following U.S. Congress members and News Media than the Swiss sample. As a result, we set higher thresholds for user inclusion: only users who followed more than 10 U.S. National News Papers or more than 30 U.S. Congress Members were included in the final list. This process resulted in a list of 251,507 users. Unfortunately, we could only collect tweets from 57,276 users as Twitter shut down access to their Academic API while we were in the midst of data collection. Nonetheless, we were still able to collect over 23.47 Million tweets from 41,656 users posted in March 2023, which is plenty to have a comparable sample with Switzerland. We classify those tweets using the Google Perspective API, specifically leveraging its endpoints for toxicity,

[7] https://github.com/maelkubli/RTwitterV2

severe toxicity, and identity attacks. In particular, a tweet is classified as hate speech if any one of these three indicators registers a score above 0.8.[8] Figure C.3 plots the results.

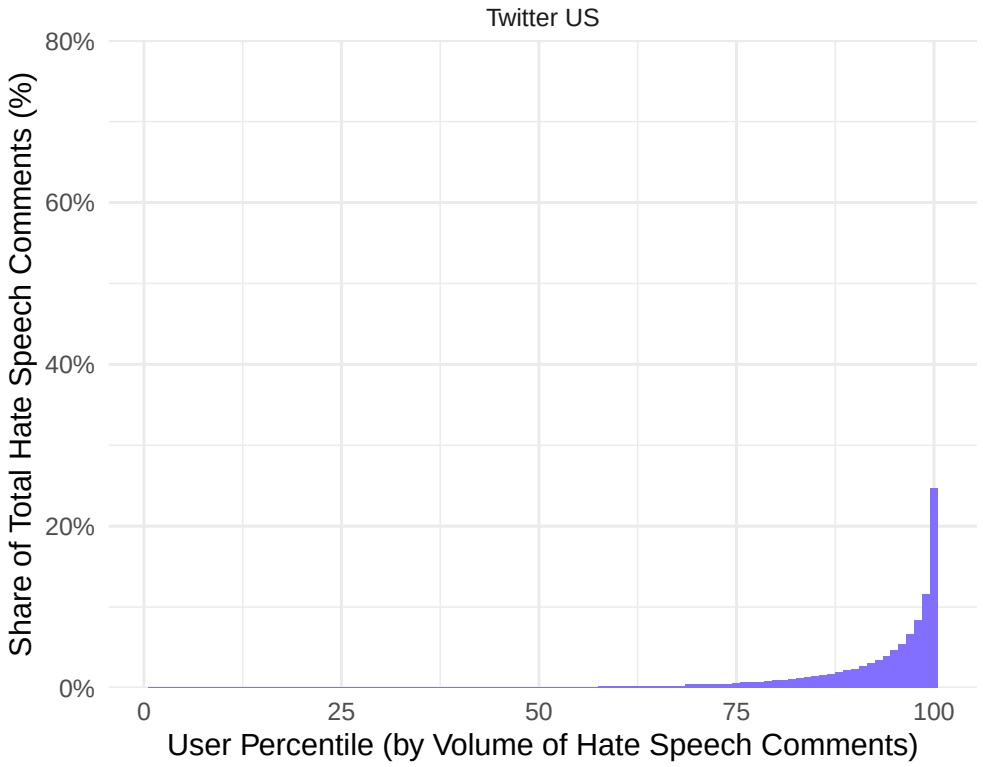


Figure C.3: Hate speech in the U.S. Twitter sample.
Prevalence indicates the share of hate speech tweets among all tweets. *Share of HS by percentile* indicates the share of hate speech tweets produced by each user percentile. *Twitter U.S.* includes all published tweets by our U.S. Twitter samples.

C.2 Swiss sample, including French Tweets

Figure C.4 shows the average percentage of hate speech tweets over all tweets, as well as their distribution across users, for the Swiss Twitter sample, including Tweets in both German and French. French is classified using the classifier by Kotarcic et al. (2022), similarly to German Tweets. We find that the share of hate speech tweets is 0.9% for the Swiss German and French sample. The distribution of hate speech displays the usual features: 1% of users

[8]We employ a threshold of 0.8 for this classification, following the guidelines of Google Perspective available at `https://www.perspectiveapi.com/`

are responsible for 49% of the hate speech produced. When expanding this observation to 5% of users, they are responsible for 85% of hate speech.

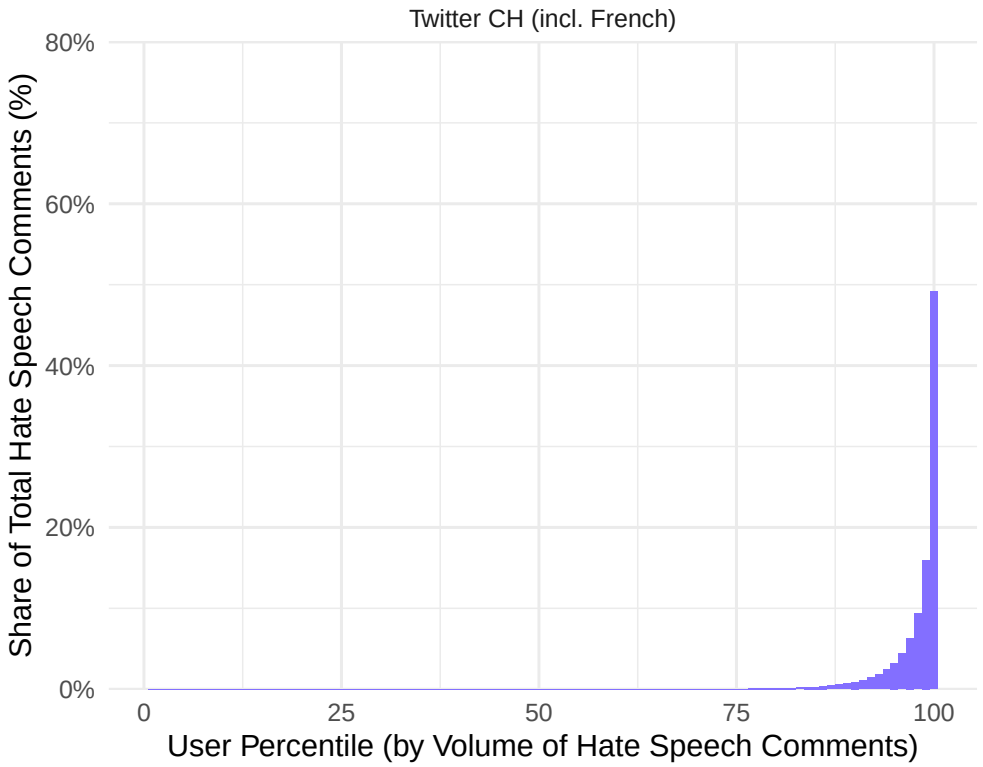


Figure C.4: Hate speech in the Swiss German- and French-language Twitter sample. *Prevalence* indicates the share of hate speech tweets among all tweets. *Share of HS by percentile* indicates the share of hate speech tweets produced by each user percentile. *Twitter CH* includes all published tweets by our Swiss German and French Twitter samples (N= 96 591). Swiss tweets are classified as hate if they score over 0.85 on the hate speech classifier developed in Kotarcic et al. (2022).

C.3 Profiling users who share hate speech

To gain a more descriptive understanding of the characteristics of users who posted hate speech, we collect several metadata variables and compare their mean values to those users who did not post hate speech. In addition to the metadata provided by the Twitter API—such as account creation date and number of followers—we construct a custom measure of anonymity.

C.3.1 Personal Identification Measures

To comprehensively classify how identifiable users are, we developed several proxies of user anonymity. Our preferred proxy consists of matching usernames with extensive lists of French, German, and English first names[9]. Names that match these data sets indicate a higher probability of authenticity (*Proportion of Verified Names* in Tables C.1-C.2). Second, we calculate the entropy of the usernames as a measure of linguistic randomness. Usernames exhibiting extremely high entropy, indicative of random character strings, or excessively low entropy, indicative of repetitive or structured numerical sequences, receive lower personal identification scores (less identifiable) (*User Name Entropy Score* in Tables C.1-C.2). Third, we consider the "Verified" account status, due to their inherent credibility and public visibility (*Proportion of Verified Accounts* in Tables C.1-C.2). We further analyze the content of profile descriptions, assigning positive scores to informative descriptions that explicitly mention professions, affiliations, or recognizable roles, while vague, nondescriptive, or empty profiles receive negative adjustments (*Anonymous Accounts Share* in Tables C.1-C.2).

C.3.2 Profiling Users with Observational Data: US and Swiss Contexts

We examine differences between users who shared some hate speech (HS) and users who never shared any hate speech tweet (non-HS) separately within the US (see Table C.1) and Swiss (see Table C.2) contexts.

In the US dataset, users who post hate speech also exhibit markedly higher overall activity as measured by total status counts. At the same time, their accounts tend to be younger, implying higher activity levels per day. HS users also maintain larger follower and friend networks, but belong to fewer Twitter lists compared to non-HS users. The prevalence of verified accounts is similar between groups, suggesting that verification alone does not

[9]French names obtained from `https://github.com/ThinkR-open/prenoms`, compiled by INSEE; German names sourced from `https://raw.githubusercontent.com/fxnn/vornamen/master/vornamen-grouped-sorted.dat` by the City of Cologne (2018); English names derived from the U.S. Social Security Administration's baby names dataset covering 1880-1997.

distinguish users who disseminate hate speech from others.

In the Swiss dataset, the patterns of differences between users who share hate speech and those who do not are largely similar to those observed in the US dataset. Specifically, we observe that users who share hate speech maintain significantly larger follower and friend networks compared to non-hate speech (non-HS) users. Furthermore, Swiss HS users exhibit notably higher posting activity (*status counts*) and have younger accounts. Notable differences appear for the *Anonymous Accounts Share* and the *Number of lists a user is a member of*, where the observed patterns are reversed compared to the US dataset. While the patterns are largely similar between the US and Switzerland, the levels of most variables are lower in Switzerland, likely reflecting a smaller and less active Twittersphere overall.

