WARWICK
ECONOMICS

Can hate speech be banned online?
The effects of shutting down toxic forums on Reddit

Adam Di Lizia and Lily Shevchenko

(This paper also appears as CAGE Discussion paper 790)

February 2026 No: 1599

Warwick Economics Research Papers

ISSN 2059-4283 (online)

ISSN 0083-7350 (print)

Can hate speech be banned online? The effects of shutting down toxic forums on Reddit*

Adam Di Lizia† Lily Shevchenko‡

February 19, 2026

Abstract

Is deplatforming effective in reducing toxicity on social media? To answer this question we study a policy change on Reddit in June 2020 which led to a simultaneous ban of thousands of forums containing hateful content, but not the users of these forums. We use data on the near universe of comments left on Reddit to examine the impact of the ban on user behaviour in a differences-in-differences design. We find that the most active users of banned subreddits comment more after the policy change and substitute to new forums in the weeks after the ban. The increase in activity persists in the long run, but is not associated with higher toxicity: instead, the comments left by affected users outside banned subreddits contain 20% fewer instances of hate speech. We do not find evidence that the policy leads to lower quality of engagement, negative spillover effects or recreation of banned subreddits elsewhere on the platform. Overall, the results suggest that moderation targeting toxic digital spaces can be effective in combating hate speech without lowering user engagement, and thus can be aligned with platforms' incentives.

*We are grateful to Manuel Bagues, Sascha Becker, Mirko Draca, Dean Eckles, Thiemo Fetzer, Thomas Hill, Ro'ee Levy, Juan Morales, Maria Petrova, Ananya Sen, Mateusz Stalinski, Gerhard Toews, Ao Wang, Ekaterina Zhuravskaya, the participants of 2025 NICEP conference, 2025 Warwick PhD Conference, 2025 PhD Conference in HEC, the 2025 PPE Conference in King's College London, 2025 CODE@MIT, as well as the participants of internal seminars in Warwick and LSE for their numerous comments and suggestions. This study was approved by the Humanities and Social Sciences Research Ethics Committee at the University of Warwick. Our work has benefited from access to the Scientific Computing Research Technology Platform at the University of Warwick. We acknowledge the support of the Centre for Advantage in the Global Economy (CAGE), Adam Di Lizia additionally thanks the ESRC for funding this research agenda. All remaining errors are our own.

†Department of Economics, University of Warwick. Email: adam.di-lizia@warwick.ac.uk

‡Department of Economics, University of Warwick. Email: liliia.shevchenko@warwick.ac.uk

1 Introduction

Reducing hate speech online is of great interest to both platforms and policymakers. In addition to harming people exposed to it directly, online hate speech has been shown to contribute to hate crimes by spreading and amplifying extreme opinions (Müller and Schwarz 2020, Müller and Schwarz 2023a). In the last decade, most social media platforms have been introducing progressively stricter moderation policies. However, due to the ease of creating new accounts and anonymity afforded to users online, the direct effects of moderation when it involves account or group bans are hard to evaluate. If being targeted by moderation makes users more extreme or leads them to learn how to evade censorship, then deplatforming could have effects opposite from the intended.

To understand the effects of deplatforming on user behaviour this work leverages a policy change on Reddit, a popular social media platform with over 500 million active monthly users, and the 18^{th} most visited website in the world.[1] All posts on Reddit must be made within topic-specific forums, aptly called *subreddits*. In June 2020, Reddit changed its terms of use to forbid hateful content and immediately banned a large number of *subreddits* content of which violated the new rules. The users of these forums themselves were not suspended, enabling us to observe their posts and comments on the platform before and after the ban.

For the empirical analysis, we use the universe of available Reddit comments in public forums in 2020 from `Pushshift` dataset (Baumgartner et al. 2020), comprising more that 2.1 billion observations. In a differences-in-differences specification, we compare users present in banned *subreddits* (henceforth "treated users") to users on the rest of the platform before and after the ban. Our main aim is to assess the effectiveness of the policy in terms of reduction of hateful content, as well see more generally what effect the restriction in access to a set of toxic "digital spaces" has on the behaviour of affected users.

We find that the ban of forums containing hateful content does not lead their users to leave Reddit. On the contrary, users most active in banned forums leave 10% more comments across the rest of the platform after the policy change, with the effect being driven by users of left-wing banned forums. The results are robust to alternative specifications and in triple-difference design using a designated set of "placebo" banned forums of similar size and patterns of user activity. We find that the increase in activity is sustained in the long run, and is due to commenting more in previously used forums and search of new *subreddits* in the weeks after the ban. We do not find evidence that banned forums are recreated elsewhere on the platform, nor do we find evidence of significant spillover effects on non-treated users.

Furthermore, we find that the ban led to an approximate 20% reduction in average number of slurs (our preferred measure of hate speech) used by the treated users outside

[1] For comparison, *Twitter* is #17, with only a slightly higher monthly user count.

banned forums. The results are robust in specifications including post- and thread-level fixed effects, suggesting that reduction in hate speech is driven by treated users becoming less toxic relative to other users in the same discussions, as opposed to sorting into different *types or topics* of discussions. The reduction is driven by right-wing users leaving fewer homophobic, transphobic and antisemitic comments. Interestingly, the effect does not vary by levels of pre-policy exposure to banned forums (share of comments left in banned *subreddits*), suggesting that the fact of exposure to banned forums matters more than the degree to which user is active in banned forums.

We do not find any negative effects on quality of engagement, proxied by text length, number of links used, probability to quote other users or number of replies to a comment. Thus, our results so far suggest that the policy appears to be effective in reducing hateful content without alienating affected users. Our current work is focused on quantifying the effects on language change beyond the use of hate speech, tracing what the characteristics of forums-substitutes are, and developing a theoretical framework of forum choice.

Related Literature The effects of deplatforming on Reddit have been examined previously in interdisciplinary literature using the same dataset. Cima et al. (2024) look at the impacts of same policy change in 2020, however exclusively on toxicity levels; Chandrasekharan et al. (2017) explore earlier bans of specific *subreddits*. Both papers find that bans of toxic forums lead to reductions in toxicity, but do not differentiate the effects from censoring from change in user behaviour (neither paper excludes banned forums from the analysis). Furthermore, since previous work does not employ counterfactual analysis with fixed effects, most of our results are quantitatively and qualitatively different, have different interpretation and implications for platforms and policy makers. More generally, our work is related to literature on content creation and moderation using Reddit data (He, Hong, and Raghu 2025, Huang, Choi, and Wan 2024, Burtch et al. 2022, and many others).

Within economics, we primarily relate to literature on the impacts of toxicity and hate speech and the effectiveness of content moderation. Within the literature on moderation of individual users or posts, Jiménez Durán, Müller, and Schwarz (2023) leverage a change in a German law regarding fines on social media platforms that host hateful content and find that Twitter is less toxic in response. Horta Ribeiro, Cheng, and West (2023) look at posts just below and above a moderation threshold on Facebook and find moderated users are less likely to be in breach of community guidelines for a time after the policy. Rizzi (2024) finds Twitter's 2020 moderation policy was effective in reducing racist hate speech use, but potentially led to negative spillovers to alternative platforms. As for the literature on moderation of communities of users, Müller and Schwarz (2023b) look at the outcome of banning Trump's Twitter account on the toxicity of his followers and find a negative effect. Thomas and Wahedi (2023) investigate a staggered ban of hate organizations' leaders on

Facebook and find a reduction in hateful content consumed and produced by audiences of these groups.

Our works differs from this literature in multiple ways. Firstly, the policy change that we analyse targeted whole communities of users (forums) but did not involve bans of individuals. Hence, we can distinguish between the effects on users *directly targeted* by the moderation and second-order contagion effects on users exposed to the moderated accounts. Secondly, we use the full history of comments in the six months before and after the policy change. This is important not just due to statistical power or level of comprehensiveness of the analysis, but also due to the fact that it allows us to control for selection into different digital spaces, with different norms and levels of toxicity.[2] Using full history of data enables us to say to which extent the effect on levels of hate speech use are driven by selection into less toxic places vs being less toxic *relative to* other users in the same discussion after a moderation policy. We can thus contribute to the literature by discussing the role of *norms* in the level of hate speech use in online spaces. Additionally, to our knowledge, this is the first paper to contribute with evidence from Reddit to the extensive literature on moderation in economics, largely focused on Twitter and Facebook.

The policy change is also a rare opportunity to study patterns of substitution in response to an exogenous shock to the "media consumption basket" of treated users. This relates us to the more general discussion in economic literature on the determinants of social media use, as well as the tension between user engagement and welfare. Notably, Allcott et al. (2020) suggest that addictive properties of social media lead users to spend more time on Facebook than would be optimal. Most relevantly, Beknazar-Yuzbashev et al. (2025) show that toxic content, while promoting engagement and thus being beneficial for platforms relying on ad shows for revenues, is not necessarily welfare maximising, and that lowering exposure to toxicity can reduce user engagement. Our results suggest that the ban of forums containing hate speech led to *increased* activity of affected users on the rest of the platform and little apparent substitution to alternative social media platforms. These conclusions are not contradictory: firstly, unlike in Beknazar-Yuzbashev et al. (2025) setting, the policy that we study did not affect *all toxic content* consumed by the exposed users on Reddit. It is likely that the substitution to the rest of Reddit that we observe after the policy has similar mechanisms to off-platform substitution observed by Beknazar-Yuzbashev et al. (2025). Our setting allows us to provide additional evidence on what drives this substitution and thus speak to the relationship between toxicity and engagement. Secondly, our work draws sharper distinction between toxic and hateful speech – reduction in toxicity was not the aim of the policy change on Reddit, even though the two are correlated, and similarly to Beknazar-Yuzbashev et al. (2025), we argue that the relationship between hateful content and levels of engagement might not be the same as

[2]On Reddit, this selection is more explicit given separation into *subreddits*, but sorting is as ubiquitous on Twitter and Facebook, since users choose which post and comment to reply to.

the relationship between toxicity and engagement. Overall, our preliminary results point to the conclusion that policies curbing the use of hate speech might not face the same issue of going against platform interests, as policies reducing exposure to toxicity do.

2 Background

Reddit is a primarily English-language social media platform founded in 2005. In 2023, it had approximately 73 million daily active users. Unlike most other major social networks, Reddit is organized not around blogs of individual users, but into forums, called *subreddits*, which are essentially boards of posts tied to a specific topic, such as `r/swimming` or `r/politics`. All activity on Reddit happens within *subreddits*, where users can post (text, images or videos), comment or react (with "upvote" or "downvote") to posts and comments of other users. The sum of upvotes and downvotes is displayed as a score, and posts, as well as comments, can be sorted by the associated score.

Users can visit individual *subreddits* or browse the front page of the website displaying most popular posts across the platform. The front page tends to reflect "general interests of the Internet": discussion of news, sport events and media, trending memes and videos, etc. A user must have an account to post and comment, but forums themselves can impose additional restrictions (f.e., being member of a *subreddit* for X days to be able to post) to prevent unwanted activity (bots, trolling, and spam).

Content moderation on Reddit *Subreddits* are run not by the platform directly, but by volunteer members of the subreddit called *moderators*. Moderators set the rules for behaviour on the forum and have the power to delete posts and ban users violating these rules[3].

Subreddits occasionally get banned by the platform, mostly for illegal activity, such as promoting online piracy. Forums that consistently engage in calls for violence are sometimes shut down as well: f.e., `r/incels` was banned for promoting violence against women. Among politics-related *subreddits* `r/altright` was shut down for publishing private information about a person who attacked a prominent right-wing public figure. However, provided that the content does not break the law or call for violence, Reddit does not ban *subreddits*, even if they contain racist or other generally offensive content. This has drawn considerable criticism, but Reddit consistently takes a stance that supports free speech.

The policy change we are analysing in this work occurred in June 2020, in the wake of George Floyd protests. Many of the moderators of some of the largest *subreddits* called for Reddit to take action and ban offensive *subreddits* across the platform. In response,

[3]In general, Reddit has been consistently promoting strong internal moderation of forums, as opposed to platform-wide policies. As a result, forum rules can be very specific to a *subreddit*: for example, `r/politics` does not allow users to post links to news outlets outside the published set of trustworthy sources.

on June 29th 2020 Reddit changed the terms of service of the site to no longer allow any form of "hateful" content, leading to the immediate ban of around 2000 *subreddits*, some of which had tens of thousands monthly users. Most banned *subreddits* can be classified as ideologically far-right or far-left and were promoting xenophobic, transphobic, misogynistic or other generally toxic content.

3 Data

The main data source for our analysis is `Pushshift`: a publicly available archive of the near universe of comments and posts on Reddit from the creation of the platform in 2005 till 2023, when the official Reddit API, where `Pushshift` used to get the data from, stopped being accessible to general public (Baumgartner et al. 2020).

`Pushshift` data contains the text of each comment, the name of the user who left the comment, *subreddit* (forum) where it was posted, creation timestamp, and identifiers allowing us to pinpoint under which post the comment was left and the relative position of the comment within the discussion thread. We also observe reactions to the comment in the form of score (sum of "upvotes" and "downvotes" left by other users on the comments), as well as some other fields.

Given that large fraction of posts are links or media with little to no text, and that it is the comments under the posts that constitute the majority of the activity on Reddit, we focus on comments as main unit of the analysis. For the majority of exercises, we consider *all comments left on Reddit in year 2020* – that is, six months before and after the policy change. The key advantage of using complete data is the ability to control for post- and thread-level fixed effects, as well as to do granular heterogeneity analysis.

One limitation of the data is that comments left in subsequently banned *subreddits* from second week of April until the ban in June 2020 are not available. The reason is that `Pushshift` saves all comments every 3 days but with a lag of 2 to 3 months. Reddit restricted API access to historical data from banned forums at the time of the policy change, and lagged `Pushshift` data stops at the same time. The advantage of the delay is that, since most activity in a thread happens within short period of time after its creation, we likely have near complete data on all comments in a post as well as close to "final" data on such variables as scores and awards.

Availability of data on comments left in banned *subreddits* in April-June 2020 would have enabled us to make the analysis somewhat richer, but its lack is not a serious issue for the empirical results presented in this paper for two main reasons.

Firstly, as is discussed in more detail in Section 4.1, we *purposefully avoid including comments from banned subreddits* in our analysis to eliminate mechanical effects of the policy on the outcomes of interest, thus comments from banned forums do not enter the analysis directly. Secondly, although lack of data for April-June 2020 does lead to incorrectly assign treated users who only appear in banned forums in that timeframe to the

control, this misattribution cannot have any meaningful impact on the estimates due to the large size of the control group. We also expect that users only commenting in banned forums from April 2020 are much more likely to be "incidental" – engaging in banned forums because of the contemporaneous political events, while our main focus is on the *core* users of banned forums, for whom the exposure to the policy change is the greatest.

Bots Bots play a prominent role on Reddit. Primarily, they serve as automatic moderators, restricting user activity in a forum or thread if spam is detected or if user behaves contrary to *subreddit* rules. They also perform all kinds of helper functions on request: displaying statistics on users and forum, conversions between measurement units, dictionary lookups and many more. To ensure that the results are not affected by observations from bots, we perform standard filtering excluding all comments left by users who comment in two different *subreddits* at the same timestamp at least once in the sample. This leads to the exclusion of about 300 million comments from the dataset.

Text processing We pre-process comment text for the whole dataset to flag and separate quotes, URLs, tables, code blocks and other non-textual content, as well as normalize text syntax, formatting and encoding. These steps ensure that variables such as text length are precise and give us access to additional information, such as number of quotes per comment or the set of links used. We then process the comments using `Spacy` to obtain tokenized and lemmatized text and identify the language of each comment using `langdetect` library.

3.1 Variable Construction

Banned subreddits The full official list of the 2000 forums banned in June 2020 was never published: the platform disclosed the names of the 10 largest banned *subreddits*, for the rest only the first letters of *subreddit* name and average daily number of users are available. We use this information to reconstruct the list of all banned *subreddits* with more than 100 average daily users. We do not expect smaller forums to significantly influence the results since the total sum of their average daily user counts does not exceed that of the single largest banned *subreddit*. The resulting set includes 112 banned forums which we additionally manually label as primarily left- or right-wing according to the political leaning reflected in their content, where such classification is meaningful. This allows us to look at the potential heterogeneity in response to the ban depending on the political preferences of the user base. Figure A.1 shows top-20 largest forums by number of comments. Most banned *subreddits* are right-wing, but left-wing forums are larger (with `r/chapotraphouse` being the main left-wing *subreddit* and also by far the largest banned forum). Right-wing *subreddits* are more heterogeneous in terms of topic, but further splitting them into categories (such as transphobic or containing offensive humour) has no bearing on the results.

