Social Norms and Xenophobia: Evidence from Facebook

Jan Sonntag[*]

Sciences Po

<jan.sonntag@sciencespo.fr>

September 13, 2019

Abstract

How do individuals influence each other's behavior online and how does this shape the type of views and information that gets shared on social networks? To answer these questions, I study the spread of hate speech on social media and attempts to contain it – a topic that has received a great deal of attention recently as lawmakers increasingly call for action against xenophobic, antisemitic and otherwise marginalizing messages. I assess the effectiveness of counterspeech, a decentralized mechanism against online hate speech that consists of moderate users speaking out against hate. I collect data on Facebook posts by large German news media as well as on user responses. Comparing the future behavior of individuals who were targeted by a counterspeech group to a set of control individuals, I find a sizable drop in their likelihood to engage in hate speech for about two weeks. These individuals are also less likely to engage in discussions in general and in contentious ones in particular. In addition, I show that news articles targeted by such interventions see more moderate individuals join the debate than control articles. I interpret my findings as evidence consistent with counterspeech acting as non-monetary punishment facilitating coordination on a social norm.

Keywords: Hate speech, Social norms, Non-monetary punishment, Text as data, Machine learning
JEL: D64, D71, D91

---

[*]Corresponding author: Département d'Economie, Sciences Po, 28 Rue des Saints-Pères, 75007 Paris.

1   Introduction

The ability of social media to shape people's views and actions has been the subject of intense debate in recent years, in particular in the aftermath of the US presidential elections and a sweeping rise of extreme right populism in many European countries. Its role as a platform for hate speech has received particular attention in this context. Recent reports of Facebook acting as a catalyst for violence against the Muslim minority in Myanmar put pressure on lawmakers and social media companies alike to rein in online hate speech (Reuters, 2018). Germany, for instance, adopted legislation that holds social media companies responsible for incendiary content posted by users on their websites. This paper evaluates the effectiveness of an alternative, decentralized approach to countering hate on social media: large scale counterspeech by users themselves.

A key motivation for curbing hate speech is that words will ultimately lead to actions. This concern is increasingly backed by empirical evidence. Yanagizawa-Drott (2014) and Adena et al. (2015) highlight the importance of mass media in fueling the genocides in Rwanda and Germany. Müller and Schwarz (2018a,b) argue that the prevalence of online hate speech increases the rate of hate crimes. Outside of the economics literature, there is evidence that hate speech has negative consequences for victims in terms of fear and participation in public debate (see Siegel (2018) for an overview).

If hate speech is so dangerous, how can it be kept in check? In principle, there are two broad approaches: centralized government intervention or a decentralized, more market based approach. The former has been adopted to some degree by many countries around the globe that have banned the most extreme forms of hate speech in order to protect minorities.[1] In modern democracies, however, the fundamental right of free speech limits the possible extent of such bans.

The second approach, favored for instance by social media companies, relies on counterspeech, a social control mechanism that requires users to step in, contradict, and speak out against hate speech in order to support victims and deter further transgressions. Organized counterspeech groups have formed and attracted sometimes large numbers of members that coordinate their efforts to counter hate speech. With the obvious risk of government intervention exceeding its aim and drifting into censorship, this decentralized approach seems of course very appealing.

The key question that this paper aims to answer is to what extent this bottom-up approach is effective and – more generally – to what extent individuals can influence each other's behavior in online debates. Does counterspeech cause targeted individuals to stop engaging in hate speech in the future? If so, what channels could explain this response? What happens to users who did not participate in the discussion before? Does counterspeech discourage new discriminatory messages?

---

[1]For example, German criminal law outlaws some of the most extreme forms of hate speech such as Holocaust denial, incitement to violence or civil unrest targeting protected groups. Similar laws exist in other European countries. In the US, on the other hand, the protection of First Amendment rights has generally outweighed concerns about the protection of minorities.

In order to study these questions, I use the Facebook pages of German language news media as a laboratory.[2] In response to widespread hate speech on these pages, a large bottom-up counterspeech group was founded in late 2016 that attracted more than 35,000 members within a few months. The group intervenes each day on 1-2 media articles that receive particularly large numbers of hateful comments. An intervention consists of members coordinating to post comments condemning hate speech on the selected articles and to respond to hateful comments directly with the stated goal of countering and ultimately reducing the prevalence of hate speech. Due to the group's size and its focus on only a few articles, their interventions can take up a significant share of comments and were thus highly visible to anyone seeing the targeted articles on Facebook.[3]

To assess the impact of these interventions, I collected six months' worth of data on large German news medias' Facebook pages. The data include all posts by the media outlets, as well as all user comments, "likes", and replies to comments that were made in public and visible to anyone. I manually annotated several thousand of these comments and leveraged recent advances in deep learning for natural language processing to infer which of the millions of comments in my data contain hate speech. In addition, the counterspeech group shared a sanitized version of their leadership's chat logs with me which contain all articles that the group considered targeting with an intervention. This allows me to identify the treatment effects of interventions by comparing a treatment and a control group of individuals. The treatment group consists of individuals who were active on Facebook articles which were subject to an intervention. The control group contains individuals who were active on "runner-up" articles which were considered by the counterspeech group as targets for an intervention but ultimately were not chosen. I obtain these posts by restricting the sample of articles from the chat log to instances when the group faced a capacity constraint and had to choose between at least two ex ante similar posts in terms of their total number of comments, likes and hate comments. This ensures that treatment status of individuals is plausibly exogenous to their behavior.

The main result of this paper is that the counterspeech interventions have a substantial but transitory moderating impact on individuals' future behavior. For about two weeks after an intervention, users in the treatment group are 5.3 percentage points less likely to write or condone a xenophobic comment in a given week than the control group. After this period, individuals appear to revert to their initial behavior. Compared to the treatment group's baseline probability of engaging in hate speech during a given week of about 25%, this corresponds to a sizable 21% reduction. Individuals who only occasionally spread hate speech before the intervention alter their behavior the most, while I find little effect on users who did so more than once a week. The effect seems to be driven at least partly by targeted individuals staying away from contentious debates prone to xenophobia. Following an intervention, they reduce the number of comments they write or like, in particular on articles pertaining to politics,

---

immigration, and related topics. Their commentary activity on sports and the weather, on the other hand, increases slightly.

Additionally, I document that articles targeted by a counterspeech intervention see an influx of more moderate users participating in their discussion. Articles experience an increase of up to 50% in the number of comments and likes of comments, only parts of which is driven by participants in the counterspeech group: interventions trigger a ripple-on effect that attracts up to 20% more new users to the article who did not previously participate in any interventions. These additional users tend to be more moderate than individuals engaging on control posts and are less likely to make hateful comments. As a result, while the total number of hateful comments remains comparable to control posts, their share in the total activity created by individuals not participating in the intervention decreases by about 3 percentage points.

I discuss three possible channels which could explain my main results. First, interventions could simply provide additional information on the topic of the article. Individuals exposed to this kind of information treatment would then be able to correct erroneous beliefs, for instance about crime rates among refugees, and adapt their behavior accordingly. Second, individuals could infer social norms from the average behavior of other users discussing news articles. The intervention would then induce individuals with a taste for conformity with other users' behavior to not express xenophobic views. Finally, counterspeech could be a form of non-monetary punishment. The members of the intervening group write messages in which they publicly disapprove of the behavior of individuals who engage in hate speech. This could lead the latter to conform with the norm conveyed through these sanctions.

While the design of the natural experiment I am using for identification does not allow me to definitively answer which of these channels are at play, the findings seem to be most consistent with the interventions acting as non-monetary punishment. First, manually classifying a sample of counterspeech messages reveals that only a small fraction of them contain new factual information, suggesting that the interventions are not pure information treatments. In addition, the effects are temporary and smaller for individuals who get treated multiple times. This set of facts seems difficult to reconcile with information provision. Second, I find no correlation between the share of counterspeech in an article's discussion and the effectiveness of the intervention. This makes it unlikely that users infer a social norm from the distribution of others' comments on a given article and adapt their behavior to be more conform with others'. Third, I find that the effect of an intervention is strongest for those users who received a counterspeech message as a direct reply to their own comment, as opposed to a general comment to the article denouncing hateful comments. As most counterspeech messages are in fact expressions of disapproval of xenophobic views, this response is consistent with the messages acting as non-monetary punishment for norm transgression. The punishments' effectiveness could in principle be explained both by the messages inducing a behavioral response to shame, and by the messages communicating the presence of a social norm. While the results presented here are consistent with both mechanisms, the social norms channel seems to be more relevant in light of previous findings in the literature on punishment.

Of course, I only observe the behavior of individuals when they comment publicly on news articles posted by German news media and I therefore cannot rule out that they simply express hateful views elsewhere, be it in private or outside Facebook. However, the data do contain articles from news media that attract very high shares of xenophobic comments and no counterspeech interventions and I find no evidence that individuals shift their activity towards these outlets. This suggests that it is unlikely that there are large displacement effects. Moreover, to the extent that one is concerned about hate speech being broadcasted to large audiences, inducing perpetrators to leave major news platforms may already be an important step.

Beyond the immediate setting I study, I argue that my results allow drawing more general conclusions. First, the fact that speaking up against hate speech has a sizable effect in a relatively impersonal online context suggests that the mechanism is likely not specific to this environment but holds true also elsewhere. This may indicate that contradicting hateful views is an effective intervention more broadly, both online and offline. Second, my findings shed light on how individuals respond to feedback by others more generally. They highlight that behavior is not hardwired by preferences, but that individuals' decisions can be affected significantly even by relatively isolated expressions of disapproval.

This paper contributes to the nascent literature on online hate speech, its drivers, and consequences. In a small scale experiment on Twitter, Munger (2017) uses false accounts to call out users for employing racist slurs and finds that they reduce their supply of hate speech temporarily. My results, on the other hand, rely on a large-scale natural experiment using actual interventions not administered by researchers which also allows me to track responses of users not directly targeted by an intervention. Müller and Schwarz (2018a) argue that anti-refugee comments on the Facebook page of Germany's populist right wing party cause an increase in the number of hate crimes in areas with high social media affinity. In the same vein, Müller and Schwarz (2018b) show that President Trump's racially charged Twitter comments may have led to a spike in hate crimes against Muslims. I build on their results and investigate the drivers of hate speech and a potential remedy.

More broadly, this research studies how social norms may affect behavior in an online context, thereby extending the existing literature on social norms in economics. Bénabou and Tirole (2006, 2012) develop a theoretical model in which individuals choose their actions in part as a response to what those actions tell others about them. The importance of individuals' concern for social norms in the context of political action has been documented by Bursztyn et al. (2017), who show experimentally that the election of Donald Trump made voters positively update their beliefs about the prevalence of xenophobic views in the population, causing xenophobes to reveal themselves more freely. Enikolopov et al. (2017) show that Russian protesters participated in demonstrations due to image concerns and provide a dynamic model of protest participation. In an experiment conducted in Pakistan, Bursztyn et al. (2019) show that men's decision to take a costly anti-American action depends in part on whether or not is observable by a moderate majority of participants.

My work is also related to the literature on non-monetary punishment in the presence of

social norms. Masclet et al. (2003) and Noussair and Tucker (2005) conduct a series of repeated public goods experiments in which they find that sanctions are effective at establishing and sustaining a norm of cooperation even when these sanctions do not directly affect participants' payoffs.[4] Xiao and Houser, 2009 and Ellingsen and Johannesson (2008) show that in dictator games even the prospect of receiving written comments by recipients induces dictators to increase transfers. By varying the degree of publicity of sanctions, Xiao and Houser (2011) establish that punishments serve to express and raise the salience of social norms, rather than relying purely on a shaming effect. The results presented in this paper suggest that the findings on non-monetary punishment matter outside the laboratory as well.

On the methodological side, I extend the growing literature applying machine learning techniques to process text as data in economics (see Loughran and Mcdonald (2016); Gentzkow et al. (2017) for an introduction and overview). In the context of political preferences, Groseclose and Milyo (2005), Gentzkow and Shapiro (2010), Jensen et al. (2013) and Gentzkow et al. (2016) use the US Congressional Record and apply simple, dictionary-based approaches to compute measures of polarization and slant based on the lexical differences between Republican and Democratic speeches. Relying more strongly on machine learning, Hansen et al. (2018) assess the impact of transparency on monetary policy deliberation. I build on more recent techniques in deep learning and use recurrent neural networks to identify hate speech in my corpus of Facebook comments.

The remainder of this paper is structured as follows: Section 2 introduces the data and provides descriptive statistics. Section 3 details the identification strategy. Section 4 reports the empirical evidence on the impact of counterspeech on the future behavior of treated users. The impact on aggregate user behavior on the targeted articles is contained in Section 5. Possible channels explaining these results are discussed in Section 6 before I conclude.