Table C.1: Profiling US hate speech users

Variable	Mean for HS Users	Mean for non HS Users	Δ	P-value
Proportion of Verified Accounts	0.052	0.050	0.001	0.528
Anonymous Accounts Share	0.064	0.061	0.003	0.258
Profile contains a Country	0.007	0.013	-0.007	0
Proportion of Verified Names	0.711	0.759	-0.048	0
Log of Followers Count	2.954	2.684	0.270	0
Log of Friends Count	3.428	3.295	0.134	0
Log of Status Count	4.312	3.461	0.851	0
No. of lists user is a member of	32.763	91.542	-58.779	0
Account Age (days)	2533.762	2866.420	-332.658	0
User Name Entropy Score	3.053	3.050	0.003	0.353

Notes. Differences between Users who have never shared any hate speech (Non-HS) and users who have shared at least once hate speech (HS) among US Twitter accounts. The table reports the average ($\bar{x}$) value of each variable for HS and Non-HS users, along with the difference between the two groups (Δ). P-values indicate that the observed differences are statistically significant.

Figure C.5 moves beyond the simple binary distinction between users who have shared hate speech at least once and those who have not. It divides users into four quartiles based on the absolute number of hate speech posts shared, and plots the standardized averages for each quartile across various user characteristic variables. In both the US context (left panel) and the Swiss context (right panel), we observe that users in the highest quartile of hate speech sharing consistently have larger networks of friends and followers, higher posting

Table C.2: Profiling Swiss hate speech users

Variable	Mean for HS Users	Mean for non HS Users	Δ	P-value
Proportion of Verified Accounts	0.022	0.017	0.005	0.000
Anonymous Accounts Share	0.081	0.091	-0.010	0.000
Profile contains a Country	0.014	0.019	-0.005	0.000
Proportion of Verified Names	0.685	0.725	-0.040	0.000
Log of Followers Count	2.414	1.887	0.525	0.000
Log of Friends Count	2.757	2.431	0.326	0.000
Log of Status Count	3.567	2.394	1.174	0.000
No. of lists user is a member of	27.733	20.562	7.171	0.000
Account Age (days)	2294.88	2498.31	-203.434	0.000
User Name Entropy Score	3.005	3.002	0.003	0.302

Notes. Differences between Users who have never shared any hate speech (Non-HS) and users who have shared at least once hate speech (HS) among Swiss Twitter accounts. The table reports the average ($\bar{x}$) value of each variable for HS and Non-HS users, along with the difference between the two groups (Δ). P-values indicate that the observed differences are statistically significant.

activity, and more recently created accounts. Given that hate speech sharing is measured in absolute terms, it is unsurprising that accounts which have shared more hate speech also tend to exhibit more intensive platform use in multiple respects. Future research should therefore consider the relative proportion of posts containing hate speech, as this measure is independent of overall posting activity. Nevertheless, the notable similarities across the two country contexts in profile characteristics plotted by quartile suggest that more in-depth profiling could yield valuable insights into the underlying patterns and drivers of hate speech sharing.

Figure C.5: Mean of standardized user-level characteristics across quartiles of hate speech posts shared among US Twitter users (left panel) and Swiss Twitter users (right panel). Error bars indicate 95% confidence intervals, which are small due to the large sample sizes.

D Robustness of the descriptive results

D.1 Validation of Hate Speech Classifiers

We conducted a comprehensive and rigorous validation of our hate speech classifiers using extensive manual annotation and stratified sampling. Because hate speech is a low-prevalence phenomenon in both Swiss and US datasets, simple random sampling would have led to a high proportion of non-hate tweets and unreliable performance estimates. To address this challenge, we use a stratified sampling design based on probabilities predicted by the classifier, following the method proposed by Tomás-Valiente (2025). This approach ensured adequate representation of tweets across the full range of model confidence scores, allowing us to more accurately estimate real-world error rates and evaluate classifier performance under varying degrees of certainty. To systematically address concerns and ensure robust

measurement of hate speech, two human coders independently manually annotated stratified samples from both Swiss and US datasets, each consisting of 500 tweets. An expert coder reviewed the independent manual annotations and consolidated them into a single label in case of disagreement. The ground truth hate speech labels obtained this way allowed us to precisely assess the strengths and limitations of the classifiers in detecting various forms of hate speech, including explicitly offensive content as well as subtle, implicit, or coded expressions that often evade automated detection systems.

The sampling method used to generate the samples involved several key steps. First, we divided the tweet populations into positive and negative categories based on classifier predictions. Following our classification protocol, Swiss tweets were classified as positive if the predicted probability of hate speech was greater than 0.85, while US tweets were classified as positive if the maximum Perspective API score exceeded 0.8 (considering identity attacks, toxicity, and severe toxicity). Next, an optimal allocation method determined the exact number of positives and negatives to sample from each dataset, minimizing the combined variance of precision, recall, and F1 score estimates. The initial parameters for this method were selected based on the prior validation results from Kotarcic et al. (2022) and lower-bound estimates recommended by Perspective API documentation for US data. The positive and negative groups were further stratified into five probability bins, ensuring proportional representation at different levels of confidence in the classifier predictions. We then applied stratified random sampling within these bins. The metrics were computed using a stratified estimator that weights tweets by the inverse of their sampling probabilities, ensuring robust and representative estimates.

Validating the classifier based on Kotarcic et al. (2022) for our Swiss Twitter dataset showed mixed performance. Our findings revealed a moderate level of precision, indicating that when the classifier identified hate speech, it was correct approximately 69% of the time. However, the recall of this classifier was particularly low, only around 10%, suggesting a significant under-detection of hate speech. This limitation became evident through the

manual annotation process, where annotators identified multiple cases of nuanced or implicit hate speech that the classifier could not detect.

The validation of Google's Perspective API classifier (where we combined identity attacks, toxicity, and severe toxicity scores) for the US Twitter dataset yielded similar results. Although precision was markedly higher (90%), recall was critically low (6%). This underscores the substantial limitations of the Perspective API in identifying hate speech that employs implicit or coded language, a common practice among users attempting to circumvent moderation efforts.

Table D.3 provides detailed metrics that capture the performance of both classifiers. Beyond standard measures such as precision, recall, and F1 score, the table includes additional robustness metrics like Matthews Correlation Coefficient (MCC), Cohen's Kappa, and Youden's J, which further contextualize classifier performance by taking class imbalance and random chance agreement into account.

Table D.3: Performance metrics of baseline hate speech classifiers

Metric	Swiss Classifier	US Classifier
Precision	0.7227	0.8955
Recall	0.1023	0.0619
F1-score	0.1792	0.1158
Accuracy	0.9147	0.9007
MCC	0.2509	0.2205
Kappa	0.1602	0.1036
Unweighted F1	0.5671	0.5316
Weighted F1	0.8844	0.8600
Youden's J (BM)	0.0983	0.0611

D.2 Alternative Fine-Tuned LLM Classifiers

To overcome the limitations of the baseline classifiers identified above, we fine-tuned several large language models (LLMs)—*LLaMA 3 8b*, *DeepSeek R1 Qwen 14b*, and *DeepSeek R1 Qwen 32b (8-bit)*—using Low-Rank Adaptation (LoRA). For fine-tuning, we used a distinct dataset of 500 expert-annotated comments, selected for their representativeness of

diverse and nuanced expressions of hate speech, including implicit and coded language forms; these expert annotations are drawn from Swiss newspaper comment data used in Umansky et al. (2024). The LoRA fine-tuning method involved freezing the original model weights and updating only a smaller set of adaptation matrices, which significantly reduces computational overhead while maintaining high performance.

Fine-tuning was conducted for three epochs, employing a learning rate of 2e-5, batch size of 16, and gradient accumulation steps to effectively manage memory usage. The training was monitored using a validation subset (10% of the fine-tuning data) to prevent overfitting and ensure optimal generalization. We validated the performance of these fine-tuned models using the same stratified and manually annotated set of 500 Swiss tweets employed in the baseline validation. Performance improvements are detailed in Table D.4.

Table D.4: Comparison of Baseline Classifiers with Fine-Tuned LLMs with the Swiss Tweets validation set

Model	Precision	Recall	F1 Score	Accuracy	F1 (Weighted)
Kotarcic et al. (2022) Classifier	0.72	0.10	0.18	0.91	0.88
Perspective API	1.00	0.00	0.01	0.90	0.87
LLaMA 3 8b	0.44	0.38	0.41	0.90	0.90
DeepSeek R1 Qwen 14b	0.50	0.36	0.42	0.91	0.90
DeepSeek R1 Qwen 32b (8-bit)	0.59	0.35	0.44	0.92	0.91

The two original classifiers exhibit relatively high precision but low recall, leading to poor F1 scores for the detection of hate speech. In contrast, fine-tuned LLMs, especially the DeepSeek R1 Qwen models, show a significantly improved balance between precision and recall, thus achieving substantially higher F1 scores. These classification scores are in line with the latest results in unbalanced hate speech detection tasks (see, e.g., Hernández-González et al. 2024; Pen et al. 2024). We opted to use the best performing model (DeepSeek R1 Qwen 32b) to annotate the additional datasets (e.g., experimental tweets, 1% sample of the Swiss Twitter Corpus, and a 10% sample of media comments),[10] which allowed more

[10]While much smaller than the full DeepSeek R1 model, the Qwen 32b version is still very large. Given the resulting runtimes, we reclassified a random 1% sample of Swiss tweets, i.e., around 560'000, and 10%

reliable and comprehensive checks of the robustness of the original results.

D.3 Distributional Changes

The fine-tuned LLM classifier (*DeepSeek R1 Qwen 32b*) detects more instances of hate speech compared to the Kotarcic et al. (2022) BERT-based classifier. However, a critical robustness check is whether this increase in sensitivity affects our primary conclusions drawn in the main analysis.

As shown in Figure D.6, the distributional patterns are very consistent between the two classifiers. In each dataset, a disproportionately large share of hate speech is consistently generated by users in the highest percentiles of hate-speechiness, underscoring the robustness of our central conclusion: hate speech production is highly concentrated among a small subset of highly active or extreme users.

As expected from the higher recall of the fine-tuned LLM classifier, it labels 6.9% of the tweets in the Swiss Twitter sample as hate speech, compared to only 1.2% identified by the BERT-based classifier of Kotarcic et al. (2022). Similarly, substantial gaps emerge in the media datasets: Newspaper 1 (9.9% vs. 0.8%), Newspaper 2 (8.7% vs. 0.5%), and Newspaper 3 (5.6% vs. 0.3%). Yet, despite these notable differences in overall prevalence, both classifiers attribute the vast majority of hate speech detected to the same small group of users.

The stability of this distributional pattern underscores the observation that although the sensitivity of the classifier vary, this does not affect our broader empirical conclusions regarding the highly skewed concentration of hate speech. This methodological robustness justifies our decision to rely on the original classifiers in the main analysis, especially given the considerable computational costs associated with employing large-scale LLM-based classification for extensive datasets.

of Swiss newspaper comments, i.e., around 580'000. Both samples are very large and therefore provide sufficiently precise estimates of the distributional characteristics we are interested in.