Treated users We define "treated users" as all users who ever leave a comment in a subsequently banned *subreddit* in January to March 2020. We additionally construct a measure of continuous treatment or "exposure" of each user to policy change as the average fraction of their activity (comments) in January-March that occurs within the set of banned forums.

Figure A.2 shows the distribution of "exposure" to the policy change across the treated users. Only about 7% of the 189 thousand treated users that we identify never post outside banned *subreddits* in January-March 2020. For those users the value of the exposure is 1. The mass at 1 remains when restricting the sample to users who post at least 50 times in January-March 2020 – that is, it is not an artefact of the data construction.

We further designate users in the 4th quartile by share of their total activity in banned forums as "core users" most exposed to the effects of the policy. To eliminate the concern that selection on share of total comments leads to overweighting of users who post few comments overall, we further restrict the sample of *core* users to those who comment on Reddit at least 50 times within the same timeframe (January to March 2020). The results in the paper are robust to alternative cutoffs.

To obtain a measure of political ideology on user level we calculate the fraction of user comments in left- and right-leaning banned *subreddits*. Users are classified as left-leaning if 95% of their activity occurs within *subreddits* classified by us as left. The vast majority of users (more than 99%) *never* post in both types of forums. This clear separation serves to further validate our classification of forum leaning.

Measures of user activity To understand the effects of the policy on overall volume of user activity we construct number of comments left per week for each user and for each user-*subreddit* pair. Additionally, we also construct the variable reflecting number of new *subreddits* user visits each week, or more precisely, number of *subreddits* that user visits each week for the very first time.[4] The variable allows us to look at whether the ban leads to a search of new *subreddits* by affected users.

Measures of hate speech Since the aim of the policy was to reduce "hateful content", to understand the full effects of the ban we need to have a measure of it. At the same time, "hateful content" is not a concept that can be feasibly summarised with a single measure.

So far, *toxicity* has been the main aspect of hate speech studied in economics of media, with Perspective API as the state-of-the-art measurement methodology (Müller and Schwarz 2023b, Jiménez Durán, Müller, and Schwarz 2023). In this paper, we instead

[4]Since we are limiting ourselves to data from 2020, that, of course, reflects forums that a user has not visited before in year 2020. The variable is still indicative of whether a *subreddit* is not a part of a user's normal "consumption basket", but its values for first weeks of the year are likely overestimated. At the same time, we do not expect this feature of data construction to affect treated and control groups in a different way.

opt for measuring incidence of hate speech using a hand-curated dictionary, for two main reasons.

Firstly, it is not feasible to produce Perspective API estimates for the full sample of more than 1.8 billion comments. As was mentioned earlier, using full history of data is of crucial importance for this analysis and is an essential part of the contribution of this paper. Hate speech estimates for the whole dataset in conjunction with granular fixed effects is what enables us to analyse to what degree the observed changes in hate speech use in response to the policy are driven by a change in behaviour within already used forums vs by sorting into different digital spaces, with different norms and levels of hate speech use.

Secondly, and more importantly, the concepts of "toxicity" and "hateful content" differ in our setting in an important way. Arguably, all hateful content is toxic, but not all toxic content is hateful. Although toxicity in its milder forms can have negative impacts, minimising it is not the direct objective of the ban implemented by Reddit.

Most measures of toxicity, including Perspective API, weigh heavily profane language that would not constitute hate speech – especially by Reddit standards. In Müller and Schwarz (2023b), the authors give an example of toxic tweets and their associated toxicity score. Both "@RandPaul You are the dumbest mother-f*cker" and "@laurenboebert lol, goddamn are you a f*cking imbecile" are given the toxicity score of 0.99 out of 1.[5] These comments are unquestionably toxic, but neither of them contains hate speech. Using toxicity measure like Perspective API to proxy hate speech would lead us to massively overestimate its prevalence (especially given the low incidence of hate speech relative to toxic language), is unlikely to be able to preserve the ordering where "true hate speech" gets consistently higher values that milder toxic language, and, in the end, can lead to drastically skewed empirical results. Measuring hate speech using a dictionary of slurs, in addition to being computationally feasible on a dataset of more than 1.8 billion comments, allows us to focus on direct, observable harms and construct numeric estimates of hate speech in a transparent way.

At the same time, using dictionary methods has its downsides: it cannot account for hateful statements that do not use "hateful vocabulary", potentially biasing empirical estimates, especially if ban evasion is prominent. We still argue that reduction in visible hate speech, as a measure of "strong harm" is important in itself, but more work on devising computationally scalable yet comprehensive measures of hate speech beyond toxicity are needed.

A further consideration is that hate speech is not completely equivalent to hateful content either. Hateful content can take other forms – such as spreading links to misinformation about targeted groups, as one prominent example. Hence, a broader analysis of

[5]They also receive 0.93 out of 1 for severe toxicity. The tweets score very high on all measures except "identity attack", precisely what slurs aims to capture.

what constitutes hateful content is needed as well.

To measure incidence of hate speech per comment, we use the publicly available dataset of hate speech compiled by SurgeAI[6], since in addition to the the keywords and their commonly used online variants it has a classification of profanities into broad types (such as racial slurs, misogynistic language, etc). We manually validate the base dataset, then for every comment we identify each word or n-gram that matches a keyword/key phrase in the dictionary of hate speech either in its raw (tokenized) or lemmatized form. We then manually validate the results to ensure that words that have multiple meanings are not skewing the hate speech estimates per *subreddit*.[7]

Measures of engagement quality To see whether the policy change leads to observable differences not only in where users comment or the use of hate speech, but perhaps more subtle differences in *how* they comment, we compute several other straightforward and computationally feasible measures for each comment in the dataset. These are length of comment text (in number of tokens), number of unique links shared in the comment and an indicator variable for whether comment quotes another comment in the post. All of them intend to proxy quality of user interactions on the platform. We do not argue that they are exhaustive, but expect that they would be able to capture drastic differences in commenting behaviour.

In addition, we also proxy engagement quality with number of all replies and direct replies to a comment, with the idea that changes in the degree to which other users engage with comments by a given user likely reflect changes in user behaviour, including differences that we might not be able to pick up using pure text-based metrics.

Comovement and spillover effects Whether the policy leads to recreation of banned forums elsewhere on the platform is an important question. At the same time, measuring potential comovement of users on a platform with hundreds of thousands forums in a counterfactual framework is a challenging task. We address this problem by constructing indicator variable for whether treated user is replying to another treated user for each comment. The variable allows us to estimate whether the probability that treated users are interacting with each other *outside banned forums* changes after the policy change.

We can utilize the same approach to estimate the potential for spillover effects on non-treated users exposed to "migrants" from banned forums. In this case, we measure probability that a *non-treated* user is replying to a treated one. If there is little change to probability of interacting with a user from banned forums, that would indicate that the

[6]https://github.com/surge-ai/profanity

[7]Validation on *subreddit* level makes the most sense since the vocabulary is largely topic-specific, rather then user-specific, and topic-specific terminology (primarily proper nouns, such as names) accounts for the vast majority of tokens incorrectly labelled as slurs. The problem can be largely circumvented by manually inspecting forums accounting for the bulk of the use of a specific word on the platform or systematically trimming N most common keyword matches by *subreddit* or discussion thread.)

potential for *large-scale* spillover effects is limited, even if such interaction does have an effect on an individual non-treated user who counterfactually would not have been exposed to a treated one absent the policy change.[8]

4 Effects on volume of activity

4.1 Empirical strategy

Our first objective is to establish the effect on the intensive margin of treated user activity outside banned forums. For this purpose, we estimate the following difference-in-differences specification:

$$NumComments_{it} = \alpha_i + \tau_t + \beta Post_t \times ShareInBan_i + \epsilon_{it} \tag{1}$$

Where $NumComments_{it}$ is the number of comments (or logarithm of the number of comments) left by user i in week t across all non-banned *subreddits*. $Post_t$ is a dummy variable equal to one in the weeks after the ban and zero otherwise. $ShareInBan_i$ is the share of user's comments left in banned forums in total number of comments left January-March 2020, measuring the degree of exposure of users to the ban, split into quartiles. α_i and τ_t are user- and week-fixed effects, respectively. $NumComments_{it}$ excludes the comments left by treated users in banned *subreddits*, for a clearer comparison of treated users' activity on the rest of the platform before and after the ban. Thus, β is the approximate percentage change in number of comments after the ban for a user who spent all their time in banned *subreddits* in January-March 2020. Throughout the analysis, standard errors are always clustered at the user level.

Identification In interpreting the results of the estimation, we follow the discussion in Callaway, Goodman-Bacon, and Sant'Anna (2024). We do not expect that "strong parallel trends" hold in this case. Specifically, while the ban itself is exogenous, users do still select into the pre-policy degree of treatment exposure: an increase in $ShareInBan_i$ reflects not only increase in treatment exposure, but also the change in *the type of user*. We expect that different types of users likely have very different characteristics and thus potential responses to the policy. Therefore, the our estimates should be interpreted as ATT on users who select into banned *subreddits*. This is, of course, precisely the effect of interest.

In our preferred specifications, we split continuous exposure to the policy $ShareInBan_i$ into quartiles and do not follow advice in Callaway, Goodman-Bacon, and Sant'Anna (2024) suggesting more parsimonious methods, such as various sieve estimators. The main reason

[8]Online users can be exposed to hate speech not only "randomly", but also through selection into the same digital spaces as toxic users. This sorting is even more prominent on forum-based Reddit, and especially since what we observe is commenting – a much stronger marker of selection than browsing content. For this reason, keeping in mind peer effects is important in the discussion on what "spillovers *due* to the policy" even are.

is that we can clearly observe that parallel pre-trends do not hold for every exposure bin. The results of sieve estimation would thus reflect effects on the policy for some groups and the pre-trend for others. Using quartiles allows us to easily check whether the effect for a given subgroup is driven by the policy and not pre-trends. We argue that, although quartiles cannot capture user heterogeneity perfectly, they do reflect a meaningful difference in policy exposure.

We do not include zeros in the regressions, so the estimates can be thought of as intensive margin effects. Intuitively, we think that it is more appropriate to treat the decision on *whether to post* on the platform in a given week as separate from the decision on *how much* to post. From the point of identification, exclusion of zeroes leads us to ignore the effects due to user drop-off from the platform; if the drop-off due to the policy was large, our estimates would have been upward biased. However, as is shown in Section 4.2, we do not find evidence of mass platform exit. Conservative Lee bounds analysis confirms that omission of zeroes cannot explain our results.

Sampling The equation is estimated both on the full sample and on the sample on users who comment at least 10 or 50 times between January and March 2020, as well as on the sample of users who comment at least 100 times in year 2020. The restriction applies symmetrically to both treated and untreated users. The purpose of the exercise is to ensure that the results are not driven by the artificially higher levels of policy exposure ($ShareInBan_i$) for users who only comment a handful of times. We also argue that the results from this restricted estimation better reflect the effects on *core* users of the platform.

Placebo estimation Importantly, defining the treatment variable based on user activity in the first quarter of the year creates selection pressure. Observing a *treated* user in any given week *also implies* observing them at some point in weeks 1 to 14. The pool of treated users is subject to constant attrition, and as the week of the year increases, the likelihood of observing a treated user who is also a more active poster overall goes up as well.

This is not a severe issue, as the immediate effect of the policy is clearly visible in the raw data (Figure 2). Nonetheless, we create a placebo set of *pseudo-banned subreddits* by matching the banned forums in the sample with similar *subreddits* based on their size (number of comments and users). We then construct the sample of placebo treated users based on the same selection criteria as for true treated (and thus facing the same selection pressure). The effects are estimated in a triple difference design comparing the additional effect of being in a *subreddit* that is banned, compared to being present in a similar (pseudo-banned) *subreddit* that was never banned relative to the control (see Equation 2 below).

$$NumComments_{it} = \alpha_i + \tau_t + \beta Post_t \times ShareInBan_i + \\ \delta Post_t \times ShareInBan_i \times TrueBan_i + \epsilon_{it} \tag{2}$$

Where $ShareInBan_i$ is the share of posts left in either the placebo or banned set of *subreddits* between January and March, and $TrueBan_i$ is 1 if the user is also present in the banned *subreddits*.

4.2 Results

We find that the ban leads to a significant increase in the number of comments left by *core* users on the rest of the platform. Figure 2 shows the effect is clearly visible in the raw data. Triple difference estimates of Equation 2 corroborate the result: Figure 1 presents the event study of the effect on log of number of comments left in each week.

In total, the policy leads to an approximately 5% increase in the number of comments left by *core* users outside banned *subreddits*. The effect is clearly sustained in the long run, suggesting that it is driven by changes in the way users engage on Reddit, rather then by the temporary response to the policy change.

A potential problem is that since the policy was a consequence of the protests after the killing of George Floyd, the effects we observe might be driven by the event itself, not the policy change. However, both raw data and the event studies clearly show that the increase in activity associated with the killing of George Floyd a month prior to the policy change disappears after a week.

Estimates for non-*core* treated users are shown in Figure B.10. As we can see, the ban does not appear to have a significant effect on the subsample of users whose normal Reddit use is "disrupted" by the policy change to a lesser extent. This discrepancy implies that the effect on *core* users is likely driven by substitution of activity in banned forums with the rest of the platform. What this substitution entails will be discussed later in this section.

Platform Exit An important question is whether the policy leads users to leave Reddit altogether, and how platform exit might in turn affect the results. Figure B.1 and Figure B.2 show the fraction of all treated and untreated users last observed in each week of 2020. To mimic the selection criteria for the treated group, we restrict the sample of untreated users to those present between January and March. Surprisingly, there is no observable difference in levels of drop-off from the platform between the treated and control groups. Still, we compute Lee bounds (Lee 2009) to ascertain that platform exit does not change our estimates (Figure B.13).

Left vs. Right Figure 3 shows the results after interacting the explanatory variable with an indicator for whether user comments in right- or left-leaning *subreddits* (see Section 3.1 for details on construction). Since the effect is driven by the *core* users we focus our attention to that group. We find a clear asymmetry across groups: the *core* left-leaning users comment approximately 15% more, with no effect for *core* right-wing users.

If the mechanism is indeed that users substitute banned *subreddits* with other forums, this relationship between the degree of substitution and political leaning might indicate that it is easier to find alternatives for left-wing banned *subreddits* than for the right.[9]

5 Effects on movement across the platform

5.1 Empirical strategy

Our next aim is to understand the effects of the ban on movement of users across the platform. Which characteristics of *subreddits* "attract" treated users who suddenly lost access to their usual forums? There are two classes of possible predictors: forum-specific characteristics (topic, size, level of activity, norms, such as level of toxicity, et cetera) and variables related to individual experience of a user within a forum, such as the degree to which comments by a user were liked or engaged with by members of a forum.

To test which characteristics predict volume of activity of treated users after the policy change, we estimate the following general specification:

$$NumComments_{ist} = \alpha_{is} + \tau_t + \beta Post_t \times ShareInBan_i \times \Gamma_{(i)s} + \epsilon_{ist} \tag{3}$$

Where $\Gamma_{(i)s}$ is a placeholder variable referring to a set of *subreddit* (Γ_s) or *subreddit*-by-user (Γ_{is}) characteristics which we expect might predict platform movement (forum size, previous activity in a *subreddit*, level of toxicity, et cetera). $NumComments_{ist}$ is the (log) number of comments left by user i in subreddit s in week t. Treatment variable $Post_t$ and $ShareInBan_i$ are defined as before. User-subreddit fixed effects (α_{is}) and week fixed effects (τ_t) are included to control for for the way a given user might engage in a specific subreddit and for general patterns in platform use (e.g.,the news cycle).

Section 5.2 presents the results of estimating Equation 3 with share of previous activity of a user in a forum and pre-policy level of hate speech use as predictors. In Section 7 we additionally report the results with subreddit virality and average use of hate speech as heterogeneity terms. In our current work we are estimating the effects for a broader set of measures, such as *subreddit* topic and existing share of treated users; as well as measures related to user experience within a forum (average number of replies and upvotes a user receives in a given forum).

[9]We further breakdown of banned forums into additional categories. We find no additional difference for users of forums containing offensive humour, anti-China *subreddits* or *r/the_Donald* specifically. However, there is a potential effect on transphobic *subreddits*. These results are shown in Figure B.14

Additionally, we estimate a version of Equation 2 with $NumberOfNewSubreddits_{it}$, referring to the number of new subreddits user i visits in week t, as the dependent variable. This measure provides a coarser, but easier for interpretation measure of reallocation of activity across the platform. Positive coefficient on β would indicate that the ban leads to a period of "search" of new forums by treated users.