2 Data

2.1 Data sources

I combine data from three different sources. First, I collected roughly 12 million public user comments and 24 million "likes" on those comments pertaining to the 249,426 posts by 85 popular German language news media on Facebook between August 1st 2017 and February 1st 2018.[5] These data were publicly available through Facebook's API and include the comments' complete text, the time of writing, as well as a unique user ID that allows identifying users across multiple Facebook pages.[6] In addition, I collected information on the exact time, the

---

[4]Gächter and Fehr (1999) similarly report an increase in cooperation when adding social feedback mechanisms to the public goods game, but only when participants were familiarized with one another before the experiment.

[5]This includes Facebook pages of news organizations with more than 60,000 followers, as well as a few manual additions of regional branches of the public broadcasting services and of the *Bild*, Germany's largest tabloid. Table XIV in the appendix contains the full list of media included.

[6]No other user information was collected. Only comments that users made on public Facebook pages are included in the data.

title and the "teaser" or description of the posts by news media, which are typically links to articles, pictures or videos.

Second, I identify all interventions of the counterspeech group *#ichbinhier* by manually collecting all calls to intervene on its internal Facebook group that is used to coordinate its roughly 35,000 members. In addition to the post ID required to identify on which post the group intervened, I also record the exact time of the call to intervene.

Finally, the organizers of the counterspeech group generously granted me access to a sanitized version of their internal chat that contains all mentions of urls along with a time stamp. As explained in more detail below, I use these urls to identify posts by news media on Facebook that the group considered as possible targets for their interventions.

2.2 Defining and identifying hateful comments

From the posts' and comments' raw text I extract the categorical variables for my econometric analysis. Most challenging in this respect is to find out which comments in the dataset contain hate speech and which ones do not.

Previous research has relied on dictionary-based techniques to identify hate speech. Munger (2017) searches the text for predefined racist slur-words, Müller and Schwarz (2018a) simply look for the word "refugee" on a right wing Facebook page to identify hateful messages. This approach proves to be insufficient in the present context for three main reasons. First, few users employ unambiguous racist slur words. Rather, they use mostly harmless language to convey racist ideas (e.g. "I think all Muslims should be forced to leave the country"). Second, the user base is not solely comprised of racists, so that occurrence of group names ("refugees", "Muslims", etc.) is not sufficiently predictive of hate speech but may in fact be used in a positive or neutral manner. Finally, users often misspell words or employ neologisms.

Instead, the approach chosen here relies on recent advances in machine learning. I first adopt a standard definition of hate speech to manually classify a set of comments and then train a deep learning algorithm on the manually categorized data. The algorithm then applies the learned categorization on the remaining data that were not manually classified.

The definition of hate speech applied here borrows heavily from Gagliardone et al. (2015) and as such is larger than the narrow definition of hate speech used by the German penal code, which views hate speech mainly as incitement to violence. A comment is counted as hate speech if it (i) insults, diminishes and/or approves discrimination of and/or violence against any group or group member based on religion, origin or ethnicity, (ii) generalizes (perceived) negative behavior or characteristics of individual group members to the group as a whole or (iii) questions intentions, honesty or ability of an individual based on group membership or spread clearly false information about the group with the intention to diminish, insult or spread prejudice.[7]

Using this definition, I manually classified a set of 15,000 comments with the help of a re-

---

[7] While I would have preferred conducting a more comprehensive analysis including homophobic and sexist hate speech, the data contain too few instances for robust statistical learning.

search assistant. About half of theses comments were sampled randomly from the entire dataset of comments, the remainder was selected by sampling from articles about refugees, immigration and crime – topics which attract a disproportionate amount of hate speech. This oversampling was necessary to insure a sufficient amount of xenophobic comments in the manually classified data as hate speech is, fortunately, a relatively rare phenomenon compared to the overall number of comments produced by users.

The hand-classified dataset thus obtained is used to train and evaluate various machine classifiers. The same reasons that rendered a keyword based approach infeasible in this context make classical bag-of-words machine learning approaches sometimes used in economics perform poorly as well (see Gentzkow et al., 2017 for an overview of these techniques). Instead, the most accurate classifier proves to be a Long Short-Term Memory algorithm similar to those often used in machine translation and text generation. These algorithms cope much better with the context dependency of the meaning of words, with misspellings and synonymy. A detailed description of the classifier is deferred to Appendix A.

The classifier achieves a level of accuracy of 94.4 percent, which means that this percentage of comments is categorized correctly. Hate speech is relatively rare (about 3.4 percent in the entire dataset) making accuracy a somewhat uninformative metric. A classifier could simply never predict hate speech and still achieve 96.6% accuracy. More informative performance metrics are therefore the area under the receiver operating characteristic curve (93.4), as well as the classifier's sensitivity (56.9%) and specificity (97.3%).[8] While the classifier is certainly not perfect and slightly to conservative in predicting if any individual comment is hateful or not, it provides sufficient accuracy for the somewhat more aggregated analyses presented here.[9]

In addition to commenting on media posts, Facebook users can choose to react to other users' comments directly with an icon meant to convey different emotions, with Facebook's signature thumbs-up symbol ("like"), or with a free text comment (sub-level comment). I treat reactions to a hateful comments with a "like" or a heart-symbol as hate speech as the user clearly expresses agreement with the comment's content. After careful manual review of more than a thousand sub-level comments, I chose to exclude them from the textual analysis as they rarely contain unequivocal approval of the initial comment or new instances of hate speech.

Media articles are assigned to topic categories applying another machine classifier trained on a small manually annotated sample. While not strictly necessary for the analysis at the core of this paper, this intermediate step allows me to report more meaningful descriptive statistics.

---

[8]The area under the receiver operating characteristic curve is the integral over the curve plotting the true-positive rate against the false-positive rate for each probability threshold of predicting positive outcome. Specificity is defined as the share of correctly predicted negatives in all negatives in the sample. Sensitivity is the share of correctly predicted positives in all positives in the sample.

[9]All key results presented in this paper are obtained using hate speech as the dependent variable so that classical measurement error should not introduce any biases.

2.3 Descriptive statistics

Users' responses to news articles

The media posts in the resulting dataset spawned reactions that differ substantially in their extent and nature depending on the posts' topic and the media outlet that produced it. The left panel of Table I shows the breakdown of posts (articles), the average number of user comments and likes, as well as the average share of these comments and likes containing hate speech by broad category of article topics. Most user reactions are attracted by articles that mention refugees. Despite the peak of the German "refugee crisis" of 2015 being long in the past during the observation period, there were almost daily articles about refugees on the social media pages of German media. These are also the articles that have the highest share of xenophobic comments. The second highest share of hateful comments is on miscellaneous articles that mostly cover crimes and lead many users to draw hasty conclusions about the origins of perpetrators.

These correlations are not merely driven by larger outlets producing more populist content. In Appendix C.1 I report the results of regressions predicting activity levels and the share of hateful comments by topic-dummies and outlet fixed effects. It shows that the positive correlation between migration-related articles and the number of both hateful and other comments and likes is highly significant even when exploiting only variation within media outlets. Moreover, the share of xenophobic comments remains a strong, positive predictor of the number of comments and likes on an article, even when including a rich set of fixed-effects.[10]

The right panel of Table I shows how user activity is distributed by topic category. Users who comment on politics and on immigration related articles do so most actively, with an average of 10.9 and 7.9 comments and likes on these topics. They are also the most likely to have written or liked at least one xenophobic comment on an article on these topics. Overall, the average individual in the sample has written or liked a total of 10.1 comments and 9.6% of these users have written or condoned a xenophobic comment at least once.[11]

Both the total number of responses to articles and the share of xenophobic responses are highly volatile over time. Figure 1 suggests that some of this volatility can be attributed to specific events. For instance, the terror attacks in Barcelona on August 17, 2017 were followed by a large spike in the share of xenophobic comments, while the day of the German federal elections saw a spike in the overall number of comments and likes. Other peaks and troughs seem to be driven by multiple smaller coinciding events.

Individuals' activity

As is to be expected, the distribution of user activity is highly skewed. The vast majority of users only makes rare appearances on the public comment sections of major news media, on

---

[10]This suggests that outlets may have an incentive to produce news stories that will trigger xenophobic comments if they want to get more activity to their pages. I elaborate on this issue in the conclusion.

[11]Table XV in the appendix provides the same breakdown by news media instead of topic.

average 1.3 times per week (Figure 2). This is true despite the fact that I only observe users in my data that commented publicly at least once.

A concern often raised in the popular press is that social media debates are often waged by bots. Inspecting the high activity users, I find little evidence that would suggest that these profiles are robots: despite being very active, there speed is not superhuman, their messages are not simple copy-pastes or highly similar messages and often contain responses to other users that would be very difficult to automate. Their German contains typos, but is not excessively error ridden. Overall, I find few longer duplicate comments in my database that would suggest the presence of bots even among the less active users. This is in line with the fact that Facebook has more stringent identity verification processes in place than other social media companies making it relatively more costly (but not impossible) to create fake profiles that can be automated.[12]

Counterspeech interventions

Figure 3 shows that the counterspeech group was quite successful at staging large interventions during the observation period. Although only relatively few posts were targeted, the majority of the 315 interventions managed to achieve a share of more than 20% of the overall comment section of that post. The average intervention involved 84 members of the counterspeech group and large ones more than a thousand. This lends credibility to the argument that these interventions did not go unnoticed by users who saw or commented on the news article on Facebook.

More generally, individuals who comment on news articles seem to also read and respond to the comments of others. One fifth of the users who write a comment on an intervention post also like or comment on another user's comment. Among those whose comments received a reply by another user, at least 41% write another reply to that comment. Anecdotally, one can see users responding to the interventions in sub-level comments and engaging in (not always friendly) discussions.

In order to provide a better sense of the content of the counterspeech messages, I manually classified approximately 600 comments into different content categories. The results of this exercise are reported in Figure 4 and in more detail in Appendix E. I considered both top-level comments that were written as a comment directly on the article and sub-level comments which are responses to other users' comments.

The largest group of comments contains common sense arguments against xenophobic views. About a quarter of the top-level comment express plain disagreement with hateful views without providing additional information or arguments. This share is much lower for responses to other users' comments. Only 14% of top-level comments contain new pieces of information, such as statistics or a link to further reading on a topic. Social norms are invoked in about 11-13% of the comments and mostly in the form of injunctive norms. A minority of comments contain ad

---

[12]I can only identify 154 profiles that are likely bots by looking for longer duplicate messages, users that post mainly links or that "never sleep". Excluding them does not change the analyses presented here.

hominem attacks on other users such as expressions of doubt about their intellectual capacity or social standing. Outright insults or foul language are the exception. Equally rare are threats to report another user to Facebook or authorities. Among the remaining comments, many call for a more fact-based debate more generally or complain about the way the article is slanted by journalists to trigger hateful reactions by users.[13]

The types of articles and media outlets that were targeted by the counterspeech group are discussed in the next section after I have introduced the identification strategy.

3   Identification Strategy

The aim of this paper is to identify the causal effects of the counterspeech intervention both on the future behavior of users that were affected by them as well as the overall incidence of hateful comments on the targeted articles. Doing so requires constructing a credible counterfactual of what would have happened in absence of the intervention. To this end, I exploit the specific way in which these interventions were carried out.

The counterspeech group counted around 35,000 members at the time the data collection started but its organization was and still is highly centralized and revolves around a small group of volunteers.[14] Every day, a handful of moderators monitors major German language news media on Facebook and picks an average of 2.1 posts a day on which to intervene. They specifically look for articles that were posted within the last two hours, attracted a lot of hateful comments and accumulated high numbers of comments and likes.[15] Once the moderators decide to intervene on a given post, they write a message on the board of the Facebook group so that it becomes visible to its thousands of members. Those who wish to follow the call to action can then write comments on the targeted article in a coordinated manner.[16] For concreteness, Appendix D contains the timeline of one specific intervention as an illustrative example.

While the group's intervention is not random per se, I argue that it has two key features that allow for the identification of causal effects. First, the group faces a capacity constraint: it only intervenes on one post at a time and only on a few posts per day in order not to divide its resources among too many interventions. Especially at times when news break which spawn many hateful reactions, this means that the group is forced to pick its battles and leave posts uncontested that it would have liked to intervene on. Second, the candidate articles that were considered as potential targets of an intervention are known and retrievable from the moderators' internal chat log. This enables me to compare actual interventions to a control group of "runner-up" posts. These are articles that the group considered as targets for an

---

[13]Among the responses to other users' comments, there is also a large share which is difficult to interpret as they refer to previous sub-level comments in often lengthy discussions.

[14]The group was founded in late December 2016. By the time I started the data collection it already had 35,000 members, a number that grew to 37,000 by the end of the observation period.

[15]As the team of volunteers did not have automated tools at the time, there is no sharp discontinuities that I could exploit for identification.