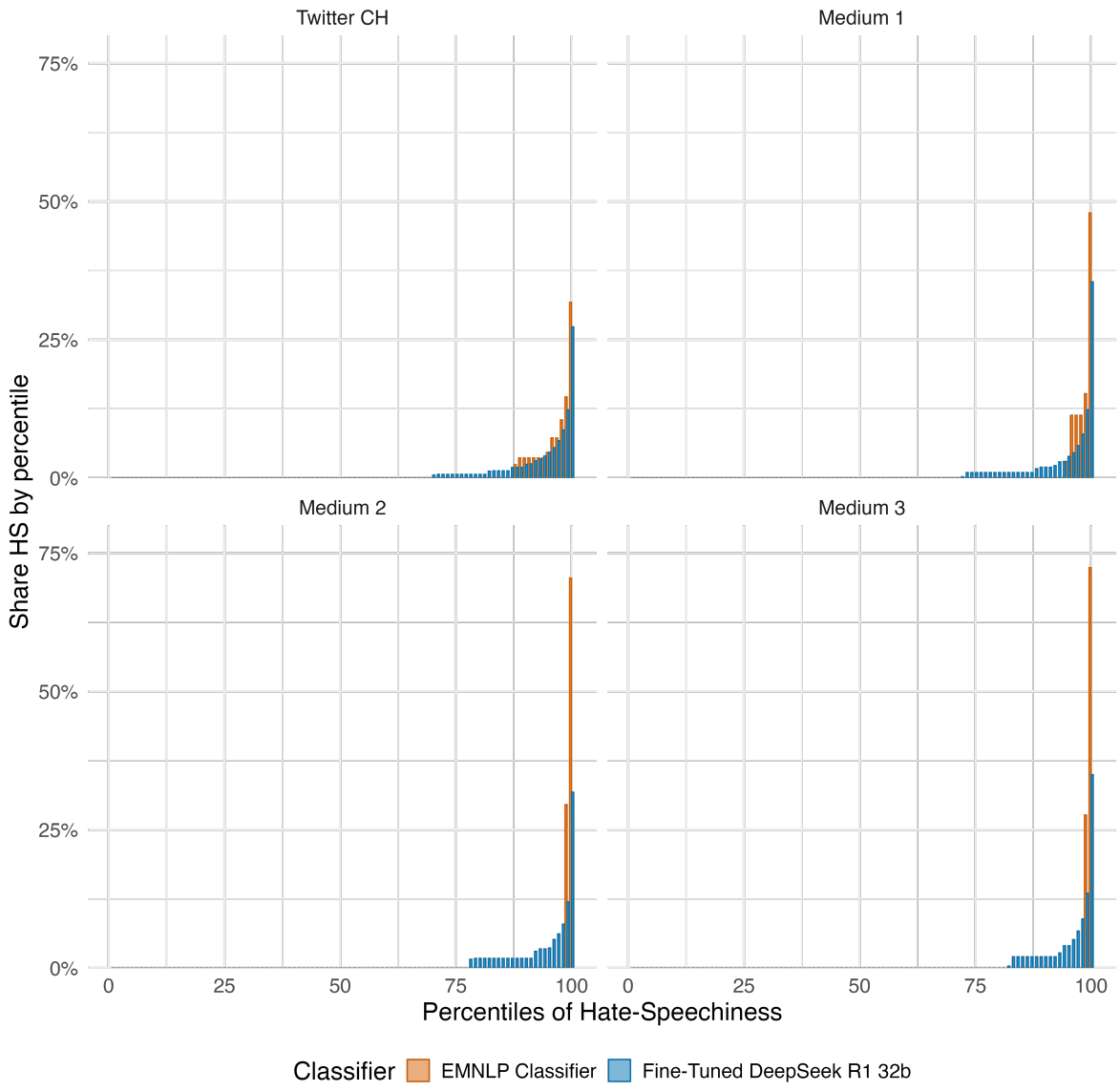


Figure D.6: The figure compares the share of hate speech produced by each user percentile across four datasets: Twitter CH (top left) and three media sources (Medium 1–3). The analysis is based on a 1% sample of all Swiss tweets and 10% samples from each media dataset. The bars represent the share of total hate speech attributed to each percentile, as classified by two models: the original EMNLP classifier of Kotarcic et al. (2022) (orange) and a fine-tuned LLM classifier (blue, DeepSeek R1 32b).

E Experiment

E.1 Data Collection Query

For the field experiment, we collected tweets from a fixed set of users to ensure an adequate sample size for the study of hate speech dissemination. In line with our pre-registration, we included accounts with at least five and fewer than 5000 followers, which are not verified, do not tweet more than 100 times per day on average since account creation, are older than one month, and have not already been treated. Since this approach yielded fewer eligible users for inclusion, we augmented the data collection process by incorporating an extensive list of keywords and hashtags specific to the Swiss context. By doing so, we could capture additional tweets from new users not initially present in the predefined user list. The list of keywords used for this purpose includes the following: *chvote, abst21, abst20, votech, srfarena, abst22, freiheitsimpfler, impfenstattschimpfen, #JaZumCovidGesetz, #NeinZumCovidGesetz, Covid19Gesetz, #JaZumCovid19Gesetz, #NeinZumCovid19Gesetz, swisscovidfail, swissboosterfail, #CovidCertificateCH, #Walliserkanne, FreiHeitsTrychler, #CovidgesetzNein, #CovidGesetzJa, CovidGesetz, #massvoll, #BREntscheid, #CoronaInfoCH, CoronaCH, COVID19CH, impfwoche, #JederPiksZählt, walkinimpfung, impfdorf, #langenNachtderImpfung, #ImpfwocheCH, impfnacht, impfdichinsknie, #Impflieferservice, #JedeImpfungZählt, #JedeImpfungZaehlt, LieberTee, frauensession, schweiz AND (maskenpflicht, booster, massnahme, 2g, 3g, idiot, impfpflicht, revolution, impfzwang, schwurbler), schweizer AND (medien, spital, impfverweigerer, nazi, idiot), polizei AND (zürich, been, schweiz, basel), taskforce AND (querdenker, schweiz, corona).* This extensive list of keywords ensured a continuous stream of new posts potentially containing hate speech, from users not yet treated.

E.2 Flowchart of Counter-speech Application

Appendix Figure E.7 reports the data collection pipeline, which includes all the different phases of hate speech identification and treatment.

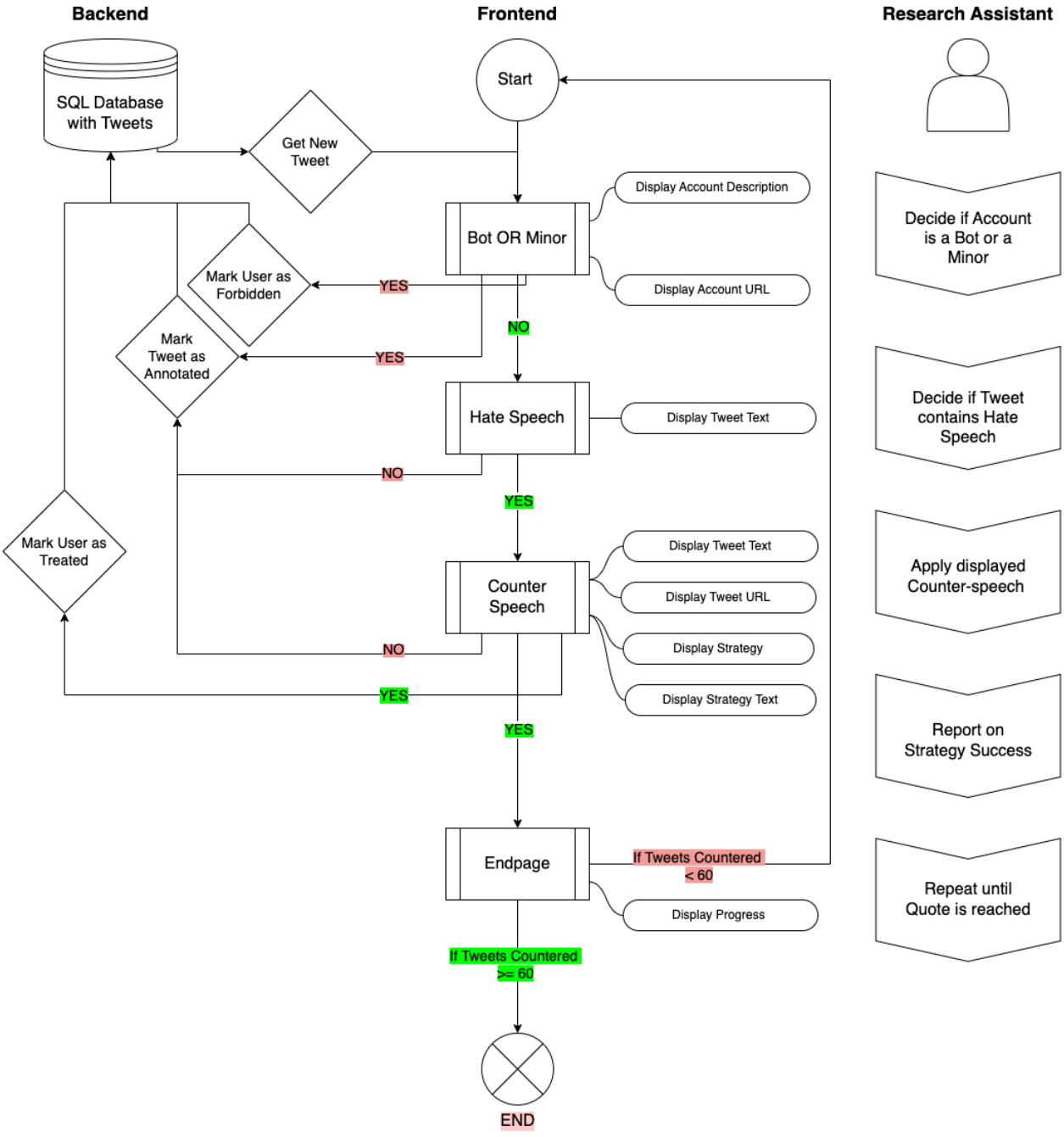


Figure E.7: Flowchart of Counter-speech Application

E.3 Treatments

Empathy. The "Empathy" condition included two possible variations, which were also randomized. In the "Empathy: Perspective-taking" condition, the message was designed to prompt subjects to put themselves in the position of the group or person against whom they had used hateful language. For example: "How would you feel if people were talking about you like this?". In the "Empathy: Perspective-getting" condition, users were provided with the perspective of a member of the targeted outgroup through the words of a potential ingroup friend or colleague. An example message would be: "When [Asian-American] friends of mine see tweets like this, it depresses them every time."