5.2 Results

To establish whether substitution is driven by increased activity in previously used forums, we estimate Equation 3 with *user* × *subreddit* fixed effects, as well as the effect on the number of new (never commented in before) *subreddits* visited by a user in a week. As before, we focus on *core* users for which we observe increase in activity after the ban. Figure 4 shows both results. It is clear that *core* users both comment more in *subreddits* they used before (Panel A) and search for new forums within the first weeks after the ban (Panel B). Estimating Equation 3 by quartile of previous "exposure" to a *subreddit*[10] shows that the effect for previously used forums is significant for *subreddits* in the second (Panel B) and third (Panel C) quartiles of previous exposure (Figure B.5). Lack of evidence for movement to most preferred *subreddits* (4^{th} quartile) can have a range of explanations, most notably, diminishing utility of interactions within a *subreddit*.

Comovement Our results showing increase in activity of *core* users could be explained by recreation of banned communities elsewhere on Reddit. If such new communities are similar to banned forums in the degree to which they promote hateful content, that would imply that the policy was ineffective at dealing with hate speech. Furthermore, if the communities are reformed in existing *subreddits*, rather then in new ones, they could result in spillover effects on existing users of "invaded" *subreddits*.

To test whether users of banned forums relocate elsewhere on Reddit, we measure probability to reply to a treated user. This measure reflects direct communication between a given user and a treated one. If banned communities are recreated somewhere on the platform, then we should see an increase in probability that a treated user is replying to another treated user relative to probability that a non-treated user replies to a treated one outside banned forums.

Figure 5 shows the results of estimating Equation 4 with probability to reply to a treated user as dependent variable. Panel A displays event study coefficients by week, Panel B shows the results after partialling out linear trend. The raw data is presented in Figure B.4.[11] All three figures show the same dynamics: a sharp increase in the probability that

[10] We take fraction of total comments by user left in a specific *subreddit* and compute quartiles of resulting measure by user.

[11] The linear trend in this case is a feature of the data. Each week some users stop using the platform, in both treated and control groups. Since the treated group, unlike the control, does not attain new users instead, the probability to reply to a treated user decreases mechanically as the total pool of treated users

two treated users are interacting outside banned forums in the week after the policy change and no effect afterwards. The initial "spike" likely reflects users talking *about the ban* in forums such as *r/subredditdrama*. We do not see a sustained increase in the probability that two treated users interact outside banned forums, therefore we do not find evidence that networks of treated users are recreated outside banned forums after the policy.

This result may seem like it contradicts our previous findings that *core* users leave more comments on Reddit after the policy change. However, it's worth considering that Reddit is a very large and active platform: the activity of tens or even hundreds of thousands treated users spread over thousands of forums is a tiny fraction of total comments left by over 20 million users in over 150 million posts comprising our sample.

This also suggests that the potential for spillovers from the policy is limited. Firstly, as can be seen in the raw data shown in Figure B.4, the probability that a non-treated user replies to a treated user does not increase after the policy. Secondly, the very act of breaking up the networks of users in banned *subreddits* over time forces a substantial drop in the probability of encountering these users: no new treated users are "created", and the normal gradual attrition from the platform constantly decreases their total number. This is without taking into account that the behaviour of treated users outside banned forums need not be the same – something that we will talk about in detail in the next section.

6 Use of hate speech

6.1 Empirical strategy

Finally, we estimate the effect of the ban on the incidence of slurs as proxy for level of "hateful content" (see Section 3.1 for a detailed discussion) in the following comment level specification:

$$NumSlurs_{cijst} = \alpha_{is} + \tau_t + \gamma_j + \beta Post_t \times ShareInBan_i + \epsilon_{cijst} \tag{4}$$

Where $NumSlurs_{cijst}$ stands for number of *unique* slurs in comment c by user i under post j in subreddit s and week t. We only count the first use of a slur in any given comment to adjust for likely autocorrelation between instances of the same slur within a comment.[12] As before, $Post_t$ is a dummy variable equal to one in the weeks after the ban and zero otherwise, $ShareInBan_i$ is the share of activity in the banned subreddits in total user's activity in January to March 2020. α_{is} is a user-subreddit fixed effect, τ_t is week-fixed effect. γ_j is a unique fixed effect for the post under which the comment is left.

In an additional specification, we also include a unique fixed effect for each discussion thread under a given post. Due to the large sample (1.8 billion comments), it is not possible

shrinks.

[12] An intuitive way of thinking about this is to distinguish the decision to use a slur in a comment from the decision on how many instances of a slur to use conditional on having decided to use it.

for us to include both α_{is} and γ_j in the same specification, however, both specifications separately produce very similar results. As before, we ensure that the results are robust in the triple differences specification (see Section 4.1 for more details on triple difference estimation).

We do not take logs of number of slurs, since there are many zeroes in the data. As suggested by Chen and Roth (2024), we instead look at the absolute changes and normalize them by the mean of the dependent variable for treated users before the policy change to get the effect sizes. Since the incidence of slurs is low relative to the number of comments, we multiply coefficients by 1000 to obtain effects in "slurs per 1000 comments" for ease of presentation and interpretation; the transformation does not affect the results. We also show that the results are robust to using Poisson estimation (see Notes to Table C.1 for the discussion on the problems with using Poisson for data with large number of groups containing only zeroes).

6.2 Results

Table 1 shows the results of estimating Equation 4 with incidence of slurs as the outcome variable. We document that the policy leads to an approximately 20% reduction in the number of slurs per comment left by treated users relative to the mean. It is important to stress that the effect is not a mechanical consequence of the policy change removing toxic forums, as observations from banned *subreddits* are completely excluded from the analysis.

The results are robust in a Poisson QML estimation (Table C.1); reassuringly, we also do not find any effect on the placebo group (Table C.2). The same pattern emerges in an event study for a difference-in-differences specification on a computationally feasible subsample (Figure C.2), as well as in the raw data (Figure C.1).

Table C.3 disaggregates the results shown in Table 1 by quartile of user activity in banned forums. Interestingly, the effect magnitudes are very similar.

The overall reduction in use of slurs can be driven not only by change in behaviour in previously visited forums, but also by *sorting* into less toxic forums. Using the universe of Reddit data allows us to disentangle the two channels by incorporating ultra-granular fixed effects. Column 2 of Table 1 includes a fixed effect for each user-subreddit pair, Columns 3 and 4 include post- and thread-fixed effects, respectively. We can thus estimate the effect on toxicity, controlling for *where the comment is left* and thus for the topic of the discussion. While the magnitude of the estimates decreases in specifications with more granular fixed effects, it is clear that most of the effect is driven by a behavioural change, not by sorting.

This also immediately rules out the possibility that the decrease in slur use is due to increased *subreddit*-level moderation in response to the ban, as the moderation would affect treated and control users in the same way within the same *subreddit*, and especially within the same post or thread.

Dictionary-based method for proxying hateful content has its limitations. Although the dictionary we use accounts for common alternative spellings of slurs (such as words with letters replaced by numbers or characters, intentional spelling errors, et cetera), there is still possibility that users evade censorship by using more elaborate neologisms which are intended to be just as offensive. Still, we argue that such a massive reduction in well-known forms of abusive language indicates that the policy reduces exposure of users to hateful content to some extent.[13]

More importantly, it is likely that "hateful content" cannot be captured completely by hateful vocabulary, even if reduction in the use of hateful vocabulary is valuable in its own right. In our current work we are developing more comprehensive metrics of hate speech that would allow us to address this limitation.

Types of slurs To understand what drives our results, we further break down the estimates by slur type. Figure 6 shows the results of estimating specification identical to one in Column 2 of Table 1 by slur type and political leaning of the user. For ease of comparison, the coefficients are normalized by the group mean of the dependent variable.

While there is a reduction across all slur categories for the left wing users, the largest relative effects are for the right wing subgroup. Specifically, there is a large fall in incidence of antisemitic, homophobic and transphobic language in comments left by right-wing users as a result of the policy. There is also a small and significant effect on the use of misogynistic language by extreme left wing users.[14]

Engagement quality An additional concern might be that the the reduction in hate speech is due to overall lower levels of engagement by treated users, in which case higher volume of comments might be explained by lower quality of the comments. We construct several proxies of comment-level engagement quality to check whether it is the case. They include text length, indicator whether a comment includes a quote of other user, number of links used in a comment, as well as number of direct and overall replies to a comment as a measure of the degree to which the comment can elicit engagement from other users. Table 3 presents the results of estimating Equation 4 with the set of engagement proxies as outcome variables. Figure D.1 shows raw data. We do not observe any negative effects on any of the measures – if anything, there is a slight increase in text length or probability to quote, although on negligible magnitude. We conclude that there is no evidence that the ban leads to lower engagement quality by treated users.

[13]It's worth keeping in mind that the ban was sudden, and in absence of off-platform interaction, coordinating on new language would be a challenging problem.

[14]The result also serves as a sanity check on the quality of vocabulary data: the relation between political leaning and predominant slur types corresponds to our existing understanding of the types of hateful speech that both groups use.

7 Mechanisms

Substitution of banned forums To summarize, we find that the ban of *subreddits* promoting hateful content did not lead to mass exit of the users of those forums. Instead, we observe that most active users of banned forums comment more on the rest of the platform, with little effect on the users engaged in the banned forums to a lesser extent. As is clear from the event studies, the effect is sustained in the long run, suggesting the results cannot be explained by the initial reaction to the policy change itself. We also do not find evidence that increased activity is associated with lower quality of engagement.

Our interpretation is that the ban leads to substitution of activity from banned forums to the rest of the platform; naturally, the magnitude of substitution for the *core* users of banned forums who are most affected by the policy change is the largest.

What is surprising, is that substitution cannot be explained by movement to the similar forums with high proportion of treated users (Figure A.3 shows 20 large *subreddits* most "exposed" to treated users). There is no evidence of recreation of networks of treated users in new forums either (see Section 5.2 for discussion). The results of estimating substitution effects by average use of hate speech in forums-substitutes, a crude approximation of similarity to banned forums in terms of hate speech use (Figure B.6) are inconclusive: we see a significant effect on movement to forums in the 3rd quartile by hate speech use with no significant effect on the 4rd quartile (although the magnitudes are larger). We also do not find evidence that treated users move to *subreddits* which were popular and highly visible in the week of the policy change (B.7). Overall, it is likely that current measures of heterogeneity do not capture real drivers of substitution.

Our next aim is to develop a measure of pairwise topic and user-network distance between each banned and non-banned *subreddit*. Both variables will help us establish more definitively whether substitution is driven by movement to forums similar to the banned. If we do not find that this is the case, the implication would be that whatever characteristics explain substitution, they are not *inherent* to banned forums and their closest alternatives.[15] If instead we do find that substitutes are in some ways similar to banned forums, then the main question is what explains the associated drastic fall in the use of hate speech.

Changes in language use A different picture emerges in the analysis of the effect of the ban on the use of hate speech. We find a significant decrease in use of slurs, even in the forums that treated users *already commented in before the policy change*, but the coefficients are not drastically different depending on the share of user activity in banned

[15]One plausible candidate for such a characteristic is degree of emotional engagement in a forum: if treated users seek out heated discussions, they might move to *subreddits* discussing, for example, sport, politics or finance, which consistently rank the highest in the degree of engagement they can elicit. Not coincidentally, these are also some of the most toxic forums on the platform.

forums. That is, what appears to matter is the fact of *being exposed to banned subreddits in the first place*, not the degree of exposure.

Robustness of the results in specifications with post- and thread-level fixed effects allows us to definitively rule out that the effect is driven by changes in *subreddit*-level moderation rules (see Section 6.2). Still, there are several potential mechanisms that could explain the reduction in use of hate speech after the ban.

First is that the effect on the use of slurs is the reaction to the policy change itself: affected users moderate their language on the rest of the platform, presumably so that they can avoid potential future bans. This explanation does not appear to be plausible, because affected users had no reason to expect a forum to be banned if it has not been targeted already as the policy was implemented. Moreover, since Reddit targeted forums, as opposed to users, changes in behaviour of individual users would have had little impact on the likelihood that the *subreddit* as a whole would be banned. Empirically, this mechanism is hard to reconcile with the observation that treated users' comments outside banned forums contain *significantly* fewer slurs even before the policy change (Table A.4). In general, learning how to evade censorship as a mechanism for the reduction in the use of hate speech is of higher concern if we expect that banned forums are recreated elsewhere on the platform; our results estimating comovement (see Section 5.2) suggest that this is not the case.

The second potential mechanism is that the effect is due to the changes in composition of content that affected users are exposed to. Namely, if we suppose that the way users interact with others in online spaces is at least partially driven not only by the communication style in a particular discussion thread, but also by the norms and language that users have been exposed to *overall*, then the policy change prevents the norms of banned toxic spaces from "bleeding into" the discussions on the rest of the platform. This mechanism can explain the fact that the decrease in the use of hate speech that we observe is *gradual* and only stabilizes in about 10 weeks after the policy change, as can be seen in event study in Figure C.2 and in the raw data (Figure C.1). It is also corroborated by the observation that we observe larger reductions in hate speech use for the categories of slurs that left- and right-wing users are exposed to the most (see Notes to Table A.5 for the discussion).

We are currently developing a measure of vocabulary composition and language similarity across *subreddits* and users, which is relying on individual-level corpus data and embedding distance between corpora. Looking at language composition in a more holistic way will allow us to establish whether the changes we observe are specific to hate speech use or affect user vocabulary in general. This, in turn, would allow us to better speak to the mechanisms driving the effectiveness of the policy.

Movement to other platforms In our setting, we do not observe users of banned forums on other social media platforms.[16] Even though the level of attrition from the platform after the policy is low, one might be concerned that the decrease in hate speech use we observe on Reddit is due to the fact that discussions which would involve hate speech use are now happening on other platforms – that is, the "aggregate" level of hate speech use across the Internet as a whole does not change.

Such dynamics have been observed in other settings. Literature has shown that the largest right-wing "alternative" platform at the time, *Parler* did see an influx in users in response to Twitter moderation during the George Floyd protests. In contrast, there is no significant increase related to ban on Reddit, nor is the ban mentioned by researchers analyzing all shocks to *Parler* membership (Aliapoulios et al. 2021).[17]

Furthermore, this would mean that treated users move discussions containing hate speech off platform *while also commenting more on Reddit itself.* The implication is that moving specific discussions from Reddit to other platforms also changes users' decision problem regarding non-banned forums on Reddit. This result is hard to explain without assuming that at least part of the utility gained from interactions in banned forums is regained after the policy by engaging more in other forums on Reddit and other platforms are not perfect substitutes to Reddit.

In any case, movement to other platforms does not directly relate to the validity of the empirical results of the paper, only to their interpretation.

Overall, we document that moderation targeting digital spaces containing hateful content can be effective in reducing incidence of hate speech, not only mechanically, through removal of content of these communities, but also via reducing hate speech use by affected users on the rest of the platform. Given that we do not observe that the policy change leads to substantial drop-off from Reddit or lower levels of engagement by affected users, the results are promising from the platform's perspective.

8 Conclusion

This paper examines the effects of restricting toxic online spaces on the behaviour of affected users. Overall, we find that the most active users of banned *subreddits* post more, move to new forums and use hate speech less after the ban of forums promoting hateful content. In that sense, the policy appears to be effective, since the removal of hateful content does not lead to lower engagement, and affected users use less hate speech outside

[16]The primary technical obstacle is inability to match user names across platforms, since, unlike on Twitter or Facebook, users of Reddit do not normally post their real names. Attempting to match accounts of people who might use the same nickname on both Reddit and, say, Parler inevitably introduces measurement error, selection, as well as some ethical concerns.

[17]The only direct example of community transplanting is unique to *r/gendercritical*. This banned *subreddit* attempted to transplant the community to *ovar.it*, a site which closed in 2025. It essentially mimicked the style of Reddit but for purely discussions about trans issues.

the banned *subreddits*. Future work aims to examine the direct effects of the policy in more detail by including more nuanced measures of text toxicity and estimating spillover effects on non-treated users. At the same time, our preliminary conclusions raise more questions about general mechanisms driving social media consumption of politically-extreme users. The ban led to substitution to new forums, persistent increase in activity and change in language of *core* users of affected forums. In our current work we try to provide empirical evidence on the behavioural mechanisms underlying these results.

References

Aliapoulios, Max et al. (2021). "A large open dataset from the Parler social network". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15, pp. 943–951.

Allcott, Hunt et al. (2020). "The welfare effects of social media". In: *American Economic Review* 110.3, pp. 629–676. DOI: 10.1257/aer.20190658. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20190658.

Baumgartner, Jason et al. (2020). "The Pushshift Reddit Dataset". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 14, pp. 830–839. URL: https://arxiv.org/pdf/2001.08435.

Beknazar-Yuzbashev, George et al. (2025). *Toxic content and user engagement on social media: Evidence from a field experiment*. Tech. rep. CESifo Working Paper.