[16]The call to action comes with a link to the article and with a link to a tool that helps identify other group members to facilitate targeted "liking" of other group members' comments. The group has a common hashtag that makes it easy to recognize these comments.

intervention but ultimately did not intervene on. I argue that between comparable candidate articles among which the moderators had to choose, the assignment of the intervention is as-if random.

Specifically, I construct the control group as follows. From the internal chat protocol of the group's moderators in charge of selecting the posts for intervention I retrieve all links that were discussed in the group. This set of links contains all the posts that were targeted by an intervention as well as all candidate posts that were considered for intervention. Of course, not at every point in time were there multiple posts that were equally likely to become the target of an intervention. Sometimes, there was just one article that clearly attracted most hateful comments and no comparable runner-up. The set of candidate posts therefore needs to be restricted in order to construct a comparable control group. I run a logistic regression to predict intervention probabilities for each post from the chat by the log number of comments and likes and the share of hateful comments over a 100 and a 30 minutes interval before intervention, as well as the stock of those numbers 100 minutes from intervention.[17] For each intervention post, I then retain only the three closest potential control posts that fall within plus or minus 5 percentage points distance in terms of the predicted intervention probability.[18]

In Appendix C.4 I report detailed robustness checks for alternative ways of constructing treatment and control group from the chat log. In particular, I test a LASSO to predict intervention probabilities, retain all treatment posts, vary the number of retained control posts and add additional temporal restrictions on the matches. I find that the key results presented in the remainder of the paper are robust to making these modifications.

The procedure results in a set of 178 treatment posts and 370 control posts, thereby leveraging 57% of the potential treatment posts mentioned in the chat log (Table II).[19] Treatment and control posts are published by similar media pages and pertain to similar topics. Table III breaks down the articles in treatment and control group by their topic category. As expected, the majority of articles in both groups talk about refugees or politics. The "miscellaneous" category is a frequent target of hate speech when articles about crimes break, leading users to speculate about the perpetrators' origins.[20] Table XVI in the appendix shows that posts in the treatment and control group were published by the same news media and in comparable proportions.

Prior to intervention, treatment and control group are comparable in a number of key

---

[17] Defining these time intervals for potential control posts requires making an assumption about when the counterfactual intervention would have happened. Here I assume the same time interval from the publication of the article to the intervention as the treatment article. I discuss this and possible alternatives in more detail in Section 5.

[18] Retaining up to three control posts for each intervention post strikes a good balance between similarity of the posts on one hand and statistical power on the other, but the results presented here are mostly robust to keeping only the closest, or the five closest potential control posts.

[19] The other interventions are not used for identification because they do not have sufficiently similar articles from the potential control group that could be used as counterfactual and were not already matched to another treatment post. In Appendix C.4, I report a robustness check in which I retain all treatment posts and obtain qualitatively similar results.

[20] While the two articles from the sports section may seem surprising at first, they in fact report incidences with extreme right wing fans during matches of the German national football team.

observables. Table IV reports summary statistics of treatment and control posts before the start of the intervention and shows no significant differences with the exception of the number of comments which is slightly lower in the control group. However, the magnitude of this difference compared to the total number of comments and reactions is relatively small. Compared to the full sample of posts collected over the observation period, treatment and control posts attracted a lot more activity and received a much higher share of xenophobic comments and likes.

Figure 5 shows that both groups' pretrends in terms of overall activity are closely aligned before intervention. Once an intervention starts, the number of comments and likes per minute almost immediately and persistently diverge.

In addition to a set of treatment and control *posts*, the identification strategy can also be used to assign treatment and control status to *individuals*. Treated individuals are those who commented on an article or liked another user's comment on an article prior to an intervention on that article. Conversely, we can assign users to the control group if they were active on a control post and thus narrowly "escaped" treatment. This assignment rules out that users self-selected into or out of treatment as it would have been very difficult for them to predict whether the counterspeech group would intervene on the treatment or the control post.

Individuals in treatment and control group are indeed almost indistinguishable in terms of observable behavior prior to treatment. Table V compares both groups of users and finds no statistically significant differences in terms of weekly activity levels or share of hateful comments. The only difference seems to be in the number of media outlets a user actively commented on: treatment users are active on slightly more media pages than control users.[21]

Note that the users in treatment and control group are not the average German Facebook user and not even the average user commenting the news on Facebook. Compared to the full sample of individuals in the database, these users are much more likely to be in the right tail of overall activity levels, with an average number of weekly comments and likes of about 5, whereas the average for users who appear at least twice in my sample is 1.3. The fact that both groups of users are so similar despite the fact that treatment and control group are selected solely on post-level characteristics gives additional assurance that they can be used for identification of the treatment effects of the interventions.

4 Impact on individuals' future behavior

4.1 Impact on propensity to engage in hate speech

Are individuals who were subject to a counterspeech intervention less likely to write or condone hateful comments in the future? To answer this question I employ a differences-in-differences strategy using the treatment and control groups of users described in the previous section: I compare users who experienced a treatment event, i.e. users who commented or liked a comment on a post that was targeted by an intervention before its start, to users who experienced a

---

[21]I did not collect demographic information on users and hence cannot provide comparisons along these dimensions.

control event, i.e. users who commented or liked a comment on a control post.

Since even the users in the treatment and control sample only write or like an average of 0.4 hateful comments per week, I aggregate the data to weekly time intervals for each user. Moreover, in most specifications I will focus on the binary dependent variable if a user wrote or liked a hateful comment in a given week, rather than studying the number of hateful comments which is skewed and contains many zeros.[22]

In order to verify that the two groups have similar activity patterns prior to the treatment or control event, I plot the event-study graph of the intervention using the following linear probability model:

$$HateSpeech_{it} = \sum_{\tau=-5}^{5} \delta_\tau \times \mathbb{1}\{t = \tau\} \times Treatment_{ie} + \alpha_i + \beta_\tau + \gamma_t + \varepsilon_{it} \tag{1}$$

Observations are indexed by individual i, the relative time to and from treatment τ during each ten-week window e constructed around treatment and control events, and the calender week t. The subscript e is needed because individuals can witness multiple treatment and control events, a fact that I will investigate further below. The dependent variable is an indicator equal to one if the individual writes or likes a hateful comment in a given week, $Treatment$ indicates whether the event falls into the treatment or control category. α_i, β_τ, and γ_t are individual fixed effects, time relative to the event dummies, and week fixed effects respectively.

Figure 6 reports the coefficients δ_τ of this regression along with 95% confidence intervals based on standard errors clustered at the user level.[23] In the weeks prior to an event, both users who experience a control event and those experiencing a treatment event are equally likely to write or condone a hateful comment in a given week, the coefficients δ_τ for $\tau = 1, ..., 5$ being small and insignificant. Activity patterns quickly diverge upon treatment: the week after being exposed to an intervention, treated users are about 7.6 percentage points less likely to engage in hate speech than the comparison group. The effect persists less strongly for another week, at 1.9 percentage points, before it decays. There still seem to be slight effects four to five weeks after treatment but they are only weakly significant and less robust to alternative ways of clustering the standard errors.

These results imply that the counterspeech interventions are highly effective at deterring users from engaging in hate speech – for a short period of time. Compared to the treatment users' baseline probability to write or like a hateful comment of 0.25, the effects' magnitude is sizable. Note that this temporary change in behavior is similar to the findings of Munger (2017) who reports that calling out users that post racist tweets reduces their likelihood of repeating this behavior for about a month before the effect wears off.

---

[22] The results are robust to the alternative approach of using a Poisson regression, as reported in Table XVII in the appendix.

[23] The standard errors are somewhat larger when clustered at the event level, but the drop at week 1 and 2 remains statistically significant at conventional levels.

4.2 Heterogeneity in interventions' effectiveness

In order to investigate the drivers of this effect and to see which users respond most to the interventions, I collapse the two weeks before and after an event into a pre- and a post-period and perform a standard differences-in-differences analysis. Column 1 and 2 of Table VI confirm the results of the event-study: In the two weeks after the intervention, treated users are on average 5.3 percentage points less likely to engage in hate speech. This result is robust to controlling for the share of hateful comments on the event post, and for an individual's average number of comments and likes per week prior to the intervention, both hateful and not. This confirms that the drop in the probability of spreading xenophobic content is not driven by individuals in treatment and control group engaging in discussions at different frequencies already before the intervention.

The effect seems to be driven mainly by users who do engage in hateful activity, but do so only occasionally. Column 3 of Table VI disaggregates the treatment effect for users who have not previously written or liked a hateful message, users who did do so but less than weekly, and users who engage in this kind of activity more than once a week. This grouping of users approximately corresponds to tertiles of hateful activity. Witnessing a counterspeech intervention has no effect on users who do not engage in hate speech in the first place. At 13.9 percentage points, the effect is strongest for users who write or like hateful comments at less than weekly frequency, i.e. moderately hateful users. Conversely, the treatment is much less effective on extreme users who very frequently engage in hateful activity.

I next turn to the question if the interventions' effectiveness depends on the relative importance of hate and counterspeech on the intervention post. The answer is not clearcut. When interacting the intervention with its size as measured by tertiles of the share of comments on the intervention post that were written by the counterspeech group, the bulk of the effect seems to emanate from medium sized interventions (column 4 of Table VI). Both small and large interventions, on the other hand, seem to have no discernible effect. It is important to note, however, that the level of participation in the interventions is likely to be endogenous. For instance, it is conceivable that users who were active on posts which attracted interventions with overwhelming support are more extreme users who could be less likely to respond to the treatment.[24] The share of xenophobic comments on the treatment post, on the other hand, is highly negatively correlated with the size of the effect. The more hate speech was pronounced on the intervention post, the higher the subsequent drop in an individual's propensity to engage in hate speech (last column). Again, it is important to stress that the variation in the share of xenophobic comments is not random and already observable by individuals at the time they decide to participate in the discussion.

The extent to which an individual responds to an intervention depends on whether that

---

[24]I considered several potential instruments for the size of the interventions: First, I computed the number of the counterspeech group's members that participated in the previous days or week with the idea that a larger group should be able to stage bigger interventions. Second, I tried to identify times of the day on different weekdays during which participation in the interventions would increase, hoping to find a lunch- or coffee break effect. Unfortunately, none of these proved to be sufficiently predictive to be used in an IV strategy.

individual has been singled out during the intervention *specifically*. Interventions consist of members of the group writing counterspeech messages as comments on the targeted article generally (top-level comments) but also of commenting directly to the hateful messages that were made by other users (sub-level comments). This allows to look at the treatment effect for those individuals who received a public reply to their comment by a member of the counterspeech group. This is true for about 54% of the users who were active before an intervention occurred. Table VII introduces an interaction term capturing this situation into the differences-in-differences regressions. While the effect of experiencing an intervention remains negative and statistically significant, its magnitude is much smaller than in the previous regressions. Individuals who received a direct counterspeech reply, however, are an additional seven percentage points less likely to engage in hate speech over the weeks after the intervention. As before, users who only occasionally write or condone xenophobic comments are affected the most by an intervention and there is little additional effect of direct replies for more hateful individuals. How many replies a user receives does not seem to alter the effectiveness of the intervention. In conjunction with the fact that I found no effect of the size of the intervention, this result suggests that the key driver of the effects here is being targeted individually by a another user.

These results are unlikely to be accounted for by deletions of comments or accounts. One could be concerned that participants in the counterspeech interventions report users who engage in hate speech to Facebook which in turn suspends or deletes these individuals. However, I confirm in Appendix C.3 that the results also hold for the subsample of individuals remaining active in the weeks following treatment or control events.

4.3   Impact on individuals' activity

How do counterspeech interventions impact individuals' activity patterns more broadly? Rather than just moderating the content of their comments, users who are affected by an intervention seem to reduce their activity in general. Table VIII contains the results of applying the differences-in-differences estimation to activity as measured by the total number of comments and likes, both hateful and not. The first column shows that treated users become 5.7 percentage points less likely to be active in a given week, which is quite close to the magnitudes of the effect on hate speech. There also seems to be a small but insignificant intensive margin effect on activity, as suggested by column 2. Combining intensive and extensive margin suggests an overall reduction of 6.8% in the number of comments and likes following an intervention.[25]

Similarly to the effects on hate speech, the reduction in activity seems to predominantly come from moderately hateful users that have written hateful messages before, but write them less than once a week. While there still is no evidence for the effects increasing in the size of the intervention, the reduction in activity is driven by the upper two tertiles of interventions, with small interventions not having a significant impact. In contrast to the results on the propensity

---

[25]I use $log(1 + \#comments\ and\ likes)$ as a dependent variable to measure the combined effect on intensive and extensive margin. In Appendix C.2 I show that the results are broadly robust to using a Poisson regression which may be statistically more appropriate but more difficult to interpret in particular in the presence of high-dimensional fixed effects.

to engage in xenophobic comments, the interaction of the intervention with the share of hate speech on the targeted article is small and statistically insignificant.