Warning of consequences. The "Warning of consequences" condition included only one variation, in which 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."

Alerting of hate speech. This condition included two variations, both of which were randomized. The "Alerting of hate speech" condition clarified that the message had crossed the line into hate speech. For instance: "Are you aware that this comment is hate speech?". The "Alerting of hate speech: Humor" condition attempts to use humour to respond to a hate speech comment in order to lighten the mood. One example is: "Thank you for the nice hate speech comment. I will embroider it onto a pillow."

E.4 Tables

Appendix Table E.5 reports the main regression results. Regression models include pre-treatment controls selected using Lasso-based post-double selection, among the following features: the number of friends and followers, past levels of activity, past hate speech and

toxicity, account age, language, whether a location or description is indicated, whether the account uses a real name or visible picture, and whether the user is a politician, a newspaper, or a politically interested user.

Appendix Table E.6 reports the unconditional treatment estimates obtained by regressing the outcome variable on treatment dummies, without additional controls.

Appendix Table E.7 reports the Lasso-based regression results for individual treatments. Regression models include pre-treatment controls selected using Lasso-based post-double selection, among the following features: number of friends and followers, past level of activity, past hate speech and toxicity, account age, language, whether a location or description is indicated, whether the account uses a real name or visible picture, whether the user is a politician, a newspaper or a politically interested user. In this case, treatment dummies indicate disaggregated treatment categories, as described in the previous Appendix section.

Appendix Table E.8 reports the unconditional treatment estimates, obtained regressing the outcome variable on treatment dummies, without additional controls. In this case, treatment dummies indicate disaggregated treatment categories, as described in the previous Appedix section.

E.5 Attrition

Appendix Table E.9 reports the attrition analysis. The first column reports the regression of a dummy variable that takes the value of 1 if the observation exited from the sample during the study period, independently from the reason, on treatment dummies. In the second column, the dependent variable takes the value of 1 if the observation was suspended by Twitter. In the third column, the dependent variable takes the value of 1 if the account was set to private. Importantly attrition does not appear to be related to treatment assignment.

Table E.5: Main results

Dependent Variable:	Original Hate Tweet Deleted	No. of hate tweets	Share of daily hate	Probability of Hate Tweets
Alerting of Hate Speech	0.110**	0.035	-0.054	-0.116**
	(0.054)	(0.056)	(0.060)	(0.049)
BH Adj. P-values	0.087	0.535	0.496	0.071
Observations	1223	1223	1223	1215
Warning of Consequences	0.143*	0.032	-0.143**	-0.123**
	(0.078)	(0.049)	(0.065)	(0.059)
BH Adj. P-values	0.091	0.515	0.075	0.075
Observations	868	868	868	857
Empathy	-0.022	-0.004	-0.089	-0.084
	(0.038)	(0.040)	(0.057)	(0.052)
BH Adj. P-values	0.761	0.919	0.231	0.231
Observations	1170	1170	1170	1156

Notes. This table reports the results of regressing the dependent variables (as indicated at the top of the table) on a dummy variable equal to 1 for each treatment condition, and 0 for control. Regression models include pre-treatment controls selected using Lasso-based post-double selection. Those controls are high-dimensional interactions among the following variables: number of friends and followers, past level of activity, past hate speech and toxicity, account age, language, whether a location or description is indicated, whether the account uses a real name or visible picture, whether the user is a politician, a newspaper or a politically interested user. Those interactions do not carry substantive meaning, and hence are not reported. *BH Adj. P-values* indicates p-values adjusted using the Benjamini- Hochberg (1995) procedure. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

E.6 Balance

Appendix Table E.10 reports the balance analysis. Each column reports the regression of the indicated pre-treatment control variable on treatment dummies. Observations are users who received a treatment status via randmization. Importantly randomization appears to have achieved a good balance of pre-treatment covariates between users assigned to the treatment and the control group. Minor imbalances are adjusted with post-double Lasso selection of controls.

Table E.6: Main results, OLS estimates

Dependent Variable:	Original Hate Tweet Deleted	No. of hate tweets	Share of daily hate	Probability of Hate Tweets
Alerting of Hate Speech	0.091	0.071	-0.018	-0.058
	(0.060)	(0.061)	(0.062)	(0.061)
Observations	1223	1223	1223	1215
R-squared	0.002	0.001	0.000	0.001
Warning of Consequences	0.113	0.033	-0.096	-0.114
	(0.081)	(0.064)	(0.073)	(0.070)
Observations	868	868	868	857
R-squared	0.003	0.000	0.002	0.003
Empathy	-0.052	0.000	-0.044	-0.042
	(0.044)	(0.047)	(0.061)	(0.063)
Observations	1170	1170	1170	1156
R-squared	0.001	0.000	0.000	0.000

Notes. This table reports the results of regressing the dependent variables (as indicated at the top of the table) on a dummy variable equal to 1 for each treatment condition, and 0 for control. The models do not include control variables. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

E.7 Heterogeneous Treatment Effects

Appendix Table E.11 reports the heterogeneous treatment effects across users with high and low pre-treatment hate speech use. Regression models include pre-treatment controls selected using Lasso-based post-double selection. Panel A reports results for the subsample of users who produced a number of hate tweets below the median in the 30 days preceding the experiment. Panel B reports results for users above the median. Moreover, Appendix Table E.12 reports the results of an interacted model, where the outcome variables are regressed on a treatment indicator, a high-hate-speech group indicator, and their interaction. In contrast to the models reported in Table E.11, the interacted models do not include controls.

Appendix Table E.13 reports the heterogeneous treatment effect across users with high and low pre-treatment status, measured as the number of followers. Regression models include pre-treatment controls selected using Lasso-based post-double selection. Panel A

Table E.7: Disaggregated treatments, Lasso-based estimates

Dependent Variable:	Original Hate Tweet Deleted	No. of hate tweets	Share of daily hate	Probability of Hate Tweets
Alert: Banning of Hate Speech	0.202**	0.040	-0.045	-0.112**
	(0.083)	(0.060)	(0.064)	(0.054)
Observations	893	893	893	886
BH Adj. P-Value	0.061	0.506	0.479	0.075
Alert: Humor	0.013	0.045	-0.057	-0.132**
	(0.054)	(0.095)	(0.079)	(0.062)
Observations	873	873	873	866
BH Adj. P-Value	0.817	0.817	0.817	0.130
Warning of Consequences	0.143*	0.032	-0.143**	-0.123**
	(0.078)	(0.049)	(0.065)	(0.059)
Observations	868	868	868	857
Adj. P-Value	0.091	0.515	0.075	0.075
Empathy: Perspective Taking	-0.036	0.025	-0.088	-0.156***
	(0.041)	(0.053)	(0.065)	(0.062)
Observations	861	861	861	851
BH Adj. P-Value	0.498	0.631	0.354	0.050
Empathy: Perspective Getting	-0.005	0.005	-0.088	-0.010
	(0.049)	(0.050)	(0.061)	(0.063)
Observations	852	852	852	842
BH Adj. P-Value	0.916	0.916	0.617	0.916

Notes. This table reports the results of regressing the dependent variables (as indicated at the top of the table) on a dummy variable equal to 1 for each treatment condition, and 0 for control. Regression models include pre-treatment controls selected using Lasso-based post-double selection. Those controls are high-dimensional interactions among the following variables: number of friends and followers, past level of activity, past hate speech and toxicity, account age, language, whether a location or description is indicated, whether the account uses a real name or visible picture, whether the user is a politician, a newspaper or a politically interested user. Those interactions do not carry substantive meaning, and hence are not reported. *BH Adj. P-values* indicates p-values adjusted using the Benjamini- Hochberg (1995) procedure. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

reports results for the subsample of users who had a number of followers below the median before the experiment. Panel B reports results for the users above the median.

Appendix Table E.14 reports the heterogeneous treatment effect across users whose account is young or old, measured as the number of days since the creation of the account.

Table E.8: Disaggregated treatments, OLS estimates

Dependent Variable:	Original Hate Tweet Deleted	No. of hate tweets	Share of daily hate	Probability of Hate Tweets
Alert: Banning of Hate Speech	0.193**	0.096	-0.039	-0.078
	(0.089)	(0.066)	(0.068)	(0.068)
Observations	893	893	893	886
R-squared	0.007	0.003	0.000	0.001
Alert: Humor	-0.016	0.045	0.005	-0.037
	(0.057)	(0.091)	(0.078)	(0.075)
Observations	873	873	873	866
R-squared	0.000	0.000	0.000	0.000
Warning of Consequences	0.113	0.033	-0.096	-0.114
	(0.081)	(0.064)	(0.073)	(0.070)
Observations	868	868	868	857
R-squared	0.003	0.000	0.002	0.003
Empathy: Perspective Taking	-0.066	0.006	-0.063	-0.119*
	(0.046)	(0.057)	(0.067)	(0.072)
Observations	861	861	861	851
R-squared	0.002	0.000	0.001	0.003
Empathy: Perspective Getting	-0.038	-0.006	-0.025	0.037
	(0.053)	(0.055)	(0.073)	(0.077)
Observations	852	852	852	842
R-squared	0.001	0.000	0.000	0.000

Notes. This table reports the results of regressing the dependent variables (as indicated at the top of the table) on a dummy variable equal to 1 for each treatment condition, and 0 for control. The models do not include control variables. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Regression models include pre-treatment controls selected using Lasso-based post-double selection. Panel A reports results for the subsample of users whose account age is below the median before the experiment. Panel B reports results for the users above the median.