Burtch, Gordon et al. (2022). "How do peer awards motivate creative content? Experimental evidence from Reddit". In: *Management Science* 68.5, pp. 3488–3506.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant'Anna (2024). *Difference-in-differences with a continuous treatment*. Tech. rep. National Bureau of Economic Research.

Chandrasekharan, Eshwar et al. (2017). "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech". In: *Proceedings of the ACM on human-computer interaction* 1.CSCW, pp. 1–22.

Chen, Jiafeng and Jonathan Roth (2024). "Logs with zeros? Some problems and solutions". In: *The Quarterly Journal of Economics* 139.2, pp. 891–936.

Cima, Lorenzo et al. (2024). "The Great Ban: Efficacy and Unintended Consequences of a Massive Deplatforming Operation on Reddit". In: *Companion Publication of the 16th ACM Web Science Conference*, pp. 85–93. URL: https://dl.acm.org/doi/pdf/10.1145/3630744.3663608.

He, Qinglai, Yili Hong, and TS Raghu (2025). "Platform governance with algorithm-based content moderation: An empirical study on Reddit". In: *Information Systems Research* 36.2, pp. 1078–1095.

Horta Ribeiro, Manoel, Justin Cheng, and Robert West (2023). "Automated content moderation increases adherence to community guidelines". In: *Proceedings of the ACM web conference 2023*, pp. 2666–2676.

Huang, Justin T, Jangwon Choi, and Yuqin Wan (2024). "Politically biased moderation drives echo chamber formation: An analysis of user-driven content removals on Reddit". In: *Available at SSRN*.

Jiménez Durán, Rafael, Karsten Müller, and Carlo Schwarz (2023). "The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG". In: *Available at SSRN 4230296*.

Lee, David S (2009). "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects". In: *The Review of Economic Studies* 76.3, pp. 1071–1102.

Müller, Karsten and Carlo Schwarz (Oct. 2020). "Fanning the Flames of Hate: Social Media and Hate Crime". In: *Journal of the European Economic Association* 19.4, pp. 2131–2167. ISSN: 1542-4766. DOI: 10.1093/jeea/jvaa045. eprint: https://academic.oup.com/jeea/article-pdf/19/4/2131/39651047/jvaa045.pdf. URL: https://doi.org/10.1093/jeea/jvaa045.

— (2023a). "From Hashtag to Hate Crime: Twitter and Antiminority Sentiment". In: *American Economic Journal: Applied Economics* 15.3, pp. 270–312. DOI: 10.1257/app.20210211. URL: https://www.aeaweb.org/articles?id=10.1257/app.20210211.

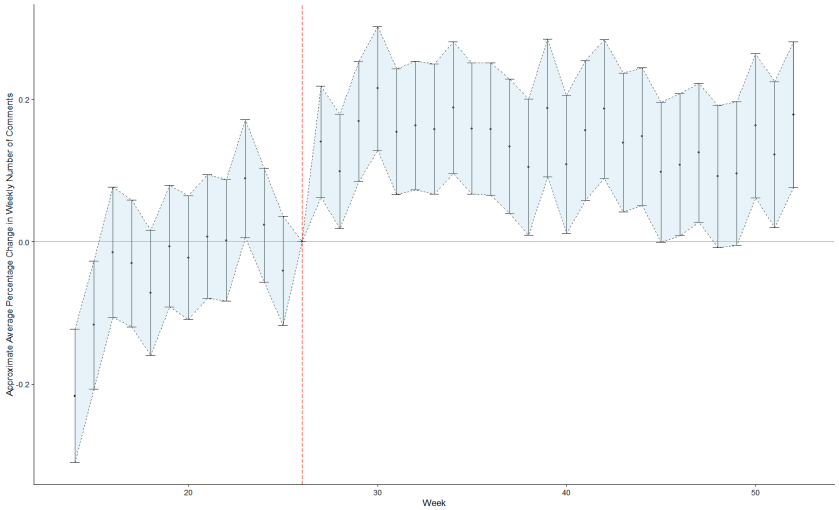
— (2023b). "The Effects of Online Content Moderation: Evidence from President Trump's Account Deletion". In: *Available at SSRN 4296306*.

Rizzi, Marina (2024). "Self-regulation of social media and the evolution of content: a cross-platform analysis". In: *Available at SSRN 5018309*.

Thomas, Daniel Robert and Laila 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.

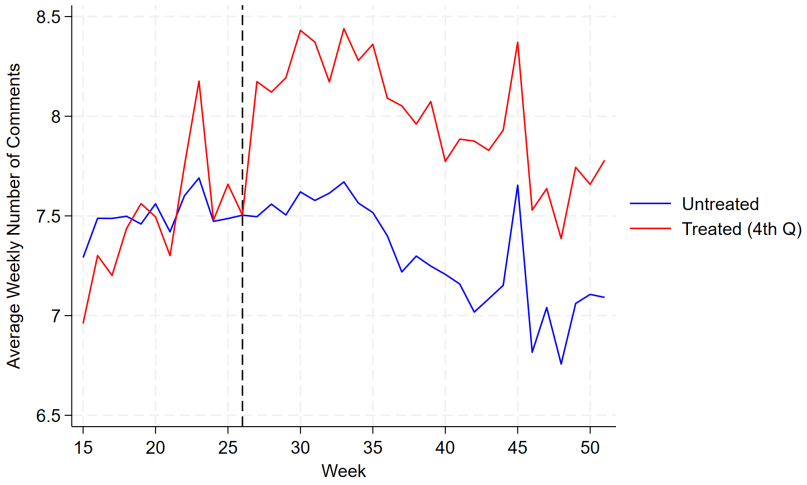
Figures and tables

Figure 1: Effect on log of number of comments left by *core* treated users: event study with sample restriction, triple difference estimates



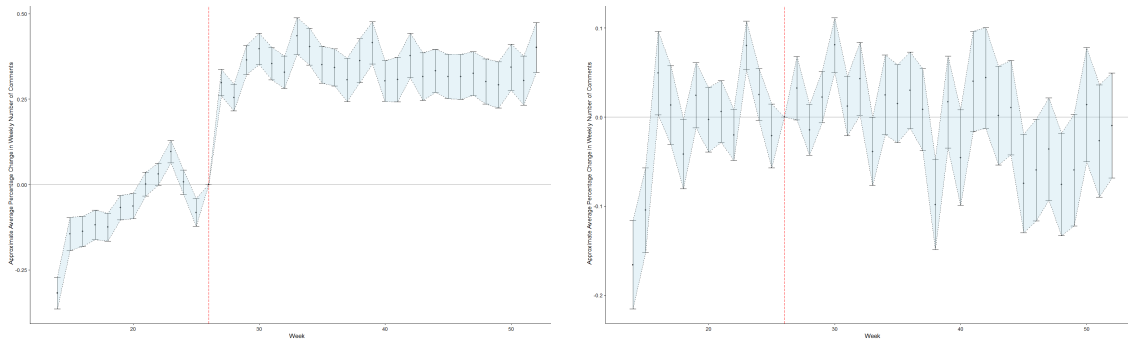
Notes: Figure 1 shows the event study of the effect of the ban on the number of comments left by *core* treated users outside banned forums (Equation 2). Number of comments is taken in logs (there are no zeroes in the data). The sample is restricted to users who comment at least 50 times between January and March 2020 (see Section 3.1 for discussion). Equivalent results with unrestricted sample are presented in Figure B.8 The y-axis represents the approximate average percentage change in number of comments left for a user in the 4th quartile of activity in banned subreddits, and the x-axis denotes weeks. The base category is the first week before the ban; the week of the ban is indicated by the dashed vertical line. Confidence intervals are at 95% coverage, clustering is at the user level.

Figure 2: Average weekly number of comments for *core* users vs. control: raw data



Notes: Figure 2 plots average weekly number of comments left by most active (*core*) users of banned subreddits (in red) and untreated users (in blue) across Reddit, excluding comments left in banned forums. To ensure comparability, the sample is restricted to users who comment at least 10 times between January and March 2020. The dashed line represents the week of the ban. For clarity of presentation, the values for the untreated group are shifted upwards to match the values for the treated in the week before the policy change. The Figure shows large and sustained increase in the number of comments left by *core* treated users after the policy change.

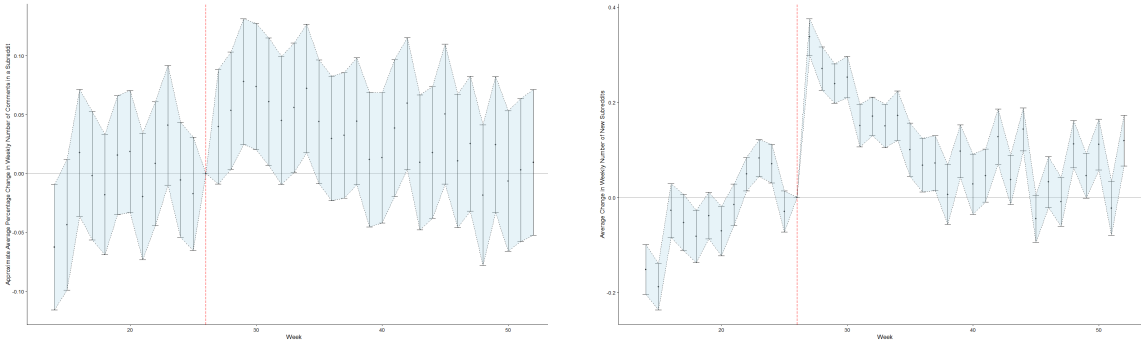
Figure 3: Effect on log of number of comments left by *core* right- and left-wing treated users: event study with sample restriction, triple difference estimates



(a) *Core* left-wing users (b) *Core* right-wing users

Notes: Figure 3 shows the effects of the ban on the weekly number of comments left by *core* left- (a) and right-wing (b) users. The results are produced by estimating Equation 2 separately for left- and right-wing users in the 4th quartile of policy exposure ($ShareInBan_i$). For all groups, the sample is restricted to users who left at least 50 comments on the platform between January and March 2020. Confidence intervals are at 95% coverage, errors are clustered at the user level.

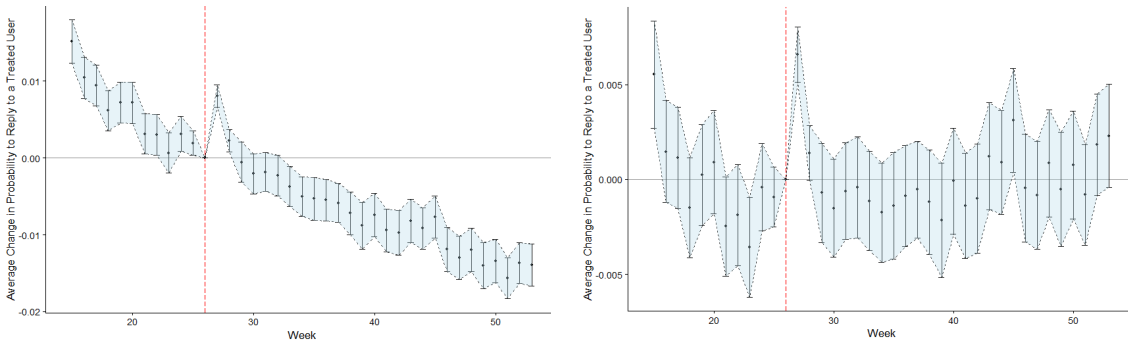
Figure 4: Effect on log of number of comments left by *core* treated users in previously visited subreddits and on the number of new subreddits used: event study with sample restriction, triple difference estimates



(a) Comments in previously used forums (b) Number of new subreddits

Notes: Figure 4 shows the effects of the policy on the number of comments in previously visited forums (a) and the number of new subreddits visited a week (b) for *core* treated users. The results for Panel (a) are produced by estimating Equation 2 with user-subreddit fixed effects. The results for Panel (b) are produced by estimating Equation 2 with number of new *subreddits* visited per week as dependent variable. The sample is restricted to users who left at least 50 comments between January and March 2020. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure 5: Effect on probability to reply to a treated user: event studies with sample restriction, difference-in-difference (detrended) estimates



(a) Raw estimates (b) Detrended estimates

Notes: Figure 5 presents the estimates of the effect on the policy on the probability that a given comment of a treated vs control user is a reply to a comment left by (another) treated user. The results are obtained by estimating comment-level Equation 4 with dummy for whether the comment is a reply to a treated user as dependent variable. The sample is restricted to users who posted at least 50 times between January and March 2020. Panel (a) shows a clear downward trend, driven by normal platform attrition (see Section 4.1 for discussion). Panel (b) presents the effects after partialling out the linear trend. Both panels show a temporary increase in probability that two treated users are interacting outside banned forums after the policy change, but the effect disappears after few weeks. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Table 1: Effects on incidence of hate speech: difference-in-differences estimates

	# of Slurs per 1000 comments			
	(1)	(2)	(3)	(4)
Post × Treated	-0.384***	-0.303***	-0.367***	-0.304**
	(0.057)	(0.104)	(0.068)	(0.127)
Fixed effect	**User**	**User×Subreddit**	**Post**	**Thread**
Weeks	53	53	53	53
Groups	29,359,218	243,757,537	160,111,081	905,773,238
Observations	1,863,841,419	1,863,841,419	1,863,841,419	1,863,841,419
Dependent var. mean	1.86	1.86	1.86	1.86

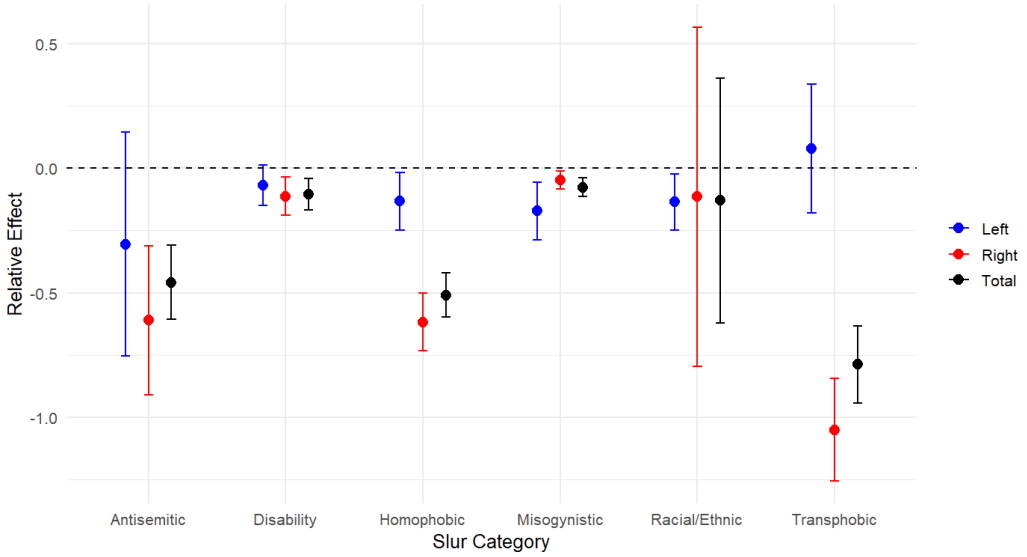
Notes: Table 1 presents the results of estimating Equation 4 across all quartiles by policy exposure. For ease of display the dependent variable is multiplied by 1000. We only count unique words due to probable autocorrelation in use of specific language within a comment. Standard errors in parentheses are clustered at the user level. The effects by quartile of policy exposure are presented in Table C.3. Triple difference estimates of the effects are presented in Table C.2 and as event study in Figure C.2. The results from Poisson QML estimation are presented in Table C.1. The stability of the coefficients in specification with ultra-granular post- and thread-fixed effects suggests that the results are not explained by treated users sorting into discussions with lower average levels of hate speech use.

Table 2: Effects on incidence of hate speech by political leaning of the users: difference-in-differences estimates

	# of Slurs per 1000 Comments			
	(1)	(2)	(3)	(4)
Post × Treated × Left	-0.328***	-0.246***	-0.215***	-0.423***
	(0.065)	(0.059)	(0.062)	(0.102)
Post × Treated × Right	-0.391***	-0.312**	-0.407***	-0.257
	(0.075)	(0.143)	(0.089)	(0.169)
Fixed effect	**User**	**User×Subreddit**	**Post**	**Thread**
Weeks	53	53	53	53
Groups	29,356,555	243,366,189	160,057,485	904,311,937
Observations	1,860,546,041	1,860,546,041	1,860,546,041	1,860,546,041
Dependent var. mean	1.86	1.86	1.86	1.86

Notes: Table 2 presents the results of estimating Equation 4 with dependent variable interacted with dummy for political leaning across all quartiles by policy exposure. For ease of display, the dependent variable is multiplied by 1000. Standard errors are clustered at the user level.