The decrease in activity is not driven solely by the reduction in hate speech. Instead, treated individuals reduce both xenophobic *and* other activity. Table XVIII in the appendix presents the results of repeating the differences-in-differences analysis using only non-xenophobic activity as the dependent variable. The magnitudes of the effects are slightly smaller than in the previous table, but the patterns in terms of heterogeneity remain broadly the same.

In addition to being less likely to write or like comments in general and xenophobic comments in particular, I find evidence that individuals who were exposed to a counterspeech intervention also shift their remaining activity towards less contentious topics. In order to see whether the interventions render individuals more or less likely to express themselves on certain matters, I disaggregate the data such that I can track each users' activity by topic category of the articles they discuss. I can thus treat each individual-topic-week triple as a separate observation. I then run the following differences-in-differences regression:

$$
\begin{aligned}
Active_{itce} = {} & \delta_{TREAT} \times Treatment_{itce} + \delta_{CONTR} \times Control_{itce} + \\
& + \sum_{\varsigma \in C} \delta_\varsigma \times Treatment_{ite} \times \mathbb{1}\{c = \varsigma\} + \alpha_{ie} + \beta_{ic} + \gamma_w + \varepsilon_{itce}
\end{aligned}
\tag{2}
$$

Here, subscript i denotes individuals, t indicates the two-week period before or after the event, c denotes the topic category in the set of categories C, and e denotes the position in the sequence of events that the individual is exposed to. The covariates are a dummy indicating whether the individual experienced a treatment event on on the given topic, a dummy indicating whether the individual experienced a control event on that topic, and a dummy whether the individual experienced a treatment on *any* topic, interacted with the topic category at hand. The δ_ς coefficients capture how individuals' activity on each topic is affected by an intervention. In addition, the regression controls for individual-event fixed effects, user-topic, and week dummies. The standard errors are clustered within users.

The estimates for the δ_ς coefficients are reported in Table IX. Both treatment and control events are followed by a drop in activity in the topic category in which the event occurred. That activity decreases for both types of events may be explained by individuals having satisfied their desire to express their opinion on a given topic. Consistent with the overall treatment effect, however, the decrease is more pronounced for treatment than for control events. The spillovers of interventions to other topic categories reveal a clear pattern: the effects are negative for topic categories prone to heated discussions and higher shares of hate speech, such as immigration, politics more broadly, and miscellaneous. The exception is formed by articles in the "other" category, for which activity drops as well. Less contentious subjects, such as business, sports, and the weather, see an increase in activity. Taken together, this suggests that interventions induce individuals to withdraw from the debates that are most prone to xenophobia and instead engage in the discussion of topics where hateful comments are more rare.

4.4 Multiple treatments

Individuals can be subject to multiple treatment and control events thus allowing for an analysis of how the magnitudes of the effects vary with the number of treatments. Whereas 74% of users in the sample are exposed to only one event during the observation period, there is still a sizable number of individuals who experience multiple events. However, the sample size quickly decreases in the number of prior events.[26]

The effect sizes could change with the number of treatments due to two possible mechanisms. The first one is a change in the effectiveness of the treatment itself. For instance, the treatment effect could wear off as users learn about a social norm or internalize information, as I will discuss further in in Section 6. The second mechanism could be selection: As I showed in the previous section, users respond to treatments by being less likely to write a hateful comment or indeed any comment at all for a period of time. As a consequence, there is a form of survivor bias in the sense that individuals who are still actively commenting after a treatment and are thus available to be treated another time are precisely those users who were less responsive to the first treatment.

While I cannot disentangle the two mechanisms, the identification strategy explained above allows to measure their combined effect. Since it implies that assignment to treatment or control is as good as random for each event, I can compare the response to treatment of individuals who have the same number of prior treatment events. I thus estimate the treatment effect conditional on remaining active after a given number of treatments rather than the average effect of receiving a given number of treatments. I do so by interacting treatment status with the number of previous exposures to treatment and estimating the following regression:

$$
\begin{aligned}
HateSpeech_{ite} = & \sum_{p=0}^{P} \delta_p Treatment_{ite} \times \mathbb{1}\{PriorTreatments = p\}_{ie} + \\
& + \sum_{p=0}^{P} \nu_p \mathbb{1}\{PriorControls = p\}_{ie} + \alpha_i + \gamma_t + \varepsilon_{it}
\end{aligned}
\tag{3}
$$

Here, subscript i denotes individuals, t indicates the two-week period before or after the event, and e denotes the position in the sequence of events that the individual experiences. The relevant parameters here are the set of δ_p which estimate the effect of interventions conditional on a users prior history of treatments.

Figure 7 plots these estimates for users who had no prior exposure to treatment during the observation period, one prior treatment event, two prior treatment events, and three or more prior treatment events. As the number of prior treatment events increases, the effect of each additional treatment decreases. The effect size of the first treatment has roughly the size of the average treatment effect, but already the second treatment is about half as effective. Beyond that, the loss of precision and significance driven by the shrinking sample size makes it difficult to say exactly how large the treatment effects still are. However, there is a clear pattern in

---

[26]Figure 11 in the appendix shows the number of event and user tuples by the number of times a given user has been previously exposed to treatment and control events at the time of the treatment or control event.

the point estimates suggesting that there is no additional treatment effect beyond the second treatment event.

4.5 Substitution towards more extreme pages

The fact that treated individuals are less likely to engage in hate speech and to comment on the public Facebook pages of news media naturally leads to the question of whether they actually reduce their levels of hate speech or if they just voice their opinions elsewhere, for instance among more like-minded people, in private comments or on fringe outlets. If we care about the size of the audience that xenophobic ideas are able to reach, then the fact that interventions reduce the amount of hate speech on articles of large news media is already very good news. However, one might also be concerned that some individuals could progressively be pushed towards a more radical fringe.

Since the scope of my data only includes the public pages of large German language news media, I cannot speak to what these individuals do or say in private Facebook groups, on other websites, or even offline. I can, however, leverage the fact that even within the 85 outlets in my data, there is considerable heterogeneity in terms of the size of the audience, the editorial mix, the prevalence of xenophobic comments and the likelihood of being subject to an intervention (see Table XV in the appendix). If individuals radicalized as a result of a counterspeech intervention, we could expect to observe a shift of their activity towards outlets that attract more like-minded users. Even if these are not completely radical fringe outlets, they would probably be more appealing to progressively radicalizing individuals than news media with more mainstream commentators.

To investigate if there is empirical support for this radicalization hypothesis I disaggregate the data such that each observation is identified by a triple of individual i, media page p, and two week period t before or after event e. Similar to the analysis of spillovers between topics, I then estimate the following differences-in-differences regression:

$$
\begin{aligned}
Active_{itpe} = {} & \delta_{TREAT} \times Treatment_{itpe} + \delta_{CONTR} \times Control_{itpe} + \\
& + \delta \times Treatment_{ite} + \alpha_{ie} + \beta_{ip} + \gamma_t + \varepsilon_{itp}
\end{aligned}
\tag{4}
$$

The dependent variable is a dummy indicating if an individual liked or wrote a comment on a given page. The covariates are a dummy indicating whether the individual experienced a treatment event on the given page, a dummy indicating whether the individual experienced a control event on that page and a dummy whether the individual experienced a treatment on *any* of page, as well as individual-event fixed effects, user-page, fixed effects and time dummies.

I find little evidence in the data that would suggest that individuals switch to pages with higher levels of hate speech or a lower probability of being targeted. The first column of Table X contains the results of estimating regression 4. The effect of an intervention is not just contained to the page on which the intervention took place, but spills over to other pages as well. The second column shows that the effect seems to be larger for pages with a bigger audience. This is presumably because users are more likely to be active on these pages in the first place, which

is why I retain this interaction term as a control in the remaining regressions.[27] In the third column, I introduce an interaction with the total number of counterspeech interventions on a page. While the point estimates of the effects are largest for pages that were especially often targeted by interventions, the differences between the coefficients are very small compared to the average magnitude of the effect: a 1.3 percentage point difference between pages that never see an intervention versus those who do so most frequently. The last column shows that after an intervention, users decrease their activity most on pages that have a high share of xenophobic comments. The magnitude of this interaction is again very small, but lends little support to the idea that individuals radicalize in response to the intervention.

5   Immediate impact on targeted articles

The identification strategy outlined in Section 3 can also be used to asses the more immediate impact of interventions on the targeted articles' discussions. The analyses presented in the previous section addressed the question of how the future behavior of individuals already actively participating in a debate responds to an intervention. This section complements these findings by shedding light on users' aggregate response to a given article, including those individuals that were not yet participating in the discussion before the intervention. I will show that counterspeech interventions do not seem to significantly decrease the total *number* of hateful comments on the targeted articles, but do attract more moderate users beyond the circles of the intervention group to the debate, thereby leading to an overall decrease of the *share* of hateful comments.

This can be achieved by comparing articles that were targeted by an intervention to the set of control articles that were considered for intervention by the group and performing event-study and differences-in-differences analyses. However, the timing of the counterfactual interventions poses an additional challenge that needs to be addressed for this purpose. The responses to articles on Facebook follow a highly cyclical pattern: when an article is posted, it quickly reaches an attention peak, usually within the first 30 minutes before activity slowly wears off over the subsequent couple of hours. As a result, identification of the impact of the intervention requires not only assuming *on which posts* the counterfactual interventions would have taken place, but also *when* they would have taken place.

A naive counterfactual timing would consist of attributing the actual time of intervention to the counterfactual intervention. However, since articles are posted at different times by different news outlets, this would lead to confounding treatment and life-cycle effects. Instead, I measure the time it took between the article's publication and the actual intervention and apply the same time difference to the control posts: if an intervention was announced t minutes after the targeted article was posted on Facebook, then I assume that the counterfactual interventions would have occurred t minutes after the publication of the control posts. An alternative strategy that leads to broadly similar results would have been to predict the time of intervention based

---

[27]Figure 12 in the appendix shows the spillover-effect for each page in the sample for completeness.

on observables of the article such as the media outlet, the number of comments and likes and the share of hateful comments. Empirically that strategy leads to less well-aligned pretrends which is why I focus on the simpler method here.[28]

Figure 8 compares the average number of comments and likes per minute that contain xenophobic messages on treatment and control posts. The two curves track each other closely prior to the intervention lending credibility to the identifying assumptions. However, contrary to the patterns in terms of the total number of comments and likes documented in Figure 5, there seems to be no sharp divergence following the intervention.

In order to test this more throughly, I estimate the following event-study model:

$$Y_{pt} = \sum_{\tau=-T}^{T} \delta_\tau \times \mathbb{1}\{t = \tau\} \times Treatment_p + \alpha_p + \beta_t + \varepsilon_{pt} \tag{5}$$

Where Y_{pt} is an characteristic of article p at t minutes after treatment, for instance the average number of comments per minute. α_p and β_t are article and time-since-announcement fixed effects. The coefficients of interest are δ_τ which will capture the difference between treatment and control group. For $\tau \leq 0$, δ_τ is expected to be indistinguishable from zero if treatment and control posts are truly comparable. If that is the case, we should be able to interpret δ_τ with $\tau > 0$ as the causal effect of the intervention on Y.

As a first step, I confirm that the announcements of interventions by the counterspeech group did trigger higher activity levels. Figure 9a plots the δ_τ coefficients of the event-study regression with the log of the number of comments and likes as a dependent variable. Consistent with the identifying assumptions, the point estimates are small and statistically insignificant prior to treatment. Within ten minutes after the interventions were announced by the group's moderators, there is an increase in the activity levels which becomes statistically significant at conventional levels 20 minutes after treatment. At its peak, the effect corresponds to a 50% increase in the number of comments and likes.

Interestingly, the increase in activity following the intervention does not seem to be entirely explained by the members of the counterspeech group carrying out the intervention. A ripple-on effect becomes apparent when estimating the event-study regression with non-intervention activity as a dependent variable, i.e. comments and likes by users who are not members of the counterspeech group and did not participate in interventions before. Figure 9a shows that there is an increase in activity not directly attributable to the group after about an hour after the intervention starts. Further inspection reported in Appendix C.5 indicates that this is driven by an influx of users who were not active on the post before, rather than by the initial commentators replying to the intervention messages.

An event-study regression on the log number of xenophobic comments and likes suggests a slight decrease in response to an intervention by up to to 17% after about an hour after the treatment starts (Figure 9b). This result should not be over-interpreted, however, as it

---

[28]Yet another alternative is to use the time of discussion among the counterspeech group's moderators instead of the articles' publishing time as reference point. The pretrends seem to align less well with this method, however.

is relatively sensitive to the exact timing of the counterfactual intervention. Still, it is clear that there is no *increase* in the absolute number of hateful comments and likes on the targeted posts, which is remarkable given the fact that there is such a strong increase in activity and the number of users on the article that are not directly attributable to the counterspeech group. The first two columns of Table XI confirm this pattern in a classical differences-in-differences regression which aggregates the data into a pre- and a postintervention period. Averaged over the entire post-period, the results suggest a 14% decline in the number of hateful comments and likes, with the previous caveat still applying.