Appendix Table E.15 reports the heterogeneous treatment effect across users whose account has low, medium or high anonymity, measured as the presence of a real-sounding name and picture. Regression models include pre-treatment controls selected using Lasso-based

Table E.9: Attrition

Dependent Variable:	Attrition (all sources)	Suspended	Set to Private
Alerting of Hate Speech	-0.014	-0.012	-0.008
	(0.016)	(0.014)	(0.009)
Warning of Consequences	-0.022	0.028	-0.013
	(0.018)	(0.024)	(0.009)
Empathy	0.000	0.026	-0.007
	(0.016)	(0.020)	(0.010)
Observations	2387	2387	2387
R-squared	0.001	0.003	0.001
Mean of DV	0.089	0.051	0.012

Notes. The first column reports the regression of a dummy variable that takes the value of 1 if the observation exited from the sample during the study period (*Attrition (all sources)*), on treatment dummies. In the second column, the dependent variable takes the value of 1 if the observation was *Suspended* by Twitter. In the third column, the dependent variable takes the value of 1 if the account was *Set to Private*. Robust standard errors are in parentheses. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table E.10: Balance

Dependent Variable:	Anonymity	Age	Followers	Friends	No. of daily hate tweets	No. of daily tweets
Alerting of Hate Speech	0.056	0.076	0.012	-0.002	0.008	-0.068
	(0.055)	(0.056)	(0.057)	(0.057)	(0.064)	(0.067)
Warning of Consequences	-0.022	0.118*	0.058	0.119	-0.068	-0.123*
	(0.068)	(0.067)	(0.067)	(0.072)	(0.044)	(0.070)
Empathy	0.045	0.086	-0.026	-0.049	-0.038	-0.110*
	(0.056)	(0.057)	(0.057)	(0.056)	(0.049)	(0.065)
Observations	2387	2260	2260	2260	2259	2259
R-squared	0.001	0.002	0.001	0.003	0.001	0.002

Notes. This table reports the results of regressing the dependent variables (as indicated at the top of the table) on dummy variables equal to 1 for each treatment condition. The control condition is the omitted category. Regressions do not include control variables. All dependent variables are standardized by subtracting the mean and dividing by the standard deviation. Robust standard errors are in parentheses. *, **, *** denote significance at the 10%, 5%, and 1% levels, respectively.

post-double selection. Panel A reports results for the subsample of users whose account has low anonymity, i.e. does not include a real name nor a picture. Panel B reports results for users with medium anonymity, i.e. with either a real name or a picture. Panel B reports

results for users with high anonymity, i.e. with neither a real name nor a picture.

Table E.11: Heterogeneous treatment effects across users with low and high frequency of pre-treatment hate tweets

Dependent Variable:	Original Hate Tweet Deleted	No. of hate tweets	Share of daily hate	Probability of Hate Tweets
Panel A: Below median pre-treatment hate speech				
Alerting of Hate Speech	0.122	-0.009	-0.093	-0.257***
	(0.080)	(0.013)	(0.099)	(0.091)
Observations	601	601	601	594
Warning of Consequences	0.232*	-0.009	-0.185	-0.222**
	(0.132)	(0.018)	(0.120)	(0.111)
Observations	443	443	443	433
Empathy	-0.034	-0.005	-0.192**	-0.161*
	(0.048)	(0.013)	(0.091)	(0.093)
Observations	613	613	613	600
Panel B: Above median pre-treatment hate speech				
Alerting of Hate Speech	0.078	0.094	0.028	0.007
	(0.073)	(0.117)	(0.070)	(0.042)
Observations	622	622	622	621
Warning of Consequences	0.035	0.063	-0.061	0.009
	(0.087)	(0.091)	(0.067)	(0.048)
Observations	425	425	425	424
Empathy	-0.004	0.063	0.031	0.018
	(0.062)	(0.079)	(0.063)	(0.042)
Observations	557	557	557	556

Notes. This table reports the results of regressing the outcome variable on a dummy variable equal to 1 for each treatment condition, and 0 for control. Regression models include pre-treatment controls selected using Lasso-based post-double selection. Those controls are high-dimensional interactions among the following variables: number of friends and followers, past level of activity, past hate speech and toxicity, account age, language, whether a location or description is indicated, whether the account uses a real name or visible picture, whether the user is a politician, a newspaper or a politically interested user. Those interactions do not carry substantive meaning, and hence are not reported. Panel A reports results for the subsample of users who produced a number of hate tweets below the median in the 30 days before the experiment. Panel B reports results for the users above the median. The sample size is indicated with N and reported at the bottom of each subsection. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table E.12: Heterogeneous treatment effects across users with low and high frequency of pre-treatment hate tweets - Interacted models

Dependent Variable:	Original Hate Tweet Deleted	No. of hate tweets	Share of daily hate	Probability of Hate Tweets
Panel A: Alerting of Hate Speech				
Treated	0.087	−0.006	−0.100	−0.182*
	(0.093)	(0.013)	(0.099)	(0.103)
High hate	−0.056	0.672***	0.045	0.067
	(0.074)	(0.067)	(0.098)	(0.097)
Treated × High hate	0.012	0.096	0.153	0.231*
	(0.120)	(0.112)	(0.124)	(0.122)
Observations	1223	1223	1223	1215
R-squared	0.002	0.108	0.005	0.013
Panel B: Warning of Consequences				
Treated	0.185	−0.019	−0.166	−0.266**
	(0.136)	(0.017)	(0.126)	(0.116)
High hate	−0.056	0.672***	0.045	0.067
	(0.074)	(0.067)	(0.098)	(0.097)
Treated × High hate	−0.145	0.091	0.141	0.300**
	(0.161)	(0.119)	(0.145)	(0.138)
Observations	868	868	868	857
R-squared	0.007	0.163	0.005	0.015
Panel C: Empathy				
Treated	−0.094	−0.001	−0.160*	−0.146
	(0.064)	(0.013)	(0.096)	(0.105)
High hate	−0.056	0.672***	0.045	0.067
	(0.074)	(0.067)	(0.098)	(0.097)
Treated ×High hate	0.088	0.029	0.249**	0.223*
	(0.088)	(0.089)	(0.120)	(0.124)
Observations	1170	1170	1170	1156
R-squared	0.002	0.187	0.012	0.011

Notes. This table reports the results of regressing the dependent variables (as indicated at the top of the table) on a dummy variable equal to 1 for each treatment condition, and 0 for control, a dummy variable equal to 1 for users with higher than median pre-treatment hate tweets, and the interaction between the two variables. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table E.13: Heterogeneous treatment effects across users with low and high status users

Dependent Variable:	Original Hate Tweet Deleted	No. of hate tweets	Share of daily hate	Probability of Hate Tweets
Panel A: Below median pre-treatment status				
Alerting of Hate Speech	0.112*	0.069	-0.015	-0.109
	(0.067)	(0.090)	(0.105)	(0.088)
Observations	612	612	612	605
Warning of Consequences	0.310**	-0.011	-0.177	-0.215*
	(0.145)	(0.076)	(0.117)	(0.115)
Observations	413	413	413	403
Empathy	0.042	0.026	-0.141	-0.148*
	(0.055)	(0.054)	(0.096)	(0.089)
Observations	602	602	602	589
Panel B: Above median pre-treatment status				
Alerting of Hate Speech	0.081	0.043	-0.067	-0.107**
	(0.084)	(0.075)	(0.063)	(0.050)
Observations	611	611	611	610
Warning of Consequences	0.020	0.171*	-0.094	-0.017
	(0.081)	(0.090)	(0.066)	(0.060)
Observations	455	455	455	454
Empathy	-0.092*	0.014	-0.043	-0.018
	(0.053)	(0.068)	(0.062)	(0.048)
Observations	568	568	568	567

Notes. This table reports the results of regressing the outcome variable on a dummy variable equal to 1 for each treatment condition, and 0 for control. Regression models include pre-treatment controls selected using Lasso-based post-double selection. Those controls are high-dimensional interactions among the following variables: number of friends and followers, past level of activity, past hate speech and toxicity, account age, language, whether a location or description is indicated, whether the account uses a real name or visible picture, whether the user is a politician, a newspaper or a politically interested user. Those interactions do not carry substantive meaning, and hence are not reported. Panel A reports results for the subsample of users who had a number of followers below the median in the 30 days before the experiment. Panel B reports results for the users above the median. The sample size is indicated with N and reported at the bottom of each subsection. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table E.14: Heterogeneous treatment effects across younger and older accounts

Dependent Variable:	Original Hate Tweet Deleted	No. of hate tweets	Share of daily hate	Probability of Hate Tweets
Panel A: Below median account age				
Alerting of Hate Speech	0.208**	0.033	0.022	-0.076
	(0.088)	(0.104)	(0.091)	(0.068)
Observations	636	636	636	632
Warning of Consequences	0.053	-0.027	-0.199**	-0.159*
	(0.094)	(0.085)	(0.090)	(0.093)
Observations	438	438	438	434
Empathy	-0.003	-0.044	-0.117	-0.137*
	(0.056)	(0.072)	(0.077)	(0.073)
Observations	598	598	598	593
Panel B: Above median account age				
Alerting of Hate Speech	-0.006	-0.003	-0.088	-0.162**
	(0.061)	(0.046)	(0.079)	(0.069)
Observations	587	587	587	583
Warning of Consequences	0.196*	0.121	0.019	-0.060
	(0.112)	(0.088)	(0.097)	(0.080)
Observations	430	430	430	423
Empathy	-0.068	0.021	-0.053	-0.018
	(0.048)	(0.044)	(0.085)	(0.073)
Observations	572	572	572	563

Notes. This table reports the results of regressing the outcome variable on a dummy variable equal to 1 for each treatment condition, and 0 for control. Regression models include pre-treatment controls selected using Lasso-based post-double selection. Those controls are high-dimensional interactions among the following variables: number of friends and followers, past level of activity, past hate speech and toxicity, account age, language, whether a location or description is indicated, whether the account uses a real name or visible picture, whether the user is a politician, a newspaper or a politically interested user. Those interactions do not carry substantive meaning, and hence are not reported. Panel A reports results for the subsample of users whose account age was below the median before the experiment. Panel B reports results for the users above the median. The sample size is indicated with N and reported at the bottom of each subsection. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table E.15: Heterogeneous treatment effects across users with different levels of anonymity

Dependent Variable:	Original Hate Tweet Deleted	No. of hate tweets	Share of daily hate	Probability of Hate Tweets
Panel A: Low anonymity				
Alerting of Hate Speech	0.133	-0.110	0.149	0.081
	(0.094)	(0.112)	(0.122)	(0.110)
Observations	248	248	248	247
Warning of Consequences	0.338*	-0.236**	-0.214*	0.041
	(0.197)	(0.118)	(0.105)	(0.106)
Observations	192	192	192	191
Empathy	0.000	-0.069	-0.009	0.037
	(0.000)	(0.097)	(0.072)	(0.114)
Observations	253	253	253	247
Panel B: Intermediate anonymity				
Alerting of Hate Speech	-0.051	0.149	0.060	-0.184*
	(0.086)	(0.092)	(0.087)	(0.109)
Observations	333	333	333	328
Warning of Consequences	0.022	0.115	0.091	0.006
	(0.095)	(0.114)	(0.120)	(0.146)
Observations	243	243	243	235
Empathy	-0.108	0.076	-0.011	-0.222*
	(0.076)	(0.092)	(0.080)	(0.124)
Observations	310	310	310	304
Panel c: High anonymity				
Alerting of Hate Speech	0.161*	0.037	-0.160*	-0.063
	(0.083)	(0.100)	(0.097)	(0.064)
Observations	642	642	642	640
Warning of Consequences	0.093	0.096	-0.228**	-0.191**
	(0.100)	(0.082)	(0.113)	(0.085)
Observations	433	433	433	431
Empathy	-0.008	0.016	-0.161*	-0.026
	(0.056)	(0.067)	(0.092)	(0.070)
Observations	607	607	607	605