Figure 6: Relative change in incidence of hate speech by political leaning of the users and type of hate speech: difference-in-differences estimates



Notes: Figure 6 shows the effect of the policy on different types of offensive language by political leaning of users. The estimates are produced by interacting the endogenous variable in Equation 4 with dummy for political leaning and indicator and estimating the effects separately for each slur type. The equation includes user-subreddit fixed effects (that is, the same formulation as in Column 2 of Table 2). As the actual incidence of slur varies significantly across types of hate speech, the coefficients are normalised by the dependent variable mean of the associated category. Confidence intervals are at the 95% coverage, clustering is at the user level. As we can see, the reduction in hate speech is driven by right-wing users leaving fewer antisemitic, homophobic, and transphobic comments. The reduction for the left-wing users is primarily due to lower incidence of misogynistic language. Table 2 reflected more similar estimate magnitudes for left- and right-wing users: the difference stems from the fact that average incidence of slurs is lower for left-wing users across all categories.

Table 3: Effects on proxies of engagement: difference-in-differences estimates

	Text Length	**# Links**	**# Direct Replies**	**# Total Replies**
	(1)	(2)	(3)	(4)
Post × Treated	1.380***	-0.000	0.004***	0.265
	(0.181)	(0.001)	(0.001)	(0.185)
Fixed effect	**User×Subreddit**	**User×Subreddit**	**User×Subreddit**	**User×Subreddit**
Weeks	53	53	53	53
Groups	241,525,510	241,525,510	241,525,510	241,525,510
Observations	1,852,554,254	1,852,554,254	1,852,554,254	1,852,554,254
Dependent var. mean	137.96	0.069	0.540	5.60

Notes: Table 3 presents the results of estimating Equation 4 with different proxies for engagement quality as dependent variables (the exact specification is equivalent to the one in Column 2 of Table 1). Standard errors in parentheses are clustered at the user level. Column 1 presents the effects on length of comment text (in number of symbols). Column 2 reflects the effect on the number unique links within comment. "Direct Replies" in Column 3 refers to number of comments left as direct replies to a given comment, "Total Replies" in Column 4 additionally includes all replies to the direct replies (that is, the whole conversation tree under a given comment). The results should be taken with caution, as it is clear from Figure D.1 displaying the raw data that pre-trends are unlikely to always hold. Nonetheless, given the small magnitude of the effects relative to variable means, we can conclude that the policy either does not have an effect on the engagement proxies or these effects are incredibly small.

Appendix

A Descriptive Statistics

Table A.1: Descriptive statistics

	Mean	SD	Min	Median	Max
Subreddits (n=937,450)					
# weeks present	6.360	12.074	1	1	53
Average # of comments per week	49.268	2,300.894	1	2	1,559,300.189
Average # of users per week	15.043	561.810	1	1	380,587.906
Banned subreddits (n=112)					
# weeks present	11.080	4.454	1	14	14
Average # of comments per week	2,793.997	15,543.876	1	58.5	124,657.071
Average # of users per week	389.899	1,571.955	1	24.893	11,244.5
Users (n=29,406,985)					
# of comments per week	63.684	486.155	1	5	894,963
Average # forums visited in a week	2.682	4.154	1	1	4,968
Treated users (n=189,950)					
# of weeks present	9.216	4.526	1	10	14
# of comments	19.052	47.654	1	6	17,846
Avg # forums visited in a week	6.505	13.562	1	3	4,968
# of banned subreddits used	1.202	0.743	1	1	36

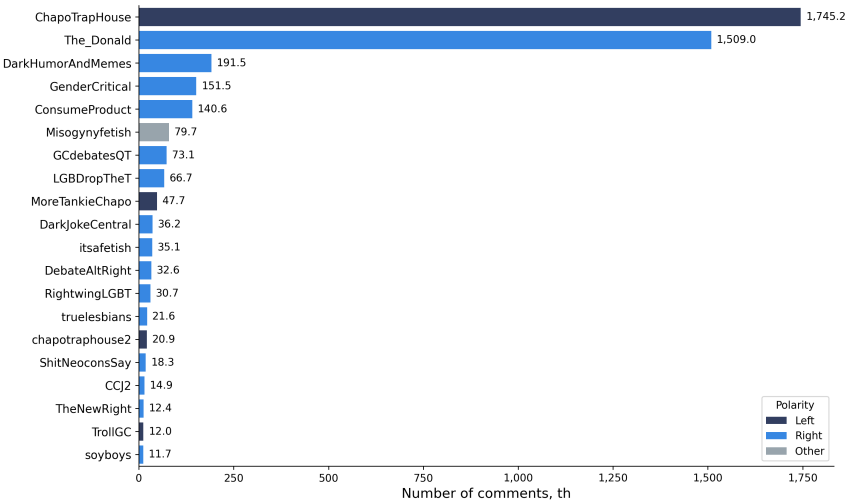
Notes: The table presents descriptive statistics for the near population of comments on Reddit in 2020. The statistics include comments left by bots or deleted accounts. User-level statistics exclude both categories. In total, we observe 2,161,449,965 comments left on the platform in 2020. 10.6% of comments in the database are missing author id: these are removed comments or comments left by deleted accounts.

Table A.2: Descriptive statistics: volume of activity of treated users by political leaning

	N	Mean	SD	Min	Median	Max
Panel A: Subreddits						
Average weekly # of comments						
Left	7	18649.412	46761.276	1	856.428	124657.072
Right	95	1850.096	11195.120	1	50.384	107788.0
Other	10	662.271	1768.312	7.428	91.214	5689.285
Average weekly # of users						
Left	7	1852.404	4100.603	1	273.857	11117.857
Right	95	303.074	1287.456	1	23.785	11244.50
Other	10	190.978	408.602	5.642	57.928	1345.642
Panel B: Users						
# of comments						
Left	50,179	17.211	42.353	1	6	4174
Right	137,050	19.279	44.009	1	7	17846
Other	2,721	38.448	154.029	1	13	11230

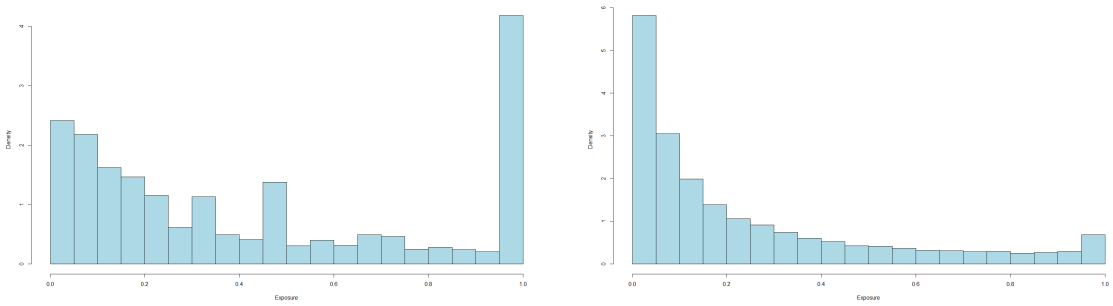
Notes: The table presents descriptive statistics showing volume of activity in banned subreddits by subreddit type (Panel A), as well as volume of activity of treated users depending on whether they post in predominantly left- or right-leaning banned subreddits (Panel B). User is classified as right- or left- leaning if 95% of their activity occurs in the corresponding subreddit type (or in banned subreddits classified as "Other"). Less than 1% of treated users comment in both left- and right-leaning forums.

Figure A.1: Top 20 largest banned subreddits



Notes: Figure A.1 shows number of comments (in thousands) left in 20 largest banned subreddits in January-March 2020 with banned left-wing forums denoted in dark blue and banned right-wing forums denoted in light blue. From 112 identified banned subreddits, 95 can be classified as right-wing, 7 as left-wing. Additionally, 10 forums belong in neither category, primarily containing extreme misogynistic sexual content.

Figure A.2: Distribution of policy exposure across treated users



(a) Baseline (b) 50 Comments Jan-Mar 2020

Notes: Figure A.2 shows the distribution of policy exposure across treated users measured as the fraction of total comments left in subsequently banned *subreddits* by week, averaged over weeks 1 to 14 of year 2020. The y-axis shows the density. Panel (a) shows the distribution of exposure across all treated users. Clustering at specific points is driven by users leaving only few comments throughout the period. Panel (b) shows the histogram of exposure for users who posted at least 50 times across Reddit over weeks 1 to 14 of year 2020. The cutoff ensures that exposure is calculated as a fraction of a large enough total number of posts to be interpretable. Average over week is preferred to simple average over the period to put larger relative weight on repeated engagement in banned forums. For reasons discussed in Section 4.1, average exposure does not take into account zeroes for weeks in which user did not leave any comments on the platform.

Table A.3: Descriptive statistics: incidence of hate speech

	N	Mean	SD	Min	Median	Max
All categories						
Control	1,766,309,688	0.00176	0.0435	0	0	61
Treated	97,531,731	0.00381	0.0660	0	0	61
Left	23,314,974	0.00404	0.0666	0	0	6
Right	70,921,379	0.00372	0.0656	0	0	61
Antisemitic language						
Control	1,766,309,688	0.0000125	0.00358	0	0	11
Treated	97,531,731	0.0000245	0.00524	0	0	11
Left	23,314,974	0.0000202	0.00451	0	0	2
Right	70,921,379	0.0000254	0.00543	0	0	11
Disability-related language						
Control	1,766,309,688	0.000376	0.0195	0	0	3
Treated	97,531,731	0.00103	0.0323	0	0	3
Left	23,314,974	0.000515	0.0228	0	0	2
Right	70,921,379	0.00119	0.0346	0	0	3
Homophobic language						
Control	1,766,309,688	0.000200	0.0144	0	0	20
Treated	97,531,731	0.000518	0.0231	0	0	7
Left	23,314,974	0.000327	0.0183	0	0	3
Right	70,921,379	0.000573	0.0243	0	0	7
Misogynistic language						
Control	1,766,309,688	0.000749	0.0282	0	0	9
Treated	97,531,731	0.00141	0.0388	0	0	7
Left	23,314,974	0.00255	0.0526	0	0	5
Right	70,921,379	0.00105	0.0334	0	0	7
Racist language						
Control	1,766,309,688	0.000385	0.0204	0	0	49
Treated	97,531,731	0.000733	0.0315	0	0	49
Left	23,314,974	0.000566	0.0248	0	0	5
Right	70,921,379	0.000775	0.0333	0	0	49
Transphobic language						
Control	1,766,309,688	0.0000350	0.00599	0	0	4
Treated	97,531,731	0.0000947	0.00984	0	0	3
Left	23,314,974	0.0000596	0.00788	0	0	3
Right	70,921,379	0.000104	0.0103	0	0	3

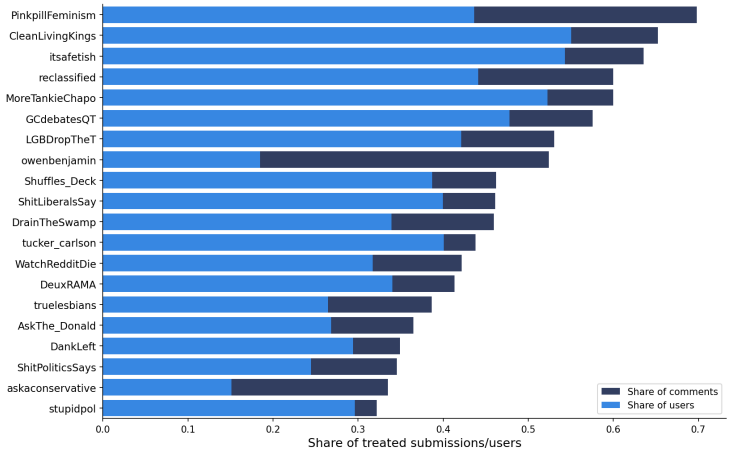
Notes: Table A.3 shows the descriptive statistics on the incidence of hate speech by language type for treatment and control groups, as well as separately for right- and left-leaning users. All values in the table refer to number of unique words of a given category per comment. Comments left by bots and users whose account has been deleted are excluded from the sample. Identical large maximum values in some groups are due to *copypasta* comments.

Table A.4: Descriptive statistics: use of hate speech by treated users in and outside banned *subreddits*

	N	Mean	SD	Min	Median	Max
All treated						
Outside banned	174,439	0.00490	0.03103	0	0	2.00
Inside banned	182,798	0.01271	0.09224	0	0	5.00
Treated 50 comm. cutoff						
Outside banned	12,658	0.00506	0.02230	0	0	0.50
Inside banned	13,327	0.00948	0.03665	0	0	1.19
Left-leaning users						
Outside banned	46,685	0.00692	0.04391	0	0	1.25
Inside banned	48,056	0.02308	0.13201	0	0	3.00
Right-leaning users						
Outside banned	125,130	0.00414	0.02479	0	0	2.00
Inside banned	132,035	0.00904	0.07307	0	0	5.00
Q1 $ShareInBan_i$						
Outside banned	47,451	0.00460	0.01701	0	0	1.00
Inside banned	45,710	0.01256	0.09425	0	0	3.00
Q2 $ShareInBan_i$						
Outside banned	47,413	0.00525	0.02559	0	0	1.24
Inside banned	45,874	0.01351	0.09213	0	0	5.00
Q3 $ShareInBan_i$						
Outside banned	47,314	0.00518	0.03437	0	0	1.33
Inside banned	46,035	0.01339	0.08915	0	0	3.00
Q4 $ShareInBan_i$						
Outside banned	32,261	0.00444	0.04565	0	0	2.00
Inside banned	45,179	0.01139	0.09336	0	0	2.00

Notes: Table A.4 presents descriptive statistics on the use of hate speech in and outside banned forums by treated users. The statistics are based on user-level incidence of hate speech calculated as $\sum_{week=2}^{14} \frac{\#Slurs}{\#Comments}$ or average number of slurs per comment in a week, averaged over weeks 2-14 for which data on activity in banned forums is available. Observations from week 1 are excluded from the sample since week 1 is incomplete (1st Jan 2020 is a Wednesday). The difference in number of observations within groups (N) is due to the fact that not all users post in both banned and non-banned forums in weeks 2-14. It is evident from the data that users are significantly more toxic in banned forums than on the rest of the platform, regardless of the subsample.

Figure A.3: Top 20 non-banned subreddits with largest share of treated users



Notes: Figure A.3 plots top 20 non-banned subreddits with the highest share of comments left by users of banned subreddits. The sample is restricted to subreddits with over 1,000 monthly users. Forums for which the user overlap with the set of banned subreddits is the largest tend to be similar in terms of their topics as well.

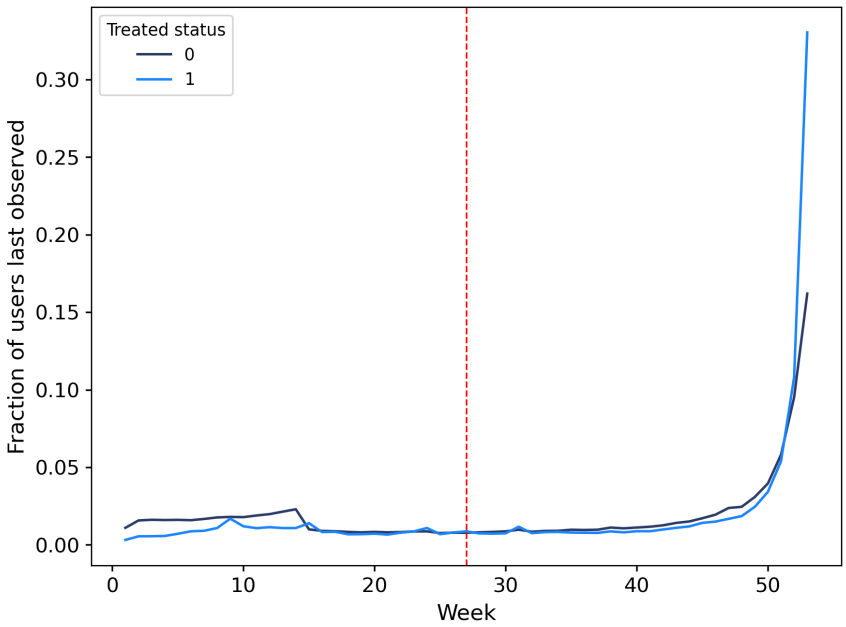
Table A.5: Descriptive statistics: exposure of left- and right-leaning treated users to hate speech in banned subreddits, by category of slur

	Average exposure to hate speech per 1,000 comments					
	Homophobic	*Disability*	*Misogynistic*	*Antisemitic*	*Racial/Ethnic*	*Transphobic*
Left	0.390	0.412	20.377	0.024	0.537	0.046
Right	3.128	2.643	1.389	0.140	3.538	0.614

Notes: Table A.5 presents the statistics on the exposure of right- and left-leaning treated users to hate speech within banned *subreddits*. Specifically, we calculate the average frequency of slurs by categories for each *subreddit*. We then compute exposure to slur category for each user, weighted by the proportion of comments left in a forum. If the ban reduces the use of hate speech through reduction of exposure to certain types of language, we should expect the effects of largest magnitudes for the types of hate speech users are exposed to the most. Figure 6 confirms this is the case: we see significant reductions in the use of antisemitic, homophobic, ableist and transphobic for the right-wing users who also have higher exposure to these hate speech categories within banned *subreddits*. For the left-wing users the largest reduction is in the use of misogynistic language, which is precisely the type of language more common in left-wing banned *subreddits*. Note that since the effect sizes displayed in Figure 6 are normalised by the mean, we do not necessarily expect that higher magnitude of "raw" exposure to a specific category would be associated with larger relative effects for this category.