As a result, the share of xenophobic activity in the comments and likes that were not written by users active in the counterspeech group declines. Columns three and four of Table XI show that this net share of hateful messages decreases by up to 3% when averaged over the entire post-intervention period. Compared to an average share of 21.2% of xenophobic messages on these posts, this effect is sizable but far from fully eradicating hate speech in this context.

Are the additional users attracted by the intervention less likely to engage in hate speech because they are more moderate in general or because they adapt their behavior depending on the comments that they find on the article? The last two columns of Table XI point to the former explanation. I compute users' average share of hate speech in all comments they made before commenting on the intervention article for all users that have not previously participated in an intervention. Using this share as a dependent variable in the differences-in-differences regression, I find that the composition of users changes in favor of ex ante more moderate users once the intervention is under way.

These results suggest that the interventions have a mobilizing effect on bystanders that lead to overall moderation of the content written in response to the articles. The intervention leads to new users joining the discussion of the targeted article who are less likely to post hateful messages than the initial set of users who were responding to the article. The results also imply that the sharp drop in the likelihood of an individual writing or condoning a hateful message observed in the previous section is not driven by a moderating effect on the intervention article itself. Rather, individuals who are subject to an intervention change their behavior going forward, even if its only for the limited timespan of about two weeks.

The fact that the share of xenophobic comments on an article decreases as a result of the intervention is encouraging. Ultimately, this share may be the more relevant metric than the absolute number of hateful comments: Especially when discussions contain lots of comments, a lower share of xenophobia implies a lower probability that an onlooker would be exposed to these ideas while browsing through the comments.

6 Possible mechanisms

Through which channels do the interventions affect individuals' behavior? Plausible candidates for mechanisms include information provision, individuals learning about social norms through the distribution of others' actions, and fostering social norm compliance through non-monetary

punishment. I discuss each of these possible mechanisms in turn.

6.1 Information provision

The simplest way to conceptualize a counterspeech intervention would be to think of it as providing information. Interventions could convey two different kinds of information. First, counterspeech messages could reveal new information on the contentious subject at hand. If initially xenophobic individuals learned that their statements were based on erroneous beliefs about the world, for instance a much larger number of refugees coming to Europe than there actually is, then an intervention could be effective simply by correcting those beliefs. Second, in a very different vein, targeted individuals could potentially infer from the outraged reactions of others that there may be a risk of having their account blocked or deleted.

The fact that the effect of the counterspeech interventions wears off for individuals who experience multiple such interventions could be interpreted as evidence consistent with this mechanism. If the first time an individual is confronted with counterspeech may still learn new information, it may have already absorbed this information the second time it is exposed to it.

While I cannot rule out that this channel plays a role, I argue that it is unlikely to explain the full extent of the effects observed here. As noted in Section 2.3, only 14% of the counterspeech messages in response to the article contain factual information on the topic such as statistics or facts relevant to the article and most of these are actually drawn from the article itself. This share is even lower, about 4%, for responses to other users' comments made by participants in the counterspeech intervention. Yet it is precisely those responses that seem to drive the largest effect.

The treatment effect's decay over a few weeks is also difficult to reconcile with information provision. We would expect either that individuals are able to retain the information learned from an intervention, or that an additional intervention refreshes their memory. Yet, neither of these patterns emerges from the analyses presented above.[29]

It is similarly unlikely that the counterspeech interventions led targeted individuals to update their beliefs about the risk of being blocked or deleted from specific media pages or from Facebook in general. First, I show in Appendix C.3 that deletions themselves cannot explain the drop in activity, which means that there would be no factual basis to sustain such a belief. Second, there are barely any threats or reminders in the counterspeech group's comments. In only 1% of the direct responses to other users' comments did I find a mention of reporting the user, and there is even fewer in the comments on the article.

---

[29]Moreover, it is not clear whether correcting users' knowledge would be sufficient to alter their behavior so dramatically. Using a survey experiment in a similar context, Barrera Rodriguez et al. (2018) find that fact-checking changes the beliefs about the state of the world of participants who were previously exposed to false claims, but does not alter the conclusions they draw from these false claims.

6.2 Inference of social norms from average behavior

A more plausible mechanism is that the interventions induce individuals to update their beliefs about a prevailing social norm. Counterspeech could shift individuals' perceptions of what type of messages the average user does or does not condone. If individuals derive utility from writing comments that do not deviate too much from average opinion, then this change in perception could lead them to adapt their conduct. This type of behavior has been documented for instance by Bursztyn et al. (2017), who show that respondents are more willing to express racist views once they have learned that more people agree with these statements than they previously thought.

This mechanism would be consistent with the fact that the overarching theme which all the counterspeech messages have in common is indeed to express disagreement with hateful views, be it by explaining the reasons why these views are unacceptable, by making injunctive norm statements or by just providing their own opinion. However, this channel too has difficulty explaining the fact that while being exposed to multiple interventions progressively reduces the treatment effect, the effect of any intervention also decays over time. If individuals inferred a social norm from the behavior of others, then these two findings would be hard to reconcile.

It would also imply that there should be a correlation between the share of counterspeech messages on the targeted article and the effectiveness of the intervention. If counterspeech makes up the vast majority of the comments, then this should shift beliefs more than if there is only a few scattered messages, all else equal. Yet, as I show in Section 4, I do not find such a correlation in the data. The reason for this might be that the size of the interventions is endogenous to the article topic. It could be conceivable that participation in the interventions is higher for topics with a broader consensus. Individuals who still post hateful messages on these topics are maybe more extreme types that would be less responsive to treatment, irrespective of how many people participate in the interventions. In the absence of a good exogenous shifter of participation it is impossible to conclude on whether this specific type of social norms channel would be consistent with the data.

6.3 Social norm enforcement through non-monetary punishment

While both of the aforementioned mechanisms could play a role in explaining the observed effects, the impact on individual behavior seems most closely aligned with the findings of the experimental literature on non-monetary punishment. In the seminal paper of this literature, Masclet et al. (2003) study a repeated public goods experiment in which participants were able to attribute non-monetary "punishment points" to other members of their group. They found that despite these points having no effect on participants' payoffs, subjects used these points to punish low contributors which in turn responded by increasing their contributions. The effectiveness of non-monetary punishment has been replicated in a series of studies and notably with written messages that participants were able to send to other group members.[30]

---

[30]See for instance Ellingsen and Johannesson (2008), Xiao and Houser (2009) and Dugar (2010).

Analogously, counterspeech can be thought of as a form of non-monetary punishment. Members of the intervening group write messages in which they publicly disapprove of the behavior of individuals who write hateful messages. The results presented here are notably consistent with the findings of the non-pecuniary punishment literature in three important aspects. First, I find that the effects of the intervention on individual behavior are primarily attributable to counterspeech comments written directly in response to a user's own comment (see Table VII). What seems to matter is the fact of being individually targeted by an expression of disapproval. Second, the decay of the interventions' effect over time resembles the results typically obtained in experimental public goods games with non-monetary punishment. A closer look at the results obtained by Masclet et al. (2003) for instance reveals that while the option to punish increases contributions rates on average, it does not seem to remedy the decay of cooperation over the course of many rounds of the repeated game. Quite to the contrary, decay might even be steeper. Finally, as noted before, the counterspeech messages consist mainly of expressions of disagreement with and disapproval of the hateful messages that were written by other users and are therefore in line with the idea of punishment. In this respect, their content is quite comparable to the messages that the participants in laboratory experiments wrote in response to other participants' contribution decisions.[31]

Whether non-monetary counterspeech sanctions are effective because their public nature induces shame in the targeted users or because they raise the awareness and salience of an existing social norm cannot be decided based on the findings presented here. While there is a longstanding literature arguing for the effectiveness of shame caused by public sanctions (see for instance Kahan and Posner (1999)), Xiao and Houser (2011) have shown that penalties are more effective when given in public than in private even when anonymity rules out shame as a mechanism. In fact, the findings on non-monetary sanctions described above are obtained entirely in settings that exclude shame as a possible explanation.[32] Rather than by a behavioral response to avoid a negative emotion such as shame, these results can be explained by the sanctions communicating a social norm (Xiao and Houser, 2011) or by raising its salience (Konow, 2000; Xiao and Houser, 2009). The fact that counterspeech is most effective when directed specifically at individual users is consistent both with inducing shame and with communicating the norm most effectively, as replies are almost sure to be read by the targeted individuals. In lack of any variation of the publicity of the sanction, I can cannot determine which of the two mechanisms is at play.

7   Conclusion

In the face of a rising tide of populism and xenophobia, people have often pointed fingers at social media in the popular press and the broader public debate. As evidence of the sometimes

---

[31]See Table 3 of Xiao and Houser (2009). For example: "[...]you need to check your priorities...it's not about the money, it's about sharing what you have and realizing you're not the center of the world" (p.399).

[32]Smith et al. (2002) and others have argued that shame requires some degree of publicity of the transgression. In Masclet et al. (2003), however, actions were communicated without any identifying information, not even anonymized player numbers or IDs.

dramatic real world consequences of online hate speech is becoming increasingly available, it has become more important than ever to understand the drivers that shape this behavior.

In this paper, I show that individuals' desire to behave in ways that will be judged favorably by others is one of these drivers. I demonstrate that German users of the world's largest social media platform, Facebook, reduce their supply of online hate speech in response to organized "counterspeech" interventions by fellow users. To do so, I collected roughly 12 million comments on the public Facebook pages of German language news media and identify hateful comments using recent deep learning techniques. Identification is obtained by comparing the treatment group of news articles and of users that were targeted by an intervention to a plausible control group that was equally likely to be targeted. The latter group is inferred from the internal chat logs of the organizers of these large scale counterspeech interventions.

Comparing targeted individuals to the control group I find that the treated individuals are substantially less likely to engage in counterspeech over a period of two weeks. Rather than moderating the opinions they express, these users are less likely to voice their beliefs altogether and tend to avoid contentious debates. While a number of possible mechanisms could explain the effectiveness of these interventions, the findings seem to be most plausibly explained by the interventions acting as a form of non-monetary punishment that makes individuals fall back in line with a prevailing social norm of acceptable behavior.

I also find that the interventions have a moderating effect on the discussion of the articles that were targeted by the intervention. While not reducing the overall number of hateful comments made on these articles significantly, interventions do attract users to the discussion that express non-hateful views, thereby reducing the share of hateful comments.

If confirmed, the findings presented in this paper carry both good news and bad. The good news is possibly that users of social media do adjust their behavior in response to others, demonstrating that maintaining a dialog even in the context of sometimes toxic online debates has clear benefits. The bad news may be that the mechanism I describe may just as well work the other way around: unchecked hate speech may set off a vicious cycle triggering more hate speech. In this context, the fact that media companies seem to have an incentive to slant news towards articles that trigger hateful reactions in order to gain attention on social media is particularly worrying and highlights the special responsibility that editorial boards need to exercise.

References

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. 2015. "Radio and the Rise of the Nazis in Prewar Germany." *Quarterly Journal of Economics*, 1885–1939.

Barrera Rodriguez, Oscar David, Sergei M. Guriev, Emeric Henry, and Ekaterina Zhuravskaya. 2018. "Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics." *Working paper.*

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." *American Economic Review*, 96(5): 1652–1678.

Bénabou, Roland, and Jean Tirole. 2012. "Laws and Norms." *IZA Discussion Paper*, 6290: 1–44.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching Word Vectors with Subword Information." *arXiv*, , (1607.04606).

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2017. "From Extreme to Mainstream: How Social Norms Unravel."

Bursztyn, Leonardo, Michael Callen, Bruno Ferman, Saad Gulzar, Ali Hasanain, and Noam Yuchtman. 2019. "Political Identity: Experimental Evidence on Anti-Americanism in Pakistan." *Journal of the European Economic Association.*

Dugar, Subhasish. 2010. "Nonmonetary sanctions and rewards in an experimental coordination game." *Journal of Economic Behavior and Organization*, 73(3): 377–386.

Ellingsen, Tore, and Magnus Johannesson. 2008. "Anticipated verbal feedback induces altruistic behavior." *Evolution and Human Behavior*, 29(2): 100–105.

Enikolopov, Ruben, Alexey Makarin, Maria Petrova, and Leonid Polishchuk. 2017. "Social Image, Networks, and Protest Participation."

Gächter, Simon, and Ernst Fehr. 1999. "Collective action as a social exchange." *Journal of Economic Behavior & Organization*, 39(4): 341–369.

Gagliardone, Iginio, Gal, Danit, Alves, Thiago, Martinez, and Gabriela. 2015. *Countering Online Hate Speech.* Paris:UNESCO.

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica*, 78(1): 35–71.

Gentzkow, Matthew, Bryan T. Kelly, and Matt. Taddy. 2017. "Text as Data." *Working Paper.*

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2016. "Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech." *NBER Working Paper*, 46.