*Notes.*This table reports the results of regressing the outcome variable on a dummy variable equal to 1 for each treatment condition, and 0 for control. Models include pre-treatment controls selected using Lasso-based post-double selection, among the following variables and their interactions: number of friends and followers, past level of activity, past hate speech and toxicity, account age, language, whether a location or description is indicated, whether the account uses a real name or visible picture, whether the user is a politician, a newspaper or a politically interested user. Those interactions do not carry substantive meaning. Panel A reports results for users with low anonymity, i.e. who have a real name and picture. Panel B reports results for the users with intermediate anonymity, i.e. who have a real name or a picture. Panel C reports results for the users with high anonymity, i.e. who do not have neither a real name nor a picture. N indicates the sample size. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

F Power and equivalence bounds

Power analysis calculations that account for our sample sizes and R-squared and a range of effect sizes suggest that we are reasonably well powered to detect effects of around 0.1 standard deviations. To avoid the pitfalls of post-hoc power calculations based on imprecisely estimated effects, we estimate power based on the pre-specified minimum and maximum effect sizes as well as the realized N and R-squared for each treatment arm. Table F.16 below presents these calculations.

Results show that for the main analysis using the full sample, we are in fact largely powered to detect effects of around 0.1. The exception to this claim is the empathy treatment, especially its effect on whether users deleted their hate speech tweet; in this case, the small effect we find (-0.02) is substantially smaller than the minimal detectable effect given the sample size and R-squared (0.1).

Furthermore, we estimated an equivalence confidence interval following Hartman and Hidalgo 2018. This analysis estimates an upper bound on the effect size (as a distance from 0), such that we can state with 95% confidence that the absolute value of the true effect is not larger than this bound. This allows us to show that for most of the outcomes in the main analysis, we can rule out effects larger than around ± 0.2 standard deviations (see table F.16). Again, the empathy treatment appears to be the least effective, and we can in fact rule out an effect larger than ± 0.08. In summary, we conclude that the experiment is powered to detect effects of aggregated treatments in the range of 0.1–0.2 standard deviations, as specified in the pre-analysis plan. Moreover, we can rule out true effects of around ± 0.2 standard deviations or larger for most outcomes and treatments. In this sense, the effects observed in the main sample are relatively small, but they are not an artifact of low power.

For the exploratory subgroup analyses, the sample sizes are smaller. For the subgroup of users who use hate speech infrequently (below-median-hate), despite the small sample size, effects estimated for that subgroup are above the minimum detectable effect (as above, the minimum detectable effect is estimated starting from the analysis sample size and R-squared).

Table F.16: Power analyses

Sample	Treatment	Outcome	Estimated effect	Minimal detectable effect	Power based on effect 0.1	Power based on effect 0.2	Equivalence CI (±)
Full sample	Alert	Hate Tweet Deleted <12h	0.11	0.10	0.76	1.00	0.20
Full sample	Warning	Hate Tweet Deleted <12h	0.14	0.12	0.61	0.99	0.27
Full sample	Empathy	Hate Tweet Deleted <12h	-0.02	0.10	0.78	1.00	0.08
Full sample	Alert	Probability of Hate Speech	-0.12	0.09	0.88	1.00	0.20
Full sample	Warning	Probability of Hate Speech	-0.12	0.11	0.75	1.00	0.22
Full sample	Empathy	Probability of Hate Speech	-0.08	0.09	0.86	1.00	0.17
Hate speech below median	Alert	Hate Tweet Deleted <12h	0.12	0.14	0.52	0.98	0.26
Hate speech below median	Warning	Hate Tweet Deleted <12h	0.23	0.17	0.38	0.91	0.45
Hate speech below median	Empathy	Hate Tweet Deleted <12h	-0.03	0.12	0.62	1.00	0.11
Hate speech below median	Alert	Probability of Hate Speech	-0.26	0.14	0.54	0.98	0.41
Hate speech below median	Warning	Probability of Hate Speech	-0.22	0.16	0.42	0.94	0.40
Hate speech below median	Empathy	Probability of Hate Speech	-0.16	0.14	0.53	0.98	0.31
Hate speech above median	Alert	Hate Tweet Deleted <12h	0.08	0.15	0.46	0.96	0.20
Hate speech above median	Warning	Hate Tweet Deleted <12h	0.04	0.19	0.31	0.83	0.17
Hate speech above median	Empathy	Hate Tweet Deleted <12h	-0.00	0.17	0.39	0.92	-
Hate speech above median	Alert	Probability of Hate Speech	0.01	0.10	0.81	1.00	0.06
Hate speech above median	Warning	Probability of Hate Speech	0.01	0.12	0.66	1.00	0.08
Hate speech above median	Empathy	Probability of Hate Speech	0.02	0.10	0.79	1.00	0.08

However, this analysis and these expectations about effect sizes were not pre-specified, so we emphasize that they should be taken as exploratory.

For the subgroup of users who use hate speech more frequently (above-median-hate), we do not find significant effects, and the effects we do find are not larger than the minimum detectable effects suggested by the sample size and R-squared. However, due to the small sample size, these null effects are not very precisely estimated (especially for the effects on deleting the original hateful tweet). This result is compatible with both a null effect or a small effect that we were unable to detect due to low power.

G Targets of hate speech

Research assistants manually coded the target group for each original hate tweet that is part of the experimental sample. We prespecified the following categories: gender, age, sexuality, religion, nationality, disability, socioeconomic status, politics, and other. Figure G.8 displays the share of hate tweets in the experimental sample that target each group. Most tweets are politically motivated, followed by those referencing socioeconomic status and nationality.

To further delineate the characteristics of our study subjects, we present descriptive statistics stratified by target groups of hate speech (see Table G.17). The analysis incorporates metrics including anonymity scores, follower and friend counts, verification status, status counts, and account age.

Generally, the differences observed between target groups compared to overall averages are modest and often statistically insignificant. However, there are notable exceptions. Users targeting groups associated with disabilities demonstrate significantly higher anonymity and younger accounts, suggesting a preference for anonymity potentially driven by concerns about social backlash or identification risks. In contrast, users who target groups related to appearance have significantly fewer friends, which implies that they may occupy more peripheral

positions within their social networks. Other groups defined by politics, nationality, and religion show a minimal deviation from the overall mean across these metrics, indicating broadly similar user characteristics.

Tables G.18, G.19, G.20, and G.21 present heterogeneous treatment effects across the four main outcomes, based on whether the original hate tweet targets any of the specified groups. While the analysis may be underpowered to detect interaction effects, and results should be interpreted with caution, no clear pattern emerges.

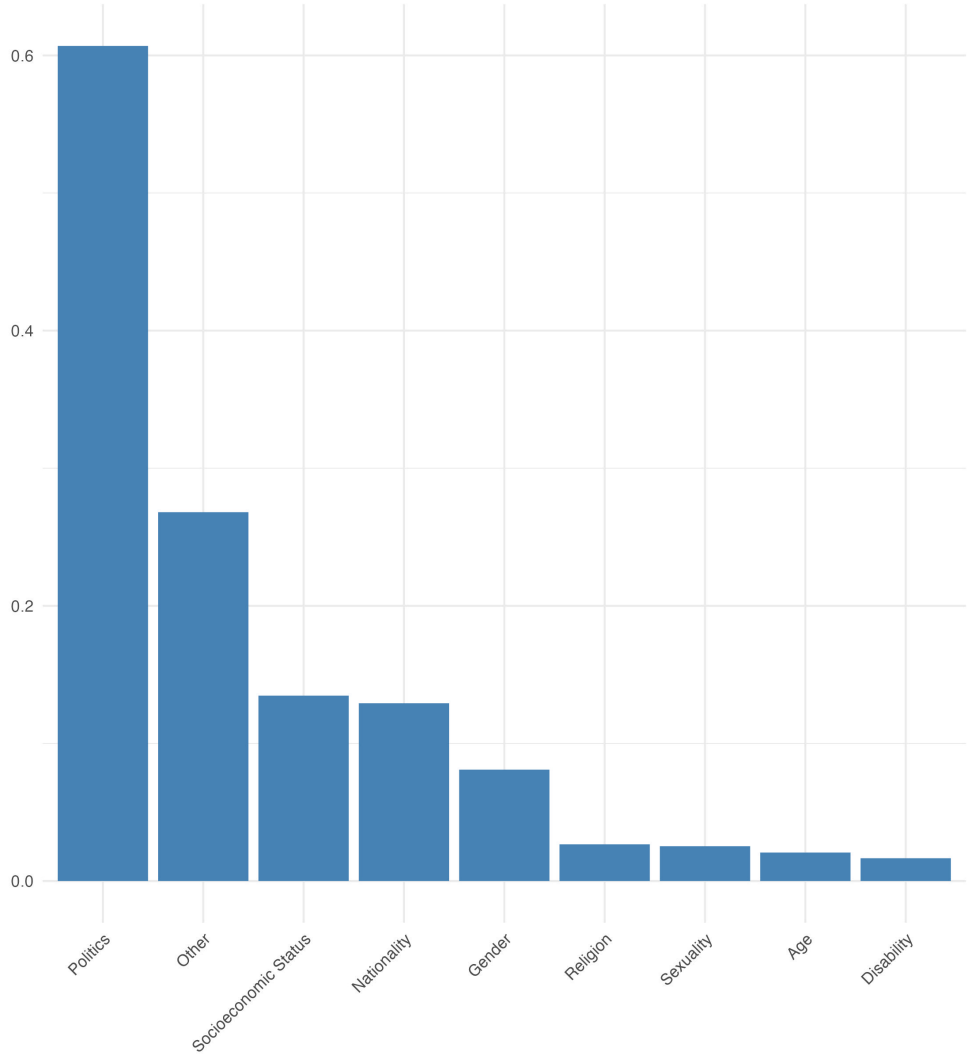


Figure G.8: Shate of hate tweet directed towards each group

Table G.17: Summary statistics of Twitter account attributes by target group. The table shows the mean and standard deviation (SD) for each variable (e.g., Proportion of Verified Accounts, Personal Identification Score, logarithm of followers, friends, status count, and account age in days). The Difference (Δ) column reports the deviation of the group mean from the overall mean (All Groups). For group comparisons (other than All Groups), a two-sample t-test was performed and the corresponding p-value and significance flag (at the 0.05 level) are presented. Cells with "—" indicate that significance testing was not applicable.