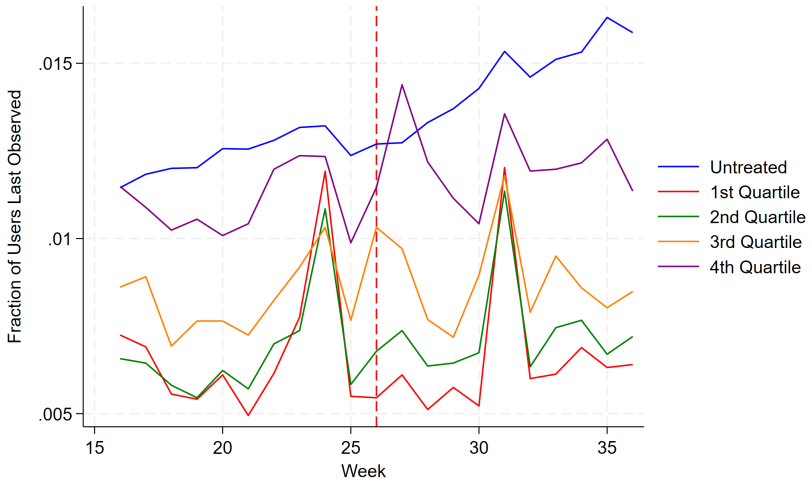
B Volume of activity & movement across the platform

Figure B.1: Weekly drop-off: treated vs non-treated users



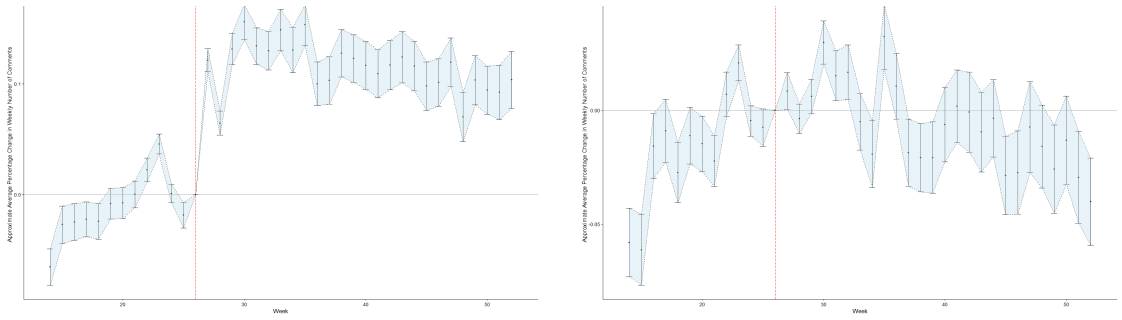
Notes: The figure shows fraction of users in treated and non-treated groups (y-axis) last observed in the corresponding week of the year (x-axis). Week 27 in which policy change has occurred is marked in red. The large uptick towards the end of the sample is mechanical, since the sample ends at week 53. If it is the case that the results are driven by most extreme treated users simply leaving Reddit after the policy change, we would expect higher drop-off after week 27 – this does not appear to be the case.

Figure B.2: Weekly drop-off: treated by quartile of policy exposure vs non-treated users



Notes: Figure B.2 shows fraction of users in treated quartiles and non-treated groups (y-axis) last observed in the corresponding week of the year (x-axis) in the 16 weeks around the policy change. Dashed red vertical line indicates the week of the policy change. The difference in platform exit between 4th quartile (*core* users) and the control group is within 0.5%. Figure B.13 confirms that drop-off from the platform does not explain the empirical results.

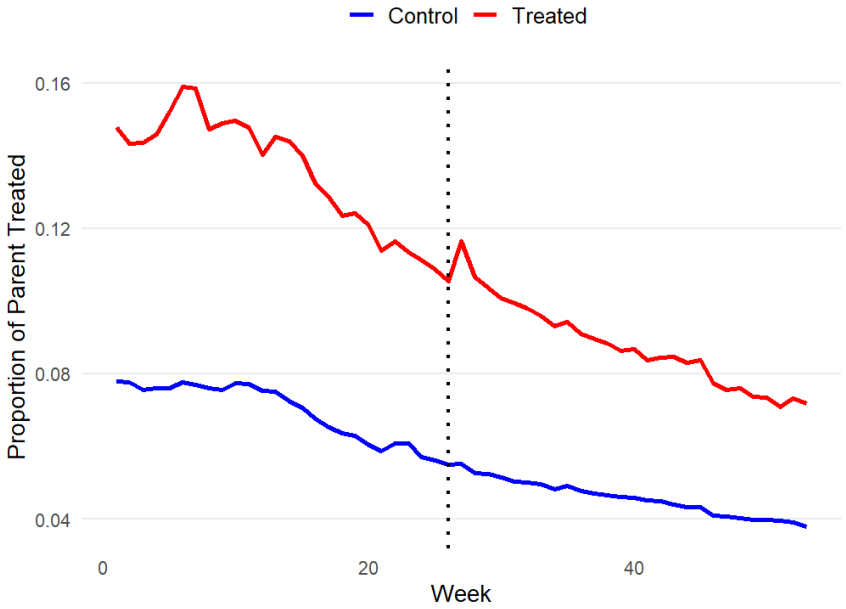
Figure B.3: Effect on log of number of comments left by *core* right- and left-wing treated users: event study, triple difference estimates



(a) "Core" left-wing users (b) "Core" right-wing users

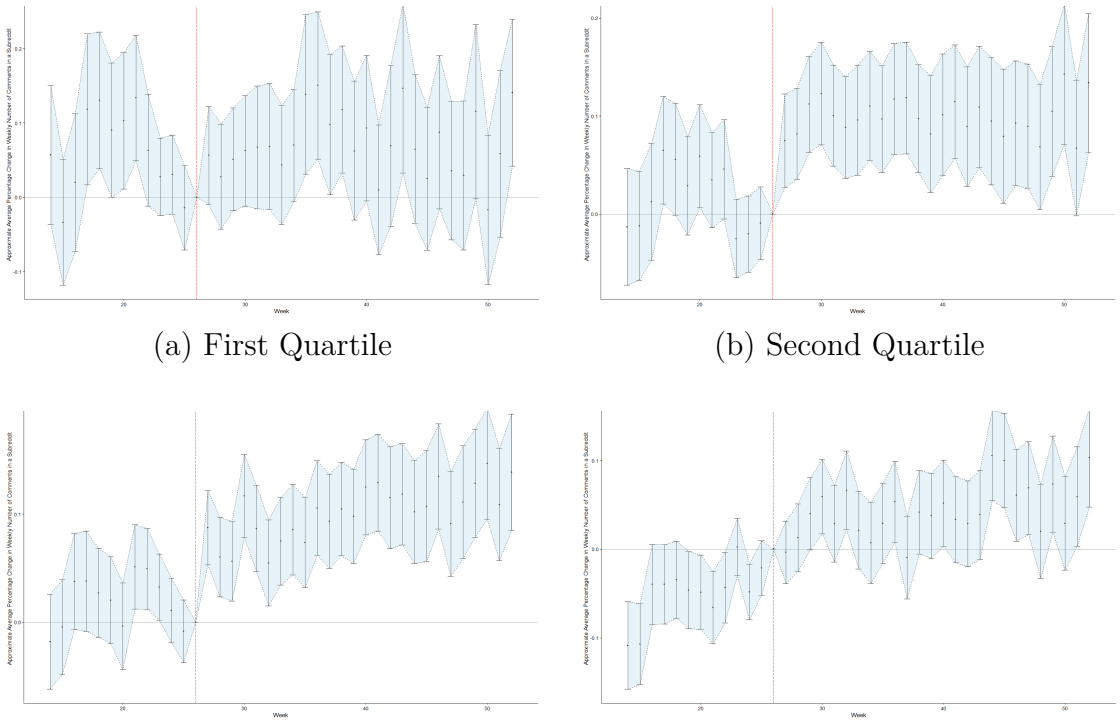
Notes: Figure B.3 shows the effect of the ban on the weekly number of comments left by *core* left- (a) and right-wing (b) users (the results of estimating Equation 2 separately for each group). "Core" users are defined as users in the 4th quartile by share of total comments left in to banned *subreddits* between January and March 2020 ($ShareInBan_i$). Dashed red vertical line indicates the week of the policy change. Confidence intervals are at 95% coverage, clustering is at the user level.

Figure B.4: Probability to reply to a treated user: raw data



Notes: Figure B.4 shows the proportion of comments in each week of the year that are replies to comments left by treated users. This takes the full sample of comments and aggregates them weekly. The continuous decay is due to normal attrition of treated users from the platform.

Figure B.5: Effect on log of number of comments left by *core* users by quartile of previous use of a subreddit: event study with sample restriction, difference-in-differences estimates



(a) First Quartile

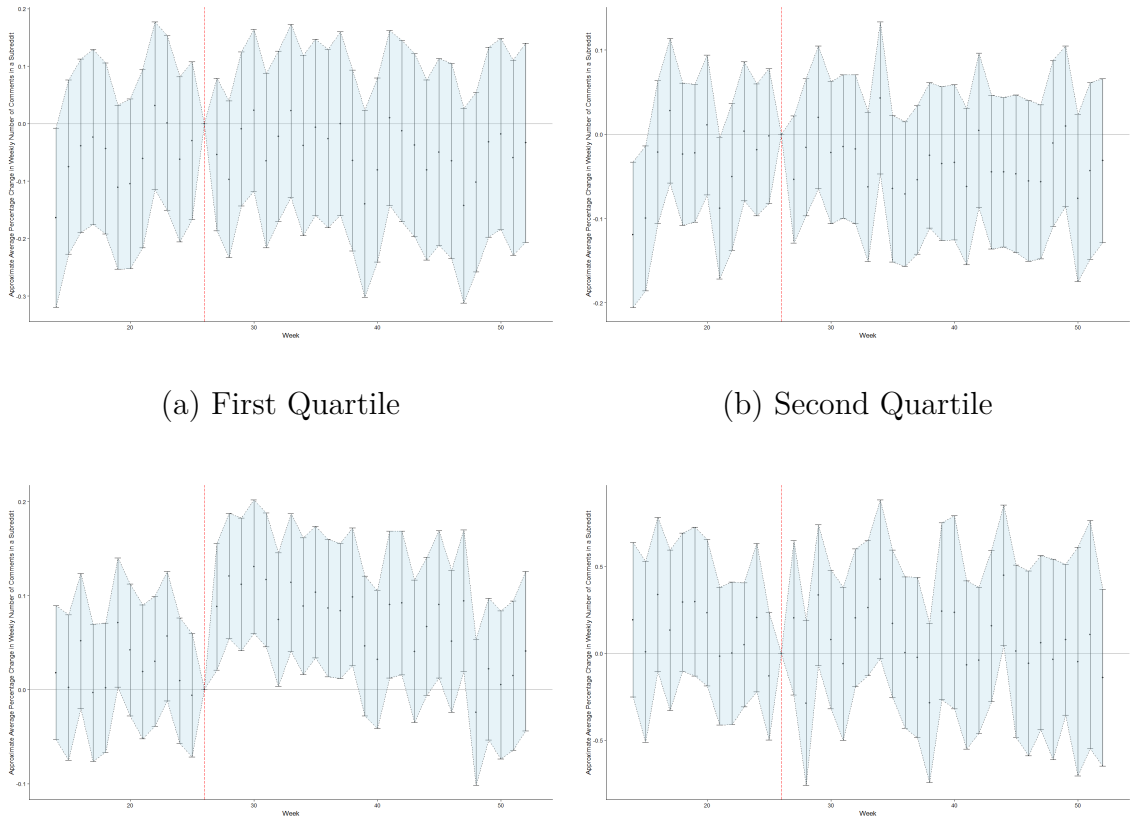
(b) Second Quartile

(c) Third Quartile

(d) Fourth Quartile

Notes: Figure B.5 presents estimates of equation 3 on the subpopulation of *core* users of banned *subreddits*, with average level of user activity in a *subreddit* (in user-level quartiles) as heterogeneity term $\Gamma_i s$. Average level of activity within a forum is calculated as fraction of user's comments left in a specific forum in a week, averaged over the four weeks before the ban (weeks 23-26). The sample is restricted to users who left at least 50 comments between January and March 2020. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.6: Effect on activity of *core* users by pre-ban quartile of hate speech incidence within a *subreddit*: event study with sample restriction, difference-in-differences estimates



(a) First Quartile

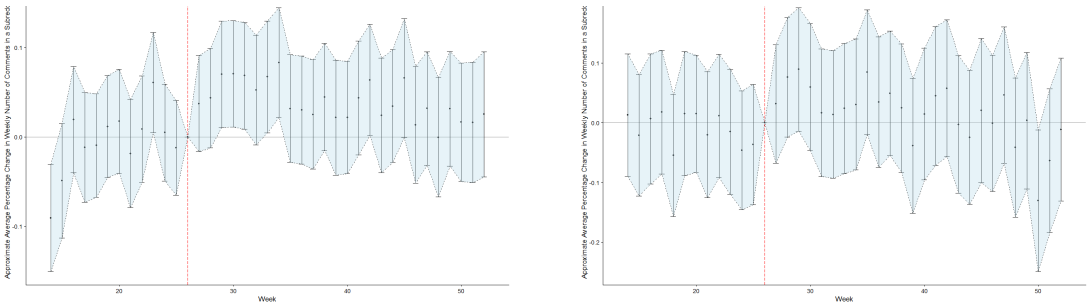
(b) Second Quartile

(c) Third Quartile

(d) Fourth Quartile

Notes: Figure B.6 presents estimates of equation 3 for the subpopulation of *core* treated users, with average hate speech incidence in a *subreddit* (in quartiles) as heterogeneity term Γ_s. Average hate speech incidence within a *subreddit* is calculated as average number of slurs per comment in a week, averaged over four weeks before the ban (weeks 23-26). The sample is restricted to users who left at least 50 comments between January and March 2020. Overall, the results do not support a suggestion that higher levels of hate speech use within a *subreddit* can predict increased activity of treated users in the *subreddit* after the ban. Confidence intervals are at 95% coverage, errors are clustered at the user level.