Groseclose, Tim, and Jeffrey Milyo. 2005. "A Measure of Media Bias." *Quarterly Journal of Economics*, 120(4): 1191–1237.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and Deliberation within the FOMC: a Computational Linguistics Approach." *The Quarterly Journal of Economics*, 133(2): 801–870.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation*, 9(8): 1735–1780.

Jensen, Jacob, Ethan Kaplan, Suresh Naidu, and Laurence Wilse-Samson. 2013. "Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech." *Brookings Papers on Economic Activity*, 2012(1): 1–81.

Kahan, Dan M., and Eric A. Posner. 1999. "Shaming White-Collar Criminals: A Proposal for Reform of the Federal Sentencing Guidelines." *Journal of Law and Economics*, 42(S1): 365–392.

Kennedy, Patrick J, and Andrea Prat. 2019. "Where do people get their news?" *Economic Policy*, 34(97): 5–47.

Konow, James. 2000. "Fair Shares : Accountability and Cognitive Dissonance in Allocation Decisions." *American Economic Review*, 90(4): 1072–1091.

Lee, Dong-Hyun. 2013. "Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks." *working paper*.

Loughran, Tim, and Bill Mcdonald. 2016. "Textual Analysis in Accounting and Finance: A Survey." *Journal of Accounting Research*, 54(4): 1187–1230.

Masclet, David, Charles Noussair, Steven Tucker, and Marie-Claire Villeval. 2003. "Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism." *American Economic Review*, 93(1): 366–380.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv*, , (1301.3781): 1–12.

Müller, Karsten, and Carlo Schwarz. 2018*a*. "Fanning the Flames of Hate: Social Media and Hate Crime."

Müller, Karsten, and Carlo Schwarz. 2018*b*. "Making America Hate Again? Twitter and Hate Crime Under Trump."

Munger, Kevin. 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior*, 39(3): 629–649.

Noussair, Charles, and Steven Tucker. 2005. "Combining monetary and social sanctions to promote cooperation." *Economic Inquiry*, 43(3): 649–660.

Reuters. 2018. "Why Facebook is losing the war on hate speech in Myanmar."

Reuters. 2019. "Germany fines Facebook for under-reporting complaints."

Siegel, Alexandra A. 2018. "Online Hate Speech."

Smith, Richard H., J. Matthew Webster, W. Gerrod Parrott, and Heidi L. Eyre. 2002. "The role of public exposure in moral and nonmoral shame and guilt." *Journal of Personality and Social Psychology*, 83(1): 138–159.

Xiao, Erte, and Daniel Houser. 2009. "Avoiding the sharp tongue: Anticipated written messages promote fair economic exchange." *Journal of Economic Psychology*, 30(3): 393–404.

Xiao, Erte, and Daniel Houser. 2011. "Punish in public." *Journal of Public Economics*, 95(7-8): 1006–1017.

Yanagizawa-Drott, David. 2014. "Propaganda and Conflict: Theory and Evidence From the Rwandan Genocide." *Quarterly Journal of Economics*, 129(4): 1947–1994.

Ziegele, Marc, Teresa K Naab, and Pablo Jost. 2019. "Lonely together? Identifying the determinants of collective corrective action against uncivil comments." *New Media & Society*.

Appendix

A Machine classification

The algorithm used to categorize users' comments into hateful and non-hateful proceeds is a sequence classification model that proceeds in two steps. First, using an unsupervised approach, the text gets a representation as a sequence of vectors. Then, this sequence is fed into Long Short-Term Memory (LSTM) and a feed-forward neural network that predicts the most likely class.

A.1 From comments to sequences of vectors

A comment can be interpreted as a sequence of vectors, where each vector represents one token in the comment. A token can be a word, an emoticon (e.g. ":)") or punctuation. As a first step, each token is represented by a m-dimensional dummy-vector e_i with the i-th component equal to one and all other components equal to zero, where m is the number of unique tokens in the corpus of comments. Using these representations to make predictions directly, is complicated by the fact that the vectors are very large, sparse and that their relative position in the vector space has little meaning.

To reduce the dimensionality of these vectors and to get more meaningful token representations, I use the Word2Vec algorithm proposed by Mikolov et al. (2013). The algorithm consists of a neural network with a single hidden layer of l neurons that is trained to predict a token e_i in a comment based on the surrounding tokens in the same comment. The l weights thus obtained for each token are normalized to unit length and used as new l-dimensional representation w_i of the token.

In addition to having lower dimensionality and being dense, these vector representations have a key feature: The cosine distance between them can be used as a measure of semantic similarity. Intuitively, this comes from the fact that two words that are surrounded by the same set of words often have similar meaning. As the same set of words is predictive for both of them, their vector representations will be similar as well. This is very useful in the context human language where the same idea can be phrased in a myriad of ways, only a fraction of which will be observed in a relatively small training sample. For example, the word "Asylant" (asylum seeker) and "Flüchtling" (refugee) describe similar concepts, yet this relation is absent in their dummy vector representations e_i, e_j. Their representation w_i, w_j, however, based on the data have cosine a cosine similarity $w_i \cdot w_j = 0.84$. In practice this helps the classifier categorize "Asylanten raus" and "Flüchtlinge raus" the same way. Since these representations are learned in an unsupervised manner, they can be learned on a larger training corpus than the hand-classified training corpus so that the classifier can make predictions on comments consisting of words that never occurred in the training sample.

To obtain 300-dimensional token representations, I train the Word2Vec algorithm on all 18.8 million user comments and comments on comments written by users, the 249,000 media

posts, as well as the full text of more than 200,000 news articles referenced int the posts. I exclude tokens that occur less than 50 times and use a context window of ten tokens to predict each token. For illustration, I list the nearest neighbors for a few tokens in table XII. They show that vector representations help deal with common typos and misspellings ("Flüchtling" vs. "Flüchling"), that they capture related concepts (Merkel is the last name of the German chancelor, "kanzlerin") and that they can help identify racist slurs ("Goldstücke", "Kulturbereicherer").

Even using a large training corpus, some words (or indeed misspellings of words) occur too rarely to compute token vectors, despite the fact that they do occur in the comments. Germans seem to be particularly prone to using composite nouns, concatenating several (frequently occurring) nouns to build a new (rarely occurring) noun. To deal with this issue, if a token is encountered that does not have a token vector associated with it, I compute a new token vector by averaging over all known tokens that I can find within the unknown token.[33]

A.2 From token vectors to comment classification

The token vectors computed in step one are then used to turn each comment into a sequence of token vectors that are fed into a LSTM network for prediction. The LSTM is a particular form of a neural network that has been shown to deal well with long-term dependencies (Hochreiter and Schmidhuber (1997)). It reads the token vectors one-by-one and updates its cell state at after each token is read in. How much each token affects the cell state is dependent on the token and all tokens previously seen by the LSTM. The final cell state at the end of a comment is a meaning encoding of that comment. In order to allow the model to also take into account tokens that

LSTM-generated encodings have a number of useful properties which are illustrated using a few example comments in figure 10 which shows the predicted probability that a comment contains hateful content. First, the same token can affect the cell state differently depending on the tokens that preceded in the sequence. The first two examples show that the word "Muslime" (muslims) increase the likelihood that a comment is classified as hateful, but less so in the first example, where it is used in a positive context. Second, they depend on the order of tokens, which avoids ambiguity. Examples three and four contain exactly the same tokens and differ only in their order, which completely changes the meaning of the sentence. In this instance, the classifier correctly classifies both of them. This property would be difficult to achieve using a bag-of-words approach without resorting to very long n-grams that further exacerbate the sparsity problem.

The encodings computed by the LSTM are then passed into a fully connected feed-forward neural network with. At the last step, a normalized exponential function (softmax) is applied so that the network returns a probability that a given comment is hateful. The model is then

---

[33]A similar idea is used by the FastText algorithm (Bojanowski et al. (2017)) that uses within-word n-grams to enrich word vectors. In my tests, however, FastText was outperformed by the approach outlined above, even when using only syllable n-grams.

trained to minimize the negative log-likelihood.

A.3   Parametrization and performance of the classifier

The model that achieved the best performance consists of two consecutive bidirectional LSTM layers, each with a 50-dimensional output space, and dense layer with 64 neurons. The input sequences are truncated to a maximum of 50 tokens. A ten percent dropout is applied to avoid overfitting in the training phase. Once the model is trained, predictions are made for 5000 thousand unlabeled comments that are not part of the training data. Comments that are assigned with at least 65 percent confidence to either class are added to the training sample along with their probabilities and the model is trained again on the augmented data. This pseudo-labeling approach inspired by Lee (2013) is repeated for two iterations and slightly improves performance. The final classifiers' average performance in a ten fold cross-validation is reported in table XIII.

I find that the exact architecture and parametrization of the model matters relatively little for the predictive power of the model, but that the most important driver of performance are the token embeddings. In a way, the part of the model that is decribed in step two seems to be the proverbial tip of the iceberg which accounts for "only" about 200,000 thousand parameters of the model while the token vectors account for roughly 25 million (approx. 84,000 tokens). What seems to matter most for performance is (1) the fact that the network is recurrent as opposed to using a purely constitutional neural network for instance and (2) that it uses domain-specific token-vectors created directly from task-related text as opposed to widely used net-crawls and Wikipedia dumps.

B   Additional tables and figures

C   Additional results and robustness checks

C.1   Article topics and user responses

Table XX reports the results of regressions predicting the share of hateful comments, the log number of comments and likes and the log number of active users on topic dummies and a media outlet × date fixed effect. Each regression is performed on the full set of pages as well as on the core sample of media outlets, i.e. those pages on which the counterspeech group at least considered to intervene at some point during the observation period.

The results show, perhaps unsurprisingly, that articles which broadly relate to foreigners receive considerably more xenophobic comments and likes than any other topic category. More interestingly, they also receive more attention by users than any other topic category. This correlation holds even within news outlets and is therefore not driven by the composition of topics and outlets. Compared to an article in the "other" category, writing about immigration can be associated with up to twice the number of comments and likes.

Pushing this idea one step further, I regress the log number of comments and likes and the log number of users active on an article on the share of xenophobic comments and likes on the article. Table XXI shows that these correlations remain strongly positive and highly statistically significant even when controlling for topic × date indicators and page × date indicators. A ten percentage point increase in the share of xenophobic comments is associated with a 19-25% increase in activity and 15-20% more users on a given posts. To the extent that the social media teams of news media try to maximize engagement with users, this suggests that they may have incentive to produce content that triggers hateful reactions.

C.2   Robustness: Poisson regression results on activity

In Section 4 I report results showing that individuals who were targeted by a counterspeech intervention are not only less likely to engage in hate speech for a period of time, but also are less active in general. For ease of interpretation, I used $1 + log(\#comments\ and\ likes)$ as a dependent variable. Table XXII shows that the key result that individuals become less active after intervention can also be obtained using a Poisson regression and is therefore not dependent on the specific functional form I imposed.

While the parameter estimates confirm the general pattern presented in the main body of the text, there are some notable differences in the heterogeneity analysis. Using the Poisson model, it does seem that there is an effect on users who write hateful messages more than once a week. Small interventions have a significant effect, while the significance on larger interventions decreases. However, the estimates should be interpreted with care as the inclusion of high-dimensional fixed effects in the regression is likely to introduce incidental parameter bias.

C.3 Deletions of accounts

In Section 4 I showed that individuals who experience a counterspeech intervention are less likely to write or like xenophobic comments in the weeks after, and to write or like comments on news articles more generally. One concern regarding these results could be that the interventions trigger Facebook to block or delete certain users and that the reduction in hate speech and activity simply reflects that these users were deleted as a result of the intervention. This could be the case for instance if the participants in the intervention in addition to writing counterspeech messages also reported users that engage in hate speech to Facebook and the company would act on these complaints.

In order to rule out that this is the main driver of the effect, I restrict the sample to users who write or like at least one comment in the weeks following a treatment or control event and repeat the differences-in-differences analysis. These users were evidently not deleted or banned from the platform and if deletions were to explain the previous findings, we would not expect to see any impact of the interventions on this subsample.

Table XXIII reports the result of the differences-in-differences regression on the subsample of users who remained active in the two weeks following a treatment or control event. There remains a highly significant 5 percentage points drop in the individual's probability of posting or condoning a hateful message. Compared to the baseline results presented in Table VI, the effect is slightly smaller but still comparable in magnitude. Deletions are are therefore unlikely to account for the observed decrease in xenophobic activity.[34]

C.4 Alternative definitions of treatment and control group

In this section I test the robustness of the results presented in the main body of the paper to alternative ways of defining treatment and control groups from the counterspeech group's internal chat log. I obtain the potential control posts by extracting all urls mentioned in the group's chat and matching them to a Facebook post using a unique post identifier contained in the links.[35] In order to identify causal effects of the counterspeech interventions, it is necessary to restrict the full set of posts that were ever discussed in the group's chat to those posts that can credibly serve as counterfactual for the interventions. Here I discuss variations in those restrictions.