Target Group	Variable	Mean	SD	Δ	p-value	Significant
All Groups	Proportion of Verified Accounts	0.000	0.000	—	—	—
All Groups	Proportion of Verified Names	0.680	0.467	—	—	—
All Groups	Username Entropy Score	2.999	0.384	—	—	—
All Groups	Personal Identification Score	1.025	1.648	—	—	—
All Groups	Profile contains a Country	0.016	0.126	—	—	—
All Groups	log(Followers Count)	2.315	0.681	—	—	—
All Groups	log(Friends Count)	2.605	0.623	—	—	—
All Groups	log(Status Count)	3.607	0.761	—	—	—
All Groups	No. of lists user is a member of	11.276	28.766	—	—	—
All Groups	Account Age (days)	2109.740	1538.305	—	—	—
Sex	Proportion of Verified Accounts	0.000	0.000	0.000	—	—
Sex	Proportion of Verified Names	0.693	0.462	0.014	0.682	No
Sex	Username Entropy Score	2.988	0.366	-0.011	0.692	No
Sex	Personal Identification Score	1.116	1.583	0.09	0.438	No
Sex	Profile contains a Country	0.030	0.171	0.014	0.258	No
Sex	log(Followers Count)	2.317	0.701	0.002	0.970	No
Sex	log(Friends Count)	2.597	0.638	-0.008	0.861	No
Sex	log(Status Count)	3.637	0.784	0.030	0.596	No
Sex	No. of lists user is a member of	10.543	22.629	-0.773	0.664	No
Sex	Account Age (days)	2063.317	1521.991	-46.424	0.677	No
Politics	Proportion of Verified Accounts	0.000	0.000	0.000	—	—
Politics	Proportion of Verified Names	0.705	0.456	0.026	0.092	No
Politics	Username Entropy Score	2.984	0.394	-0.015	0.243	No
Politics	Personal Identification Score	1.105	1.643	0.008	0.144	No
Politics	Profile contains a Country	0.017	0.129	0.001	0.864	No
Politics	log(Followers Count)	2.326	0.687	0.011	0.620	No
Politics	log(Friends Count)	2.621	0.601	0.016	0.423	No
Politics	log(Status Count)	3.593	0.778	-0.014	0.585	No
Politics	No. of lists a user is member of	11.42	26.885	0.144	0.874	No
Politics	Account Age (days)	2185.425	1552.219	75.685	0.140	No
Nationality	Proportion of Verified Accounts	0.000	0.000	0.000	—	—
Nationality	Proportion of Verified Names	0.675	0.469	-0.004	0.885	No
Nationality	Username Entropy Score	3.013	0.383	0.014	0.559	No
Nationality	Personal Identification Score	0.891	1.59	-0.135	0.163	No
Nationality	Profile contains a Country	0.02	0.14	0.004	0.654	No
Nationality	log(Followers Count)	2.324	0.669	0.010	0.813	No
Nationality	log(Friends Count)	2.654	0.621	0.049	0.191	No
Nationality	log(Status Count)	3.650	0.753	0.043	0.340	No
Nationality	No. of lists a user is member of	10.344	24.584	-0.932	0.537	No
Nationality	Account Age (days)	2029.914	1487.418	-79.826	0.376	No

Continued on next page

Table G.17 – *Continued from previous page*

Target Group	Variable	Mean	SD	Δ	p-value	Significant
Social Status	Proportion of Verified Accounts	0.000	0.000	0.000	—	—
Social Status	Proportion of Verified Names	0.657	0.476	-0.023	0.425	No
Social Status	Username Entropy Score	3.050	0.357	0.051	0.019	Yes
Social Status	Personal Identification Score	1.04	1.573	0.014	0.881	No
Social Status	Profile contains a Country	0.007	0.081	-0.01	0.067	No
Social Status	log(Followers Count)	2.325	0.679	0.010	0.800	No
Social Status	log(Friends Count)	2.604	0.651	-0.001	0.981	No
Social Status	log(Status Count)	3.646	0.709	0.040	0.355	No
Social Status	No. of lists a user is member of	12.967	36.496	1.691	0.434	No
Social Status	Account Age (days)	2022.970	1600.286	-86.770	0.367	No
Others	Proportion of Verified Accounts	0.000	0.000	0.000	—	—
Others	Proportion of Verified Names	0.647	0.478	-0.033	0.132	No
Others	Username Entropy Score	2.976	0.396	-0.023	0.208	No
Others	Personal Identification Score	1.005	1.707	-0.02	0.796	No
Others	Profile contains a Country	0.018	0.132	0.001	0.807	No
Others	log(Followers Count)	2.276	0.674	-0.039	0.209	No
Others	log(Friends Count)	2.585	0.595	-0.020	0.467	No
Others	log(Status Count)	3.571	0.757	-0.035	0.307	No
Others	No. of lists a user is member of	11.489	35.642	0.213	0.893	No
Others	Account Age (days)	2162.104	1551.266	52.364	0.460	No
Age	Proportion of Verified Accounts	0.000	0.000	0.000	—	—
Age	Proportion of Verified Names	0.723	0.451	0.043	0.455	No
Age	Username Entropy Score	3.069	0.341	0.070	0.109	No
Age	Personal Identification Score	0.969	1.639	-0.056	0.768	No
Age	Profile contains a Country	0	0	-0.016	0	Yes
Age	log(Followers Count)	2.395	0.655	0.081	0.330	No
Age	log(Friends Count)	2.607	0.689	0.002	0.978	No
Age	log(Status Count)	3.626	0.743	0.020	0.832	No
Age	No. of lists a user is member of	10.908	17.477	-0.368	0.869	No
Age	Account Age (days)	2050.477	1584.734	-59.263	0.766	No
Disability	Proportion of Verified Accounts	0.000	0.000	0.000	—	—
Disability	Proportion of Verified Names	0.672	0.473	-0.007	0.909	No
Disability	Username Entropy Score	3.043	0.374	0.044	0.378	No
Disability	Personal Identification Score	0.586	1.864	-0.439	0.08	No
Disability	Profile contains a Country	0.017	0.131	0.001	0.949	No
Disability	log(Followers Count)	2.289	0.704	-0.026	0.784	No
Disability	log(Friends Count)	2.512	0.753	-0.093	0.355	No
Disability	log(Status Count)	3.623	0.696	0.016	0.864	No
Disability	No. of lists a user is member of	7.483	13.521	-3.793	0.044	Yes
Disability	Account Age (days)	1603.069	1356.280	-506.671	0.007	Yes
Appearance	Proportion of Verified Accounts	0.000	0.000	0.000	—	—
Appearance	Proportion of Verified Names	0.662	0.476	-0.017	0.752	No
Appearance	Username Entropy Score	2.994	0.338	-0.005	0.907	No
Appearance	Personal Identification Score	0.7	1.61	-0.325	0.078	No
Appearance	Profile contains a Country	0	0	-0.016	0	Yes
Appearance	log(Followers Count)	2.173	0.654	-0.141	0.060	No
Appearance	log(Friends Count)	2.432	0.750	-0.173	0.044	Yes
Appearance	log(Status Count)	3.491	0.801	-0.115	0.206	No

Continued on next page

Table G.17 – *Continued from previous page*

Target Group	Variable	Mean	SD	Δ	p-value	Significant
Appearance	No. of lists a user is member of	11.287	31.662	0.012	0.997	No
Appearance	Account Age (days)	1927.612	1550.971	-182.128	0.303	No
Religion	Proportion of Verified Accounts	0.000	0.000	0.000	—	—
Religion	Proportion of Verified Names	0.637	0.484	-0.042	0.444	No
Religion	Username Entropy Score	3.049	0.354	0.050	0.215	No
Religion	Personal Identification Score	0.825	1.734	-0.200	0.031	No
Religion	Profile contains a Country	0.025	0.157	0.009	0.617	No
Religion	log(Followers Count)	2.405	0.685	0.090	0.250	No
Religion	log(Friends Count)	2.625	0.699	0.020	0.801	No
Religion	log(Status Count)	3.765	0.731	0.158	0.060	No
Religion	No. of lists a user is member of	8.95	14.004	-2.326	0.162	No
Religion	Account Age (days)	1894.938	1412.695	-214.803	0.184	No
Sexuality	Proportion of Verified Accounts	0.000	0.000	0.000	—	—
Sexuality	Proportion of Verified Names	0.590	0.495	-0.090	0.117	No
Sexuality	Username Entropy Score	3.056	0.345	0.057	0.153	No
Sexuality	Personal Identification Score	0.987	1.694	-0.038	0.845	No
Sexuality	Profile contains a Country	0	0	-0.016	0	Yes
Sexuality	log(Followers Count)	2.332	0.679	0.017	0.829	No
Sexuality	log(Friends Count)	2.539	0.643	-0.066	0.372	No
Sexuality	log(Status Count)	3.625	0.711	0.019	0.818	No
Sexuality	No. of lists a user is member of	11.718	21.954	0.442	0.862	No
Sexuality	Account Age (days)	2058.346	1317.862	-51.394	0.736	No

Table G.18: Heterogeneous treatment effects on *Original hate tweet deleted* across groups targeted by hate speech