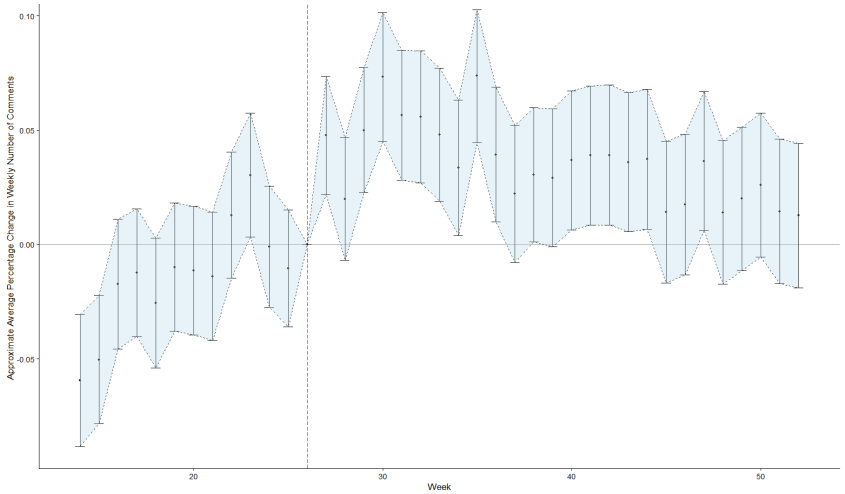
Figure B.7: Effect on activity of *core* users for *viral* and *non-viral subreddits*: event study with sample restriction, difference-in-differences estimates



(a) Non-viral in week 26 (b) Viral in week 26

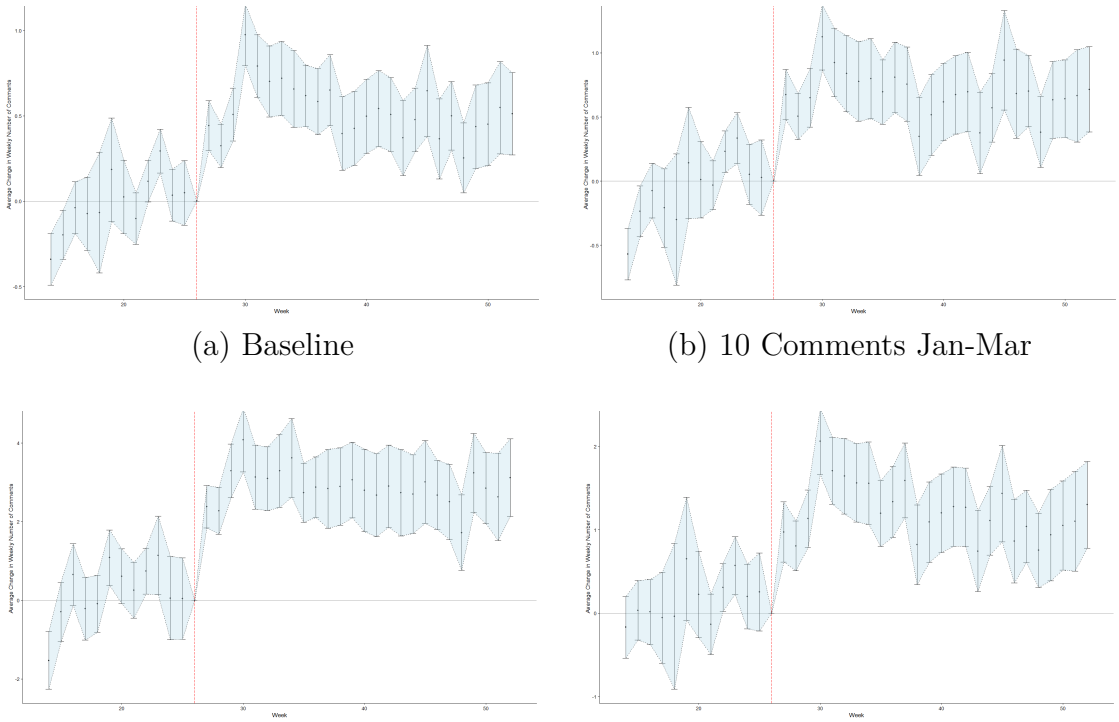
Notes: Figure B.7 estimates Equation 3 for the subpopulation of *core* treated users, with level of *subreddit* virality as heterogeneity term Γ_s. Virality is calculated as the maximum score (sum of upvotes and downvotes) on a comment within a *subreddit* in the week before the ban (week 26). Since a comment can receive a large score only if the post under which it is left receives a lot of attention (that is, is likely on the *front page of Reddit* in that week), the exercise captures whether treated users tend to move to forums which are popular at the time of the ban. Since the statistic behaves according to power law, rather than split the measure into quartiles, we take an absolute cut-off 10,000. This leaves us with around 200 *subreddits* in the "viral" category, which seems to be a correct order of magnitude for the set that could plausibly feature on Reddit's front page within a week. The sample is restricted to users who left at least 50 comments between January and March 2020. The results suggest movement of treated users across the platform is not easily explained by flocking to the popular *subreddits*. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.8: Effect on log of number of comments left by *core* treated users: event study, triple difference estimates



Notes: Figure B.8 shows the event study of the effect of the ban on the number of comments left by *core* treated users outside banned forums (Equation 2). Number of comments is taken in logs (there are no zeroes in the data). The y-axis represents the approximate average percentage change in number of comments left for a user in the 4th quartile of activity in banned subreddits, and the x-axis denotes weeks. The base category is the first week before the ban; the week of the ban is indicated by the dashed vertical line. Confidence intervals are at 95% coverage, clustering is at the user level.

Figure B.9: Effect on number of comments left by *core* treated users: event studies with different sample restrictions, triple difference estimates



(a) Baseline

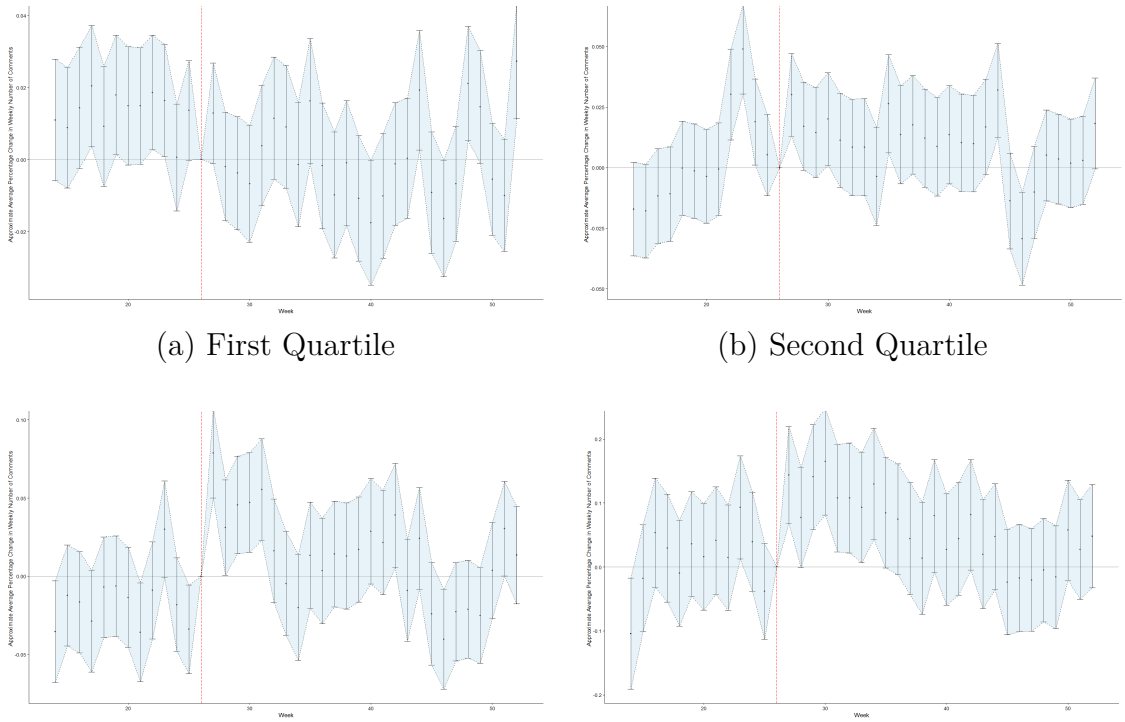
(b) 10 Comments Jan-Mar

(c) 50 Comments Jan-Mar

(d) 100 Comments Jan-Dec

Notes: Figure B.9 shows the effect of the ban on weekly number of comments left by treated users for different subpopulations depending on the volume of activity on the platform. Coefficients are produced by estimating Equation 2. The base category is the first week before the ban; the week of the ban is indicated by the dashed vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.10: Effect on log of number of comments left by treated users by quartile of policy exposure: event study with sample restriction, detrended triple difference estimates



(a) First Quartile

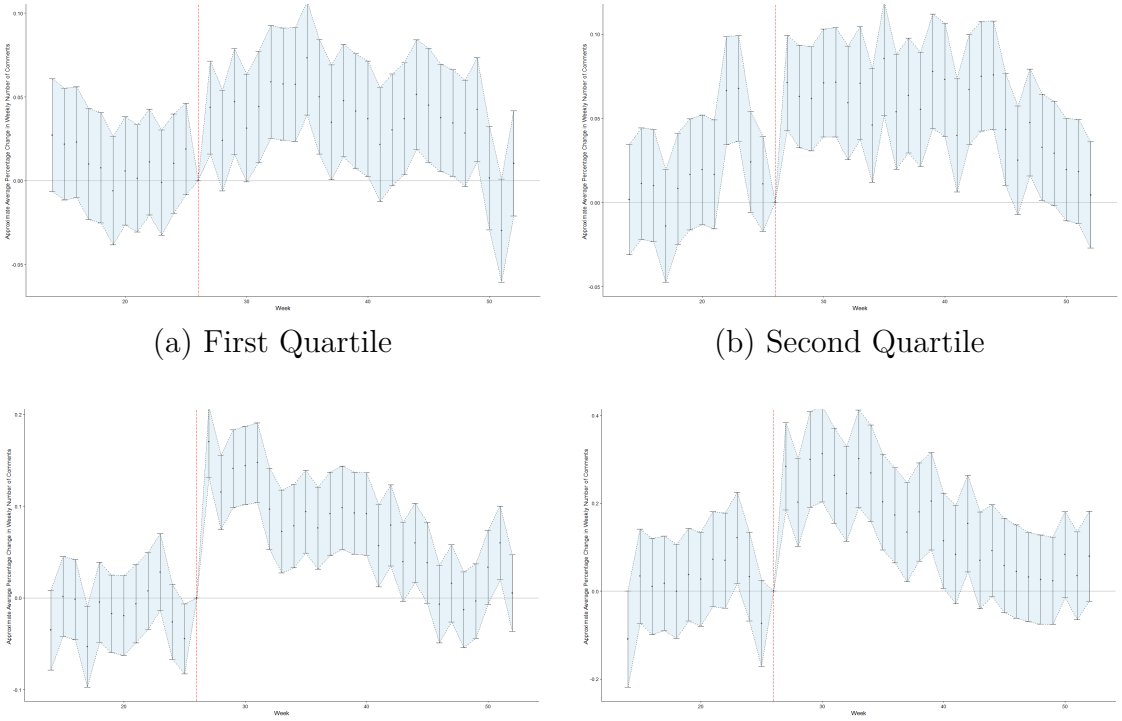
(b) Second Quartile

(c) Third Quartile

(d) Fourth Quartile

Notes: Figure B.10 presents event studies of the effect of the ban on weekly number of comments for different quartiles. Coefficients are estimated using the triple-difference specification described in Equation 3. In addition, quadratic trends are removed to account for the fact that parallel trend assumption does not hold otherwise for the first two quartiles. The sample is restricted to users who left at least 50 comments between January and March 2020. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.11: Effect on log of number of comments left by treated left-wing users by quartile of policy exposure: event study with sample restriction, detrended triple difference estimates



(a) First Quartile

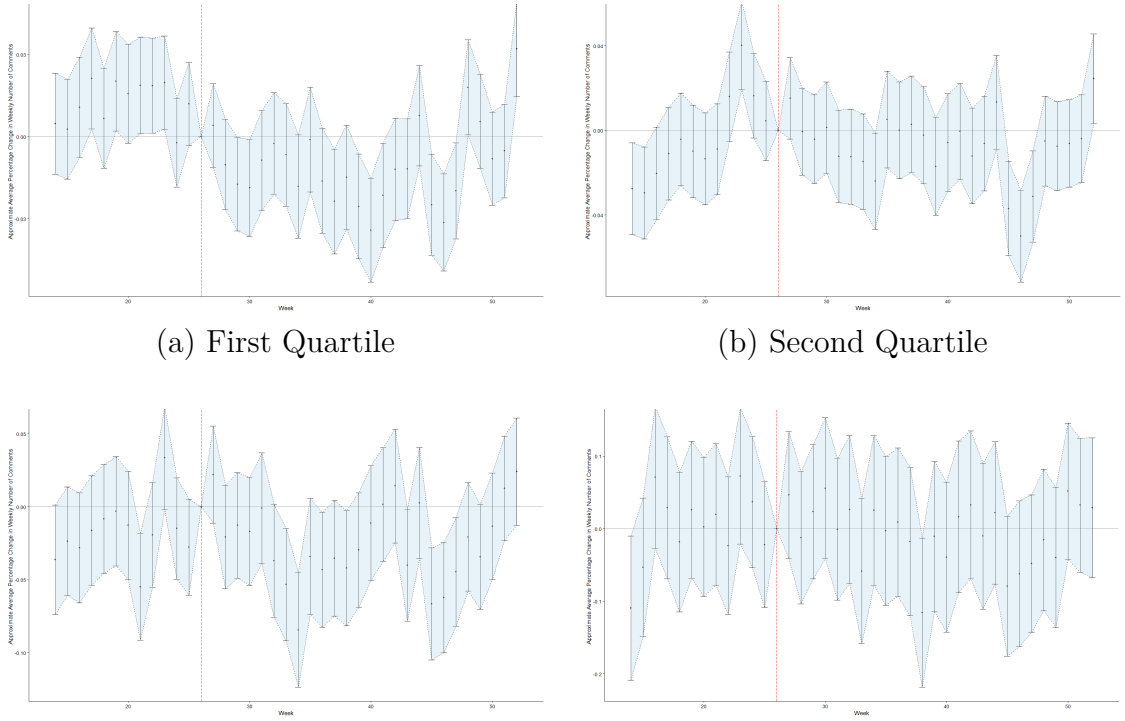
(b) Second Quartile

(c) Third Quartile

(d) Fourth Quartile

Notes: Figure B.11 presents event studies of the effect of the ban on weekly number of comments left by *left-wing users* by quartile of policy exposure ($ShareInBan_i$). The coefficients are produced by estimating Equation 2 separately by subgroup and removing quadratic trends. The sample is restricted to users who left at least 50 comments between January and March 2020. Quadratic trends are removed. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.12: Effect on log of number of comments left by treated right-wing users by quartile of policy exposure: event study with sample restriction, detrended triple difference estimates



(a) First Quartile

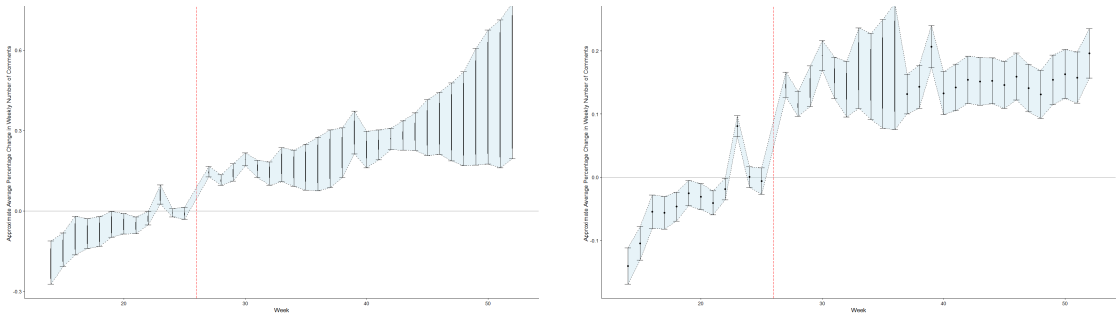
(b) Second Quartile

(c) Third Quartile

(d) Fourth Quartile

Notes: Figure B.12 presents event studies of the effect of the ban on weekly number of comments left by *right-wing users* by quartile of policy exposure ($ShareInBan_i$). The coefficients are produced by estimating Equation 2 separately by subgroup and removing quadratic trends. The sample is restricted to users who left at least 50 comments between January and March 2020. Quadratic trends are removed. Confidence intervals are at 95% coverage, errors are clustered at the user level.