Baseline procedure

In Section 3 I outlined a two step procedure to get to a set of treatment posts and likely runner-up posts to be used as controls. In the first step, I predict the intervention probability for all posts in the chat log using a logistic regression based on pre-intervention observables.

---

[34]Anecdotally, there have been many complaints about Facebook's lack of responsiveness to hate speech in general and reports filed by users in particular. Recently, German authorities even decided to fine the company for failure to adequately report instances of hate speech (Reuters, 2019).

[35]I remove posts with less than 10% xenophobic comments, as these were likely posted in the group for other reasons. The results are robust to setting this threshold to 5%.

Specifically, I include the log activity levels at 100 and 30 minutes before intervention, the log number of comments and likes over the last 30 minutes prior to the intervention decision and the share of hateful comments just before the intervention decision and 30 minutes before the intervention decision as predictors. The first column of Table XXIV reports the coefficients of this regression. The difference in the predicted treatment probability is a measure of two posts' similarity. For each treatment post I retain only candidate control costs that are within plus or minus five percentage points of predicted intervention probability. In the second step, I further restrict the set of potential control posts by retaining only the closest three of these control posts for each treatment post.

To test the main results' sensitivity to the restrictions in each of these two steps, I modify them in turn and repeat the complete set of analyses presented above.

Estimating treatment propensities using LASSO

To modify the first step, I run a LASSO logistic regression instead of the standard logit with manually chosen predictors. I tie my hands by letting the algorithm choose from a rich set of pre-intervention observables. The regularization parameter λ is chosen to maximize the area under the receiver operating characteristic curve (AUC) in 5 fold cross-validation. The second column of Table XXIV contains the resulting coefficients as well as the full set of possible predictors. While the results turn out to be very similar to the standard logit in terms of included regressors and predictive power, I chose the standard regression as baseline because it results in slightly better balance in the treatment and control posts and is numerically more stable.

Table XXV presents descriptive statistics for the baseline definition used in the main paper along with the six alternative definitions described in this section. The sample obtained by using a LASSO regression and presented in the second supercolumn is almost indistinguishable from the baseline sample. Treatment and control groups balance well both in terms of articles and users.

Figure 13 replicates the post-level event-study plots presented in Section 5 for the different definitions of treatment and control posts. Broadly, all definitions yield a similar overall pattern comparable to the baseline definition: Interventions lead to a substantial increase in the number of comments and likes on targeted posts and potentially to a slight decrease in xenophobic activity. The results obtained using the LASSO are no exception (light blue diamonds in the figure). The only difference to the baseline results is that the coefficients 40 minutes before intervention are different from zero, but this seems to stem mostly from the fact that some posts have not entered the sample yet at that point in time, rather than from actual pretrends.

The event-study plot in Figure 14 shows the impact of counterspeech interventions on individuals' probability to write or like a hateful comment in a given week. It replicates the results presented in Figure 6 in the main body of the paper. The strong drop in that probability the week following the treatment is highly significant in all alternative definitions of treatment and control group. The LASSO specification (light blue diamonds in the figure) tracks the baseline

results almost exactly, produces comparable persistence, and exhibits even a slightly larger drop in xenophobic activity.

Finally, Table XXVI summarizes the key differences-in-differences regression results from the main paper under the alternative definitions. The top panel of the table summarizes the individual level results. While the magnitudes of the coefficients vary slightly across the definitions, the main finding persists that interventions reduce the probability of users to engage in hate speech. Moreover, this effect seems to be strongest for individuals who were directly targeted with a sub-level comment to their own comment by the counterspeech group. The LASSO results are again very similar to the baseline results. The bottom panel of the same table reports the results of the article-level regressions. Here, the LASSO produces a smaller and statistically insignificant effect of interventions on the number of xenophobic comments on the targeted article, explaining why I cautioned about its interpretation in the main text. The remaining results mimic the baseline.

Retaining all treatment posts

As another modification to the first step, I retain *all* treatment posts and their closest matching control post, irrespectively if they meet the other restrictions. Indeed, I cannot use all treatment posts in the baseline specification because they might not have a suitable control that is not already matched to another treatment. Here I check that the results hold at least qualitatively when relaxing this restriction.

The descriptive statistics presented in the third supercolumn of Table XXV show that keeping all treatments and their closest control posts results in a larger sample in terms of both articles and users, but this comes at the cost of balance. For instance, individuals in the treatment group are significantly more active than those in the control group. Compared to the baseline sample, both subsamples are slightly less active.

This imbalance also becomes apparent in the post-level event-study plots of Figure 13 (blue crosses). There might be slight pretends making the common trends assumption more difficult to defend than in the baseline results. Still, the broad pattern clearly emerges that interventions were followed by a substantial increase in activity while the number of xenophobic comments and likes stayed relatively flat.

The blue crosses in Figure 14 show that as in the baseline specification, retaining all treatment posts yields drop in individuals' propensity to engage in hate speech in response to an intervention. The pre-intervention coefficients are slightly more volatile, however, suggesting that the treatment might be less well isolated than in in the baseline.

The differences-in-differences regression results presented in the third supercolumn of Table XXVI confirm the key results of the baseline specification. While the magnitudes of the effects both at the individual level and at the article level are somewhat smaller than in the baseline, the qualitative patterns described in the main body of the paper remain the same.

Matching only within a time window

As a modification to the second step, I add an additional temporal restriction to the matching by first only looking for matches between articles that were considered for intervention on the same day and then by only retaining those that were discussed within two hours of the actual intervention.

Supercolumns four and five of Table XXV present descriptive statistics for the treatment and control groups thus obtained. Drawing control posts only from the set of posts discussed the same day as the treatment post halves the sample size and leads to slightly bigger differences between treatment and control posts even if these are not statistically significant. The user-level differences remain fairly small although the overall sample exhibits higher levels of xenophobia than in the baseline definition. Further restricting the set of possible matches to only posts that were discussed within a two hour window dramatically diminishes the sample size while simultaneously increasing the differences between treatment and control posts. Moreover, the users in both group are much more likely to write or condone xenophobic messages than in the baseline definition.

Figure 13 shows that the results from the article-level event-study plots presented in Section 5 are robust to only allowing for matches within the same day (green solid triangle) or within a two hour window (green empty triangle). Using these narrow definitions of treatment and control group even suggests a slight decline in the number of hateful activity. The variability in the estimates of the first event-study dummy seem to stem again mostly from sample imbalance rather than actual pretrends.

Similar to the baseline, the individual-level event-study plot in Figure 14 indicates that users are less likely to engage in hate speech as a result of an intervention, even with temporal restrictions on the matches (triangles in the figure). The persistence of the effect might be lower than in the baseline, however. With the small sample obtained by only keeping matches within a two hour window, there is already a slight decrease in hate speech propensity in the period before the intervention which does not appear in the other specifications.

Supercolumns four and five of Table XXVI replicate the key regression results from the main body of the paper using the additional restrictions on the matches. As with the previous robustness checks, the baseline results go through with only slight differences. At the individual level, the effect of experiencing an intervention without being targeted by a direct reply is no longer distinguishable from zero. At the article level, the effect on the total number of hateful comments on targeted articles looses its statistical significance, which is why I have cautioned about its interpretation in the main text.

Varying the number of control posts

As a second modification to the second step, I vary the number of closest matching control posts. In the baseline procedure, I keep up to three of the closest matches for each treatment posts. Here I vary this number by first keeping only the closest one and then by keeping the five closest matches.

The last two supercolumns of Table XXV present descriptive statistics for the samples obtained with these modifications. Retaining only the closest matches in terms of intervention propensity results naturally in a much smaller sample in terms of both posts and users. While in terms of the post-level observables, treatment and control group are even more closely comparable, the characteristics of individuals are less well aligned, which is probably explained by a 25% drop in sample size. Retaining the five closest matches, on the other hand, significantly increases the sample size. With this definition, control posts have somewhat lower levels of activity compared to treatment posts. Treatment and control users remain quite comparable.

Again, the article-level results that interventions lead to an increase in activity while the number of xenophobic comments does not change much remain robust, as can be seen from Figure 13. Including the five closest control posts for each treatment post leads to almost the same results as the baseline despite a much larger sample (empty yellow squares), but the common trends assumption becomes more difficult to defend using more restrictive definition (solid pink square). In general, the pattern that emerges is that the smaller the samples, the more strongly pretends appear in the graphs.

This robustness check too leads to a similar pattern in terms of the impact of interventions on individuals' future behavior (squares in Figure 14). There is a sharp decline in individuals propensity to write or condone xenophobic comments following the treatment. When retaining up to five control posts for each treatment post, this effect seems to be even slightly more persistent than in the baseline.

The last columns of Table XXVI reproduce the main regressions when retaining more or fewer control posts. The baseline results appear to be qualitatively robust for both the individual level regressions presented in the tpo panel and the article level regressions in the bottom panel. Keeping up to five control posts also quantitatively matches the baseline results, while restricting the number of controls to one per treatment lowers the magnitudes of the individual-level effects.

In sum, the robustness checks presented in this section leave me confident that the results of my paper are not driven by selecting specific posts to be included in the analysis but can be obtained using different selection procedures.

C.5 Interventions attract new users to targeted articles

In Section 5 I document that interventions by the counterspeech group increase the activity levels in two ways. First, there is the mechanical effect of the intervention which consists precisely of users writing couterspeech messages. Second, there is a less obvious ripple-on effect consisting of activity by users that did not directly participate in counterspeech interventions before.

Here, I disentangle the drivers of the ripple-on effect by reporting the results of two additional sets of regressions. First, I repeat the event-study analysis using the log of the number of users who comment on the specific article for the first time as a dependent variable. Panel (a) of Figure 15 shows that the intervention announcement leads to an influx of new users to

the article, which is to be expected given that the intervention takes place. It also shows that with a slight time delay, there is an influx of users who have not previously been active in a counterspeech intervention.

Second, I plot the event-study graph for the activity levels of users who had already commented or liked a comment on the article before the intervention was announced. As can be seen in panel (b) of Figure 15, the intervention is not associated with an increase in activity of those users. If anything there seems to be a slight decrease in their activity levels more than one and a half hours after the start of the intervention, which would be consistent with the individual level results presented in Section 4.

Taken together, these two pieces of evidence suggest that the ripple-on effect is in fact driven by new users who are attracted to the article by the counterspeech intervention rather than by users who were already actively commenting on the article and try to shout back at the counterspeech group.

D   Example timeline of an intervention

In order to illustrate the process that leads to a counterspeech intervention, this section provides a specific example of the first intervention of the day on August 8, 2017. In this case the article on *N24/Welt* becomes a treatment post while the article published by *Focus Online Politik* becomes a candidate for the control group.

E Content of the counterspeech comments

To disentangle what exactly drives the effect of the counterspeech interventions, it is useful to look at the messages that were sent by members of the counterspeech group during these interventions. I distinguish between top-level comments that are written in response to the media post, and sub-level comments which are written directly in response to another user's comment.

I manually classified approximately 600 comments into different content categories. The results of this exercise are reported in Figure 17. Each comment was assigned to at least one of the following content types:

- Arguments against hateful statements (*Argument*):
  Author provides a common-sense based argument against hateful comments or a specific hateful comment. Example: "Our ancestors were immigrants, too. So what is your problem with immigrants?"

- Disagreement with hateful statements without arguments (*Opinion*):
  Author voices her disagreement with the hateful comments without providing arguments for her view. Example: "Sad! I did not expect to live to see the day where there is so much xenophobia in this country!"

- New piece of Information to debunk hateful comment (*Info*):
  Author provides a new fact, statistic or source material to the debate. Example: "There are 500 Million people in the EU and we are talking about 40.000 refugees here. Those who claim to feel no longer at home in their own country massively distort the facts."

- Injunctive norm statement (*Injunctive*):
  Author makes a statement invoking a moral norm that ought to be followed. Example: "I do not want racism to become acceptable in this country. It is scientifically baseless, morally wrong and destructive. Racism hurts all of us"

- Descriptive norm statement (*Descriptive*):
  Author argues that the majority of people do not accept hateful comments. Example: "I disagree with you but I accept your opinion. It illustrates, however, why social networks are problematic. In surveys, only 17% of Germans are in favor of the death penalty. The likes for this comment give the wrong impression that the share is much higher."

- Call for more facts (*Debate*):
  Author calls for a more fact based debate, to read the article before commenting or to hold back with sweeping statements until more facts are available. Example: "You should back your speculative statements with facts, otherwise they are worthless!"

- Complaint about the article itself (*Complaint*):
  Author complains about sensationalism or shortcomings of the article. Example: "N24

seems to love to draw attention by publishing incomplete articles that draw attention to specific parts of the population."