Moderator (Targeted group):	Gender	Age	Sexuality	Religion	Nationality	Disability	Socioeconomic status	Politics	Other
Alerting of hate speech	0.089	0.093	0.093	0.094	0.082	0.079	0.051	0.179**	0.087
	(0.063)	(0.061)	(0.062)	(0.061)	(0.062)	(0.059)	(0.064)	(0.091)	(0.073)
Warning of consequences	0.128	0.116	0.114	0.118	0.151	0.091	0.123	0.163	0.038
	(0.088)	(0.082)	(0.083)	(0.083)	(0.092)	(0.079)	(0.091)	(0.123)	(0.086)
Empathy	−0.054	−0.066	−0.055	−0.053	−0.043	−0.053	−0.076	0.026	−0.049
	(0.047)	(0.043)	(0.045)	(0.045)	(0.047)	(0.045)	(0.048)	(0.061)	(0.058)
Moderator	−0.098**	−0.093**	−0.095**	−0.094**	0.033	−0.094**	−0.106**	0.086	−0.052
	(0.040)	(0.038)	(0.039)	(0.038)	(0.127)	(0.038)	(0.043)	(0.068)	(0.072)
Alerting × Mod	0.060	−0.093	−0.093	−0.094	0.062	1.305	0.281	−0.144	0.012
	(0.162)	(0.061)	(0.062)	(0.061)	(0.204)	(1.517)	(0.176)	(0.121)	(0.127)
Warning × Mod	−0.128	−0.116	−0.114	−0.118	−0.271*	0.663	−0.123	−0.083	0.261
	(0.088)	(0.082)	(0.083)	(0.083)	(0.152)	(0.796)	(0.091)	(0.163)	(0.202)
Empathy × Mod	0.054	0.527	0.055	0.053	−0.078	0.053	0.175	−0.129	−0.006
	(0.047)	(0.477)	(0.045)	(0.045)	(0.130)	(0.045)	(0.111)	(0.085)	(0.080)
Observations	2175	2175	2175	2175	2175	2175	2175	2175	2175
R-squared	0.005	0.006	0.005	0.005	0.006	0.010	0.007	0.005	0.006

Notes. This table reports the results of regressing *Original hate tweet deleted* on a treatment dummies, a dummy variable equal to 1 if the original hate tweet targets a specific group (as indicated in the column name), and their interactions. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table G.19: Heterogeneous treatment effects on *Number of hate tweets* across groups targeted by hate speech

Moderator (Targeted group):	Gender	Age	Sexuality	Religion	Nationality	Disability	Socioeconomic status	Politics	Other
Alerting of hate speech	0.074	0.075	0.066	0.070	0.086	0.069	0.101	0.155**	0.010
	(0.065)	(0.062)	(0.063)	(0.062)	(0.068)	(0.062)	(0.066)	(0.072)	(0.061)
Warning of consequences	0.007	0.027	0.024	0.017	0.025	0.031	0.071	0.173	−0.008
	(0.065)	(0.064)	(0.065)	(0.063)	(0.069)	(0.065)	(0.068)	(0.118)	(0.074)
Empathy	−0.007	−0.004	−0.012	−0.004	0.021	−0.011	0.041	0.046	−0.022
	(0.049)	(0.048)	(0.048)	(0.046)	(0.050)	(0.047)	(0.049)	(0.064)	(0.058)
Moderator	−0.108	−0.082	−0.277***	0.301	0.061	−0.067	0.223*	0.132**	−0.140**
	(0.144)	(0.183)	(0.051)	(0.476)	(0.119)	(0.159)	(0.133)	(0.067)	(0.065)
Alerting × Mod	−0.009	−0.202	0.085	−0.014	−0.109	0.112	−0.216	−0.138	0.231
	(0.173)	(0.213)	(0.118)	(0.512)	(0.150)	(0.539)	(0.176)	(0.114)	(0.180)
Warning × Mod	0.310	0.274	0.236	0.335	0.045	0.076	−0.299	−0.228	0.144
	(0.275)	(0.701)	(0.299)	(0.704)	(0.194)	(0.290)	(0.194)	(0.140)	(0.147)
Empathy × Mod	0.110	0.192	0.390	0.031	−0.174	0.796**	−0.301**	−0.075	0.083
	(0.178)	(0.255)	(0.273)	(0.569)	(0.133)	(0.361)	(0.149)	(0.091)	(0.094)
Observations	2175	2175	2175	2175	2175	2175	2175	2175	2175
R-squared	0.002	0.002	0.002	0.005	0.002	0.003	0.003	0.003	0.003

Notes. This table reports the results of regressing *Number of hate tweets* on a treatment dummies, a dummy variable equal to 1 if the original hate tweet targets a specific group (as indicated in the column name), and their interactions. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table G.20: Heterogeneous treatment effects on *Share of daily hate* across groups targeted by hate speech

Moderator (Targeted group):	Gender	Age	Sexuality	Religion	Nationality	Disability	Socioeconomic status	Politics	Other
Alerting of hate speech	−0.006	−0.013	−0.009	−0.040	−0.009	−0.004	−0.005	0.017	−0.078
	(0.064)	(0.063)	(0.063)	(0.063)	(0.068)	(0.063)	(0.068)	(0.080)	(0.078)
Warning of consequences	−0.068	−0.090	−0.087	−0.125*	−0.105	−0.087	−0.094	−0.021	−0.162*
	(0.077)	(0.074)	(0.074)	(0.073)	(0.081)	(0.074)	(0.077)	(0.101)	(0.085)
Empathy	−0.050	−0.049	−0.040	−0.047	−0.024	−0.041	−0.028	−0.032	−0.089
	(0.062)	(0.062)	(0.061)	(0.062)	(0.067)	(0.061)	(0.066)	(0.075)	(0.079)
Moderator	0.073	−0.149	0.143	−0.081	0.051	0.511	0.041	0.140	−0.263***
	(0.253)	(0.197)	(0.418)	(0.182)	(0.132)	(0.486)	(0.147)	(0.095)	(0.084)
Alerting × Mod	−0.158	−0.239	−0.309	0.850**	−0.069	−0.901*	−0.092	−0.057	0.206
	(0.279)	(0.220)	(0.436)	(0.353)	(0.169)	(0.496)	(0.167)	(0.119)	(0.127)
Warning × Mod	−0.325	−0.246	−0.329	0.887*	0.056	−0.466	−0.011	−0.124	0.236
	(0.270)	(0.212)	(0.451)	(0.475)	(0.175)	(0.547)	(0.252)	(0.143)	(0.163)
Empathy × Mod	0.041	0.225	−0.127	0.106	−0.174	0.008	−0.121	−0.016	0.169
	(0.286)	(0.246)	(0.478)	(0.222)	(0.150)	(0.583)	(0.163)	(0.115)	(0.109)
Observations	2175	2175	2175	2175	2175	2175	2175	2175	2175
R-squared	0.002	0.002	0.001	0.009	0.002	0.004	0.001	0.004	0.005

Notes. This table reports the results of regressing *Share of daily hate* on a treatment dummies, a dummy variable equal to 1 if the original hate tweet targets a specific group (as indicated in the column name), and their interactions. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table G.21: Heterogeneous treatment effects on *Probability of hate tweets* across groups targeted by hate speech

Moderator (Targeted group):	Gender	Age	Sexuality	Religion	Nationality	Disability	Socioeconomic status	Politics	Other
Alerting of hate speech	−0.050	−0.051	−0.054	−0.076	−0.098	−0.054	−0.033	−0.020	−0.091
	(0.064)	(0.062)	(0.062)	(0.062)	(0.066)	(0.062)	(0.066)	(0.096)	(0.070)
Warning of consequences	−0.076	−0.110	−0.101	−0.144**	−0.191**	−0.112	−0.076	−0.028	−0.162*
	(0.073)	(0.070)	(0.071)	(0.070)	(0.075)	(0.071)	(0.074)	(0.117)	(0.084)
Empathy	−0.041	−0.043	−0.037	−0.041	−0.057	−0.052	−0.004	−0.086	−0.057
	(0.065)	(0.064)	(0.063)	(0.064)	(0.069)	(0.063)	(0.069)	(0.098)	(0.075)
Moderator	0.227	−0.157	0.181	0.070	−0.131	0.138	0.164	0.035	−0.186*
	(0.204)	(0.370)	(0.325)	(0.221)	(0.126)	(0.269)	(0.143)	(0.100)	(0.111)
Alerting × Mod	−0.154	−0.364	−0.084	0.647*	0.300*	−0.252	−0.181	−0.061	0.108
	(0.238)	(0.422)	(0.392)	(0.340)	(0.166)	(0.298)	(0.170)	(0.124)	(0.141)
Warning × Mod	−0.493*	−0.131	−0.446	0.828*	0.563***	−0.107	−0.307	−0.140	0.171
	(0.258)	(0.656)	(0.423)	(0.453)	(0.195)	(0.428)	(0.220)	(0.146)	(0.151)
Empathy × Mod	−0.086	0.126	−0.105	−0.047	0.119	0.829	−0.277	0.075	0.064
	(0.256)	(0.422)	(0.564)	(0.304)	(0.159)	(0.618)	(0.172)	(0.128)	(0.140)
Observations	2154	2154	2154	2154	2154	2154	2154	2154	2154
R-squared	0.003	0.003	0.002	0.009	0.006	0.005	0.003	0.003	0.004

Notes. This table reports the results of regressing *Probability of hate tweets* on a treatment dummies, a dummy variable equal to 1 if the original hate tweet targets a specific group (as indicated in the column name), and their interactions. Robust standard errors are reported in parenthesis. *,**, *** denote significance at the 10%, 5%, and 1% levels, respectively.

References

Hangartner, D., G. Gennaro, et al. (2021). "Empathy-based counterspeech can reduce racist hate speech in a social media field experiment". In: *Proceedings of the National Academy of Sciences* 118.50, e2116310118.

Hartman, E. and F. D. Hidalgo (2018). "An equivalence approach to balance and placebo tests". In: *American journal of political science* 62.4, pp. 1000–1013.

Hernández-González, A., J. Madera-Quintana, and A Simón-Cuevas (2024). "Uc-cujae at homo-mex 2024: Detecting hate speech against the lgtb+ community using transformers on imbalanced datasets". In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS. org.*

Kotarcic, A., D. Hangartner, F. Gilardi, S. Kurer, and K. Donnay (2022). "Human-in-the-Loop Hate Speech Classification in a Multilingual Context". In: *Findings of the As-*

sociation for Computational Linguistics: EMNLP. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 7414–7442.

Munger, K. (2017). "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment". In: *Political Behavior* 39.3, pp. 629–649.

Pen, H., N. Teo, and Z. Wang (2024). "Comparative Analysis of Hate Speech Detection: Traditional vs. Deep Learning Approaches". In: *2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, pp. 332–337.

Tomás-Valiente, F. (2025). *Uncertain Performance: How to Quantify Uncertainty and Draw Test Sets When Evaluating Classifiers*. Working Paper, presented at CompText 2025.

Umansky, N., M. Kubli, K. Donnay, F. Gilardi, D. Hangartner, A. Kotarcic, L. Bronner, S. Kurer, and P. Grech (2024). *Enhancing Hate Speech Detection with Fine-Tuned Large Language Models Requires High-Quality Data. OSF Preprint*, available at `https://doi.org/10.31219/osf.io/7kbqt` *(accessed 19 April 2025)*.