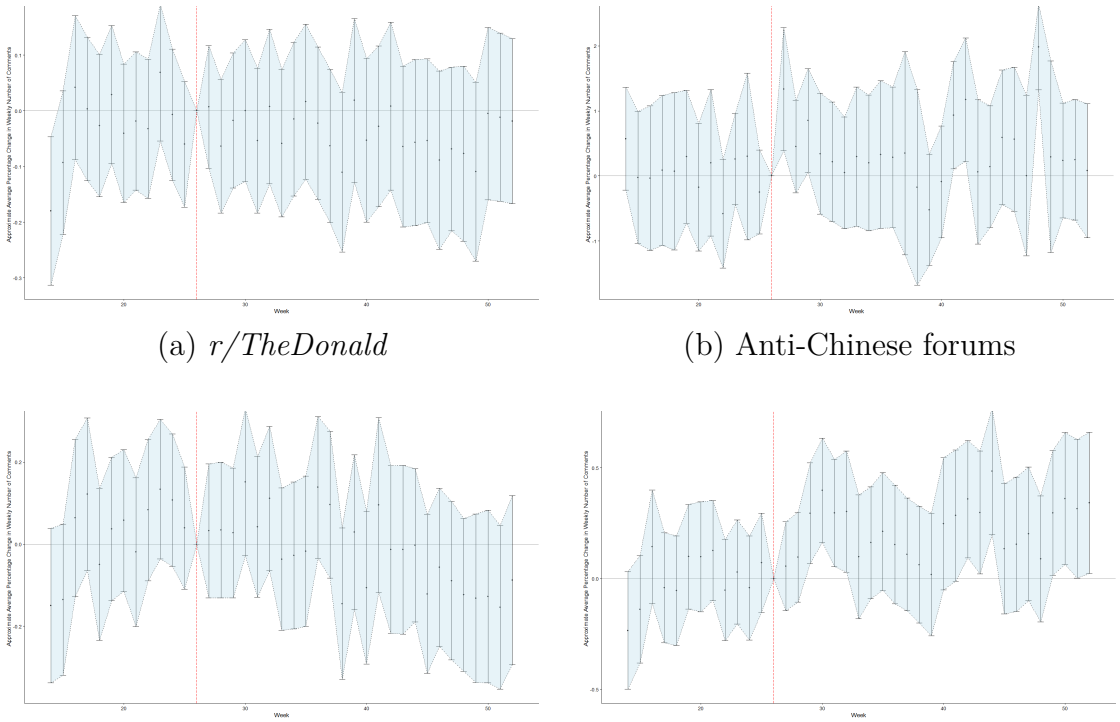
Figure B.13: Effect on log number of comments: event study with sample restriction, difference-in-differences specification, Lee bound estimates



(a) Conservative (b) Baseline

Notes: Figure B.13 presents the results of estimating Equation 1using a Lee bounds style (Lee 2009) trimming procedure. The exercise aims to ensure that the results are not driven by platform exit (although, as we discuss in the paper, the magnitudes of drop-off are small). The drop-off rates for each week are calculated and the top or bottom proportion of either treatment (4th quartile) or control are trimmed to match the samples. Standard errors are calculated by bootstrapping the full procedure. In Panel (a) the event study approach assumes all differential drop-off (even before the ban) could be caused by the policy change and could affect the intensive margin. This is the reason the identified set grows bigger across weeks, as there is higher relative drop-off over time. Panel (b) restricts the potential for set identification to subsample after the policy change and weeks where the treated users have a higher drop-off than the control group - that is, assuming that the policy could only affect drop-off of the treated users after the ban. The sample is restricted to users who left at least 50 comments between January and March 2020. The results confirm that the effect is not driven by platform exit and that estimates from Equation 1 reflect the intensive margin effect of the policy.

Figure B.14: Effect on log of number of comments by right-wing forum type: event study with sample restriction, triple difference estimates

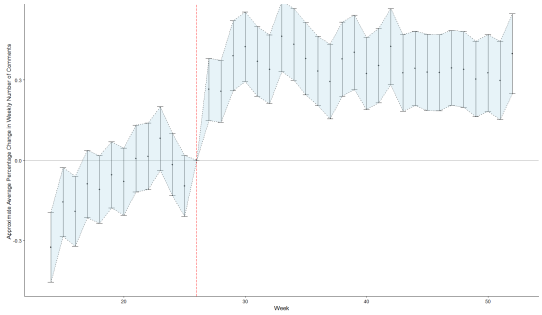


(a) *r/TheDonald* (b) Anti-Chinese forums

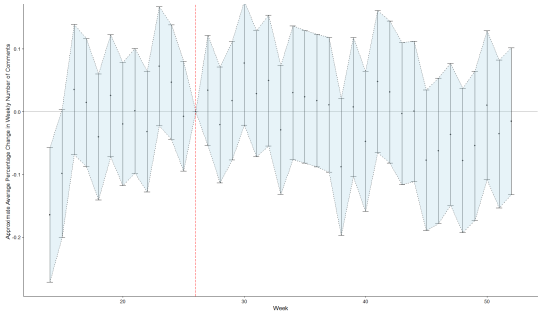
(c) Offensive humour (d) Transphobic content

Notes: Figure B.14 presents the effect of the ban on log weekly number of comments across users of different right-wing forums to account for the fact that "right-wing" subpopulation is significantly more heterogeneous than the left-wing one. The coefficients are produced by estimating Equation 3 separately for each subgroup. Panel (a) shows the effect of the policy on users of *r/TheDonald*, Panel (b) on the users of *subreddits* promoting anti-Chinese sentiment, Panel (c) on *subreddits* banned for offensive humour, and Panel (d) on *subreddits* banned for transphobic content. Given the smaller groupings, we identify users of these forums as users who left at least 80% of all their comments in the forums of the specific subgroup. The sample is restricted to users who posted at least 50 times between January and March 2020. The base category is the first week before the ban; the week of the ban is indicated by the dashed vertical line. Confidence intervals are at 95% coverage, errors are clustered at the individual level.

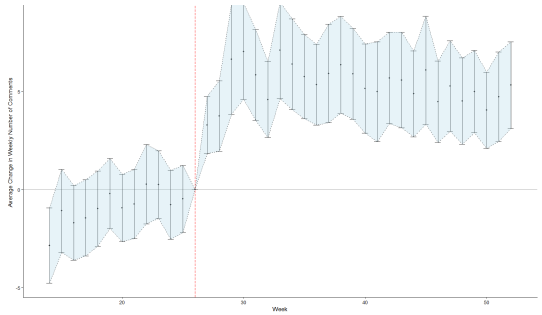
Figure B.15: Effect on log of number of comments by *core* left- and right-wing users: event study with sample restriction, triple difference estimates (alternative quartile definitions)



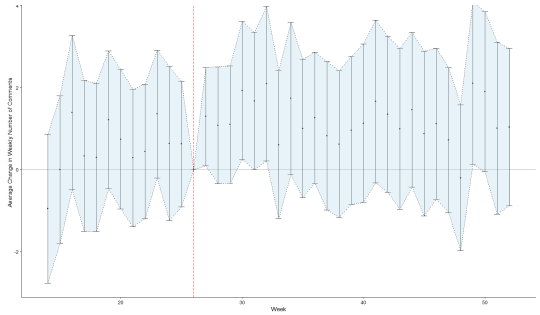
(a) Log # of comments, Left



(b) Log # of comments, Right



(c) # of comments, Left



(d) # of comments, Right

Notes: Figure B.15 presents event studies of the effect of the ban on log and raw weekly number of comments by *core* left- and right-wing users (the results of estimating Equation 2 by corresponding subgroup). The quartile definitions are changed to be group-specific to account for the fact that left wing users are on average more active. This ensures that selection criteria for both left and right subpopulations are more comparable. Panels (a) and (b) show the effects on log number of comments; Panels (c) and (d) on the raw number of comments. The sample is restricted to users who left at least 50 comments between January and March 2020. The base category is the first week before the ban; the week of the ban is indicated by the dashed vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level. The change in definitions does not appear to have a significant impact on the relative effects.

C Use of hate speech

Table C.1: Change in the incidence of hate speech: Poisson QMLE

	# of Slurs per 1000 comments			
	(1)	(2)	(3)	(4)
Post × Treated	-0.083***	-0.061*	-0.086***	-0.078***
	(0.018)	(0.032)	(0.016)	(0.018)
Fixed effect	**User**	**User×Subreddit**	**Post**	**Thread**
Weeks	53	53	53	53
Groups	1,240,315	2,024,956	1,987,752	2,934,579
Observations	782,174,250	266,776,435	314,589,548	87,164,240
Dependent var. mean	1.86	1.86	1.86	1.86

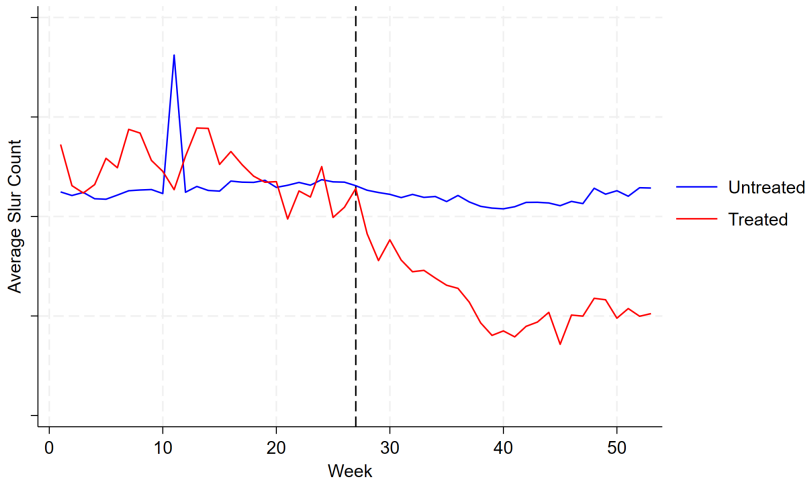
Notes: Table 1 presents the results of estimating Equation 4 with number of unique slurs (incidences of hate speech) per comment as dependent variable using Poisson QMLE. We count only unique words to avoid overestimating comment toxicity, due to probable autocorrelation in use of specific slurs within a comment. Standard errors are clustered at the user level and are in parentheses. The results are different in magnitude to the normalised coefficients in Table 1. This is due to the fact that Poisson QMLE estimator implicitly drops groups which contain only zeroes. This particularity does not effect the coefficients in fixed effects estimation, but massively changes the estimate of the average incidence of hate speech. Therefore, the coefficients are smaller, though still economically and statistically significant.

Table C.2: Change in the incidence of hate speech for treated vs placebo treated groups: difference-in-differences estimates

	# of Slurs per 1000 comments		
	(1)	(2)	(3)
Post × Placebo	-0.042***	-0.025*	-0.025*
	(0.011)	(0.013)	(0.013)
Post × Treated	-0.372***	-0.292**	-0.366***
	(0.074)	(0.139)	(0.086)
Fixed effect	**User**	**User×Subreddit**	**Post**
Weeks	53	53	53
Groups	29,343,181	241,201,445	159,742,315
Observations	1,840,118,635	1,840,118,635	1,840,118,635
Dependent var. mean	1.86	1.86	1.86

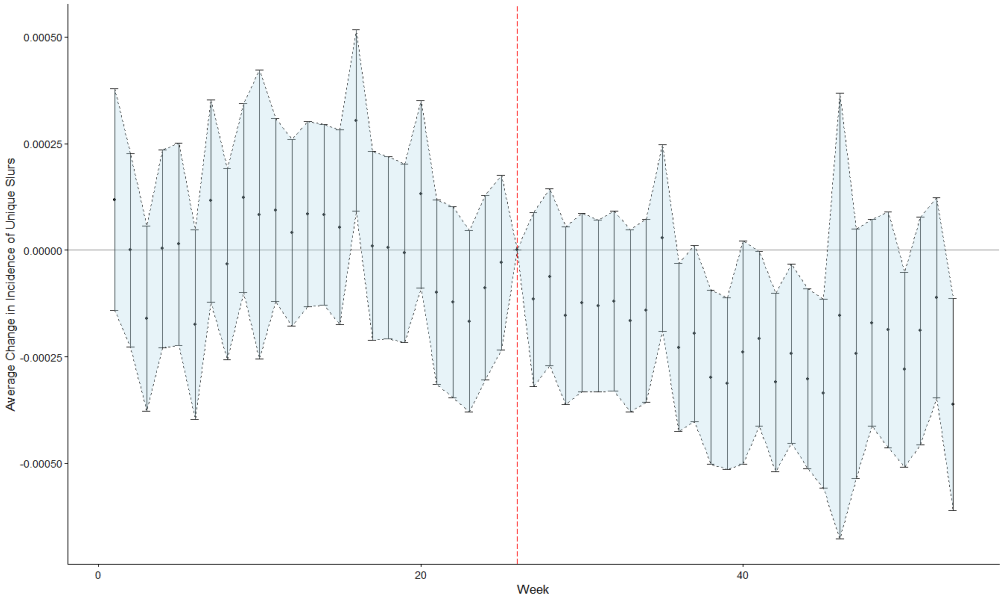
Notes: Table C.2 presents the results of estimating Equation 4 for treated and placebo-treated groups (see Section 4.1 for construction details). For ease of display the dependent variable is multiplied by 1000. Treatment group includes placebo, so the coefficient on the the treated group can be interpreted as the additional effect of being treated over and above the effect of being in the placebo treated group. Standard errors in parentheses are clustered at the user level.

Figure C.1: Incidence of hate speech in comments by treated vs. non-treated users: raw data



Notes: Figure C.1 shows the average number of unique slurs per comment used by members of treated (red) and control (blue) groups over the weeks of the sample. For clarity of presentation the values for both groups are shifted to be zero in the week of the policy change. The week of the policy is denoted in dashed line.

Figure C.2: Effect on incidence of hate speech per comment: event study with sample restriction, difference-in-differences estimates



Notes: Figure C.2 shows the results of estimating Equation 4 as an event study. The outcome variable is the incidence of hate speech per comment. The sample is restricted to users who commented at least 50 times between January and March 2020. Dashed red vertical line represents the week of the policy change. Confidence intervals are at 95% coverage, errors are clustered at the user level.

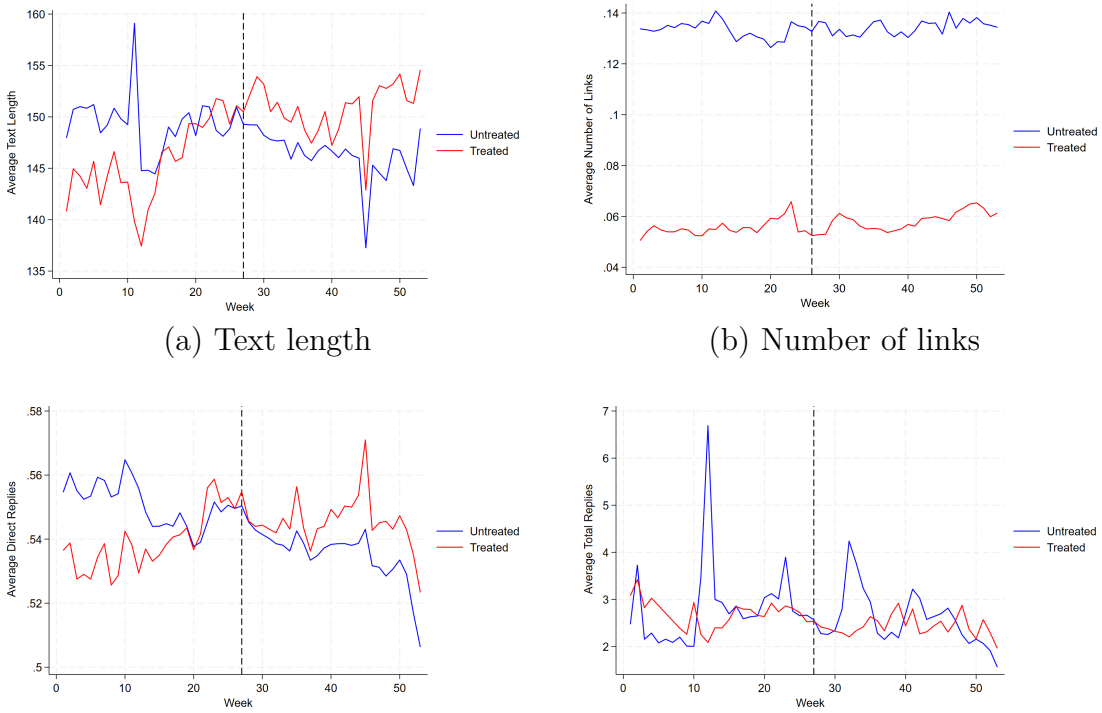
Table C.3: Change in incidence of hate speech by quartile of activity in banned forums: difference-in-differences specification

	# of Slurs per 1000 comments			
	(1)	(2)	(3)	(4)
Post × Q1 $ShareInBan_i$	-0.316^{***}	-0.241	-0.285^{***}	-0.207
	(0.082)	(0.148)	(0.099)	(0.189)
Post × Q2 $ShareInBan_i$	-0.576^{***}	-0.513^{***}	-0.638^{***}	-0.601^{***}
	(0.054)	(0.059)	(0.061)	(0.084)
Post × Q3 $ShareInBan_i$	-0.430^{***}	-0.311^{***}	-0.452^{***}	-0.418^{***}
	(0.079)	(0.084)	(0.104)	(0.124)
Post × Q4 $ShareInBan_i$	-0.324^{**}	-0.263	-0.211	-0.132
	(0.135)	(0.221)	(0.143)	(0.182)
Fixed effect	**User**	**User×Subreddit**	**Post**	**Thread**
Weeks	53	53	53	53
Groups	29,359,218	243,757,537	160,111,081	905,773,238
Observations	1,863,841,419	1,863,841,419	1,863,841,419	1,863,841,419
Dependent var. mean	1.86	1.86	1.86	1.86

Notes: Table C.3 presents the results of estimating Equation 4 by quartile of policy exposure ($ShareInBan_i$) with progressively more granular fixed effects across columns (1)-(4). The outcome variable is number of slurs within a comment. For ease of display the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level.

D Engagement quality

Figure D.1: Measures of engagement: raw data



Notes: Figure D.1 shows the effect of the policy on proxies of engagement quality in the raw data. Panel A displays the average length of a comment, Panel B displays average number of links per comment. Panels C and D display average number of direct and total replies to a comment. The corresponding regression results are presented in Table 3.