- Discredit or insult hateful user (*Discrediting*):
Author launches ad hominem attack on users who made hateful comment or tries to discredit them. Example: "No that's not a fact, it's only a thing one of the five voices in your head is telling you."

- Threat of complaint or legal action (*Report*):
Author tells users of hateful comments that they can or will be reported to Facebook, an employer, or law enforcement. Example: "I will report your comment."

- Unintelligible reference to another comment (*Ref*):
Author responds or refers to another comment that makes it not clearly interpretable in the sense of these categories. Example: "That is what I am reproaching you"

- Other comments (*Other*):
Author's comment does not fall into any of the categories above. Example: "So you are sitting there in the comfort of your own home and feel entitled to comment."

List of Figures

Figure 1: User comments, likes and share of hate speech

Note: This figure reports the daily number of comments and likes (dark blue, right scale) and the xenophobic share therof (light blue, left scale) for each day in the observation period. For illustration, a few key dates are highlighted: the terrorist attack in Barcelona (Aug 17, 2017), the German federal elections (Sep 24, 2017) and the day the talks between major parties to form a new German government failed (Nov 20, 2017).

Figure 2: Histogram of the number of weekly comments and likes by users
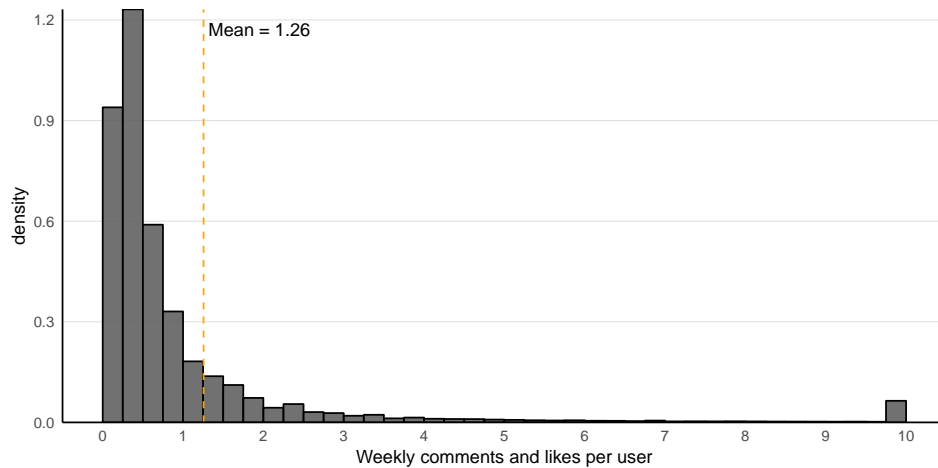
Note: Histogram showing the distribution of the weekly number of comments an likes by user for all users active during the observation period whose first and last activity are at least five days apart.

Figure 3: Counterspeech interventions by share of comments in the post
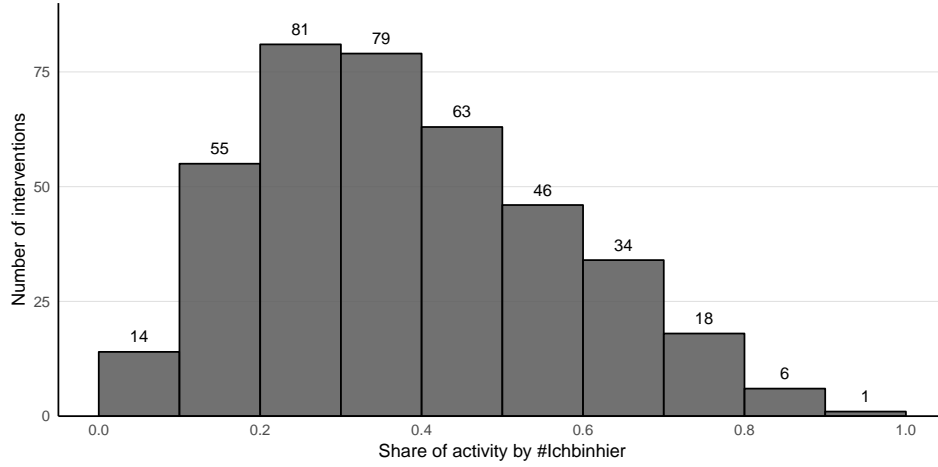
Note: Histogram showing the distribution of the share of activity attributable to members of the counterspeech group *#ichbinhier* on posts the group decided to intervene on.

Figure 4: Content of counterspeech comments
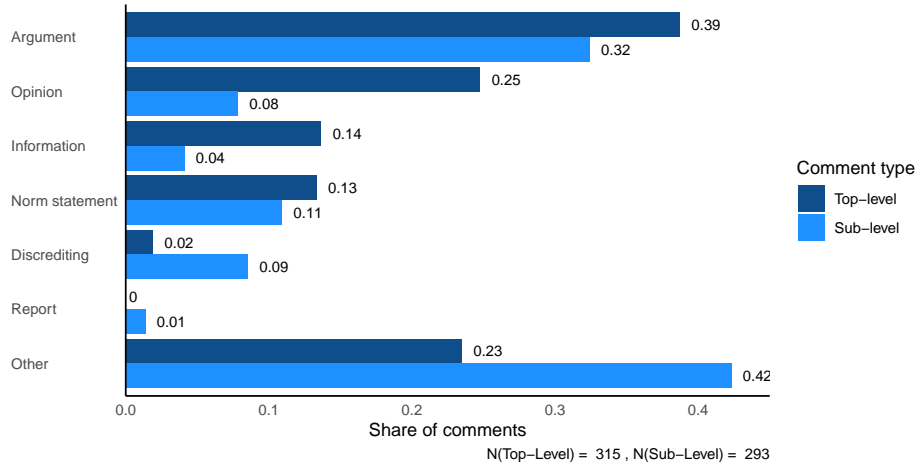
Note: Manually classified comments by participants in counterspeech interventions by type of comment and content of the message. A comment can be assigned to multiple categories. *Argument* contains comments with common-sense based arguments against xenophobic statements, *Opinion* contains comments in which the author voices disagreement with hateful comments without providing arguments, *Information* counts comments that contain a new factual element to the debate, such as a statistic, *Norm statement* contains comments in which author invokes social norm, injunctive or descriptive, *Discrediting* contains ad hominem attacks, *Report* contains threats to report another user to site moderators, Facebook or the police. A more detailed breakdown and examples are available in appendix E.

Figure 5: Evolution of the number of comments and likes per minute in treatment and control group
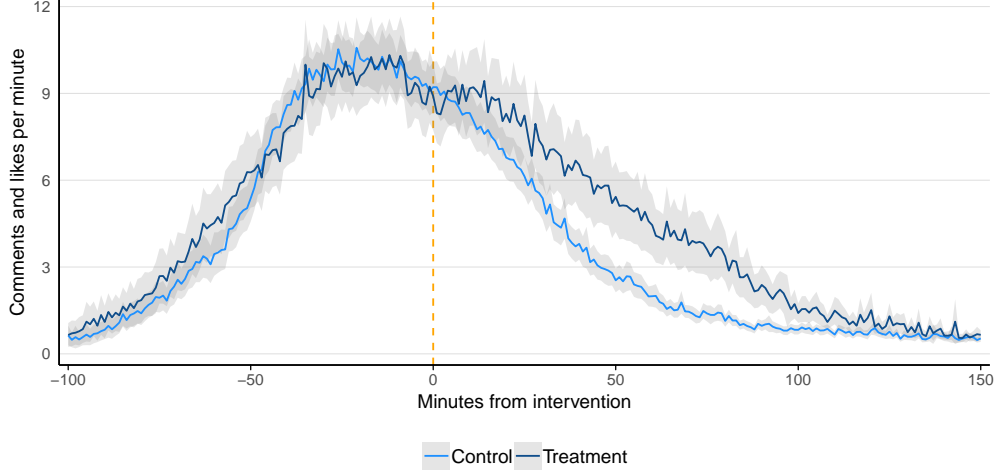
Note: Number of comments and reactions per minute on treatment and control posts by minutes to the announcement of an intervention by the counterspeech group. The shaded gray areas correspond to 95% confidence intervals.

Figure 6: Intervention impact on individuals' propensity to write or like xenophobic comments
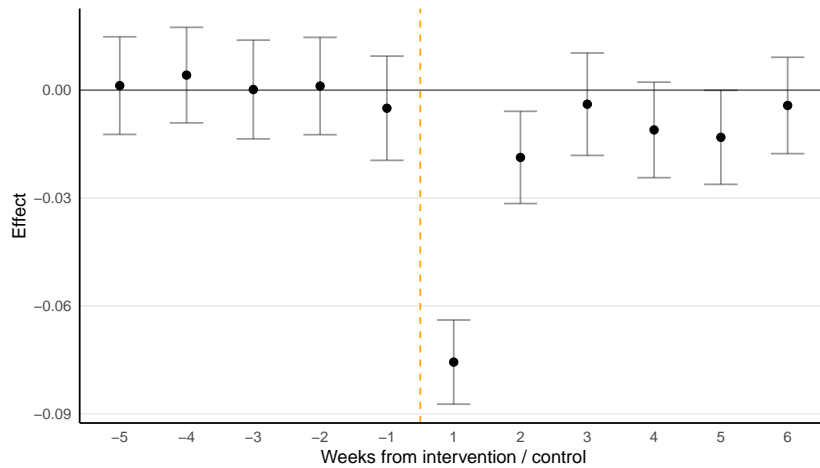
Note: This event-study plot graphs the δ_τ coefficients from regression (1) along with 95% confidence intervals based on user-clustered standard errors. It corresponds to the second column of Table XVII in the appendix, which also contains robustness checks using logistic and Poisson regressions.

46

Figure 7: Diff-in-Diff estimate on xenophobic comments & likes by treatment history
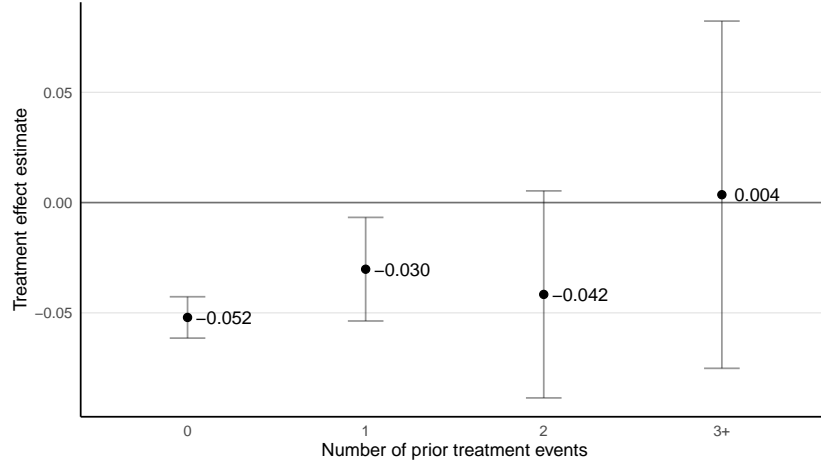
Note: Regression coefficients δ_p from equation 3 with 95% confidence intervals based on standard errors clustered at the user level. The regression results are also reported in more detail in Table XIX in the appendix.

Figure 8: Xenophobic comments and likes per minute in treatment and control group
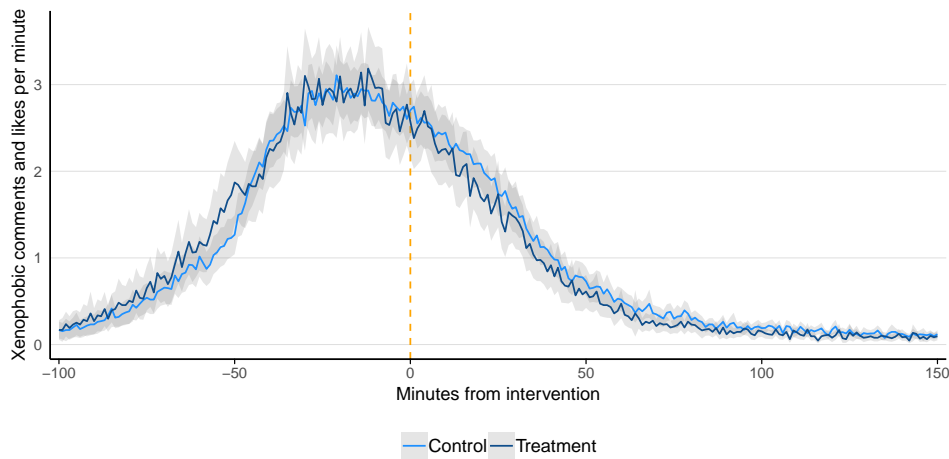

Note: Number of xenophobic comments and reactions per minute on treatment and control posts by minutes to the announcement of an intervention by the counterspeech group. The shaded gray areas correspond to 95% confidence intervals.

Figure 9: Event-study graphs at the article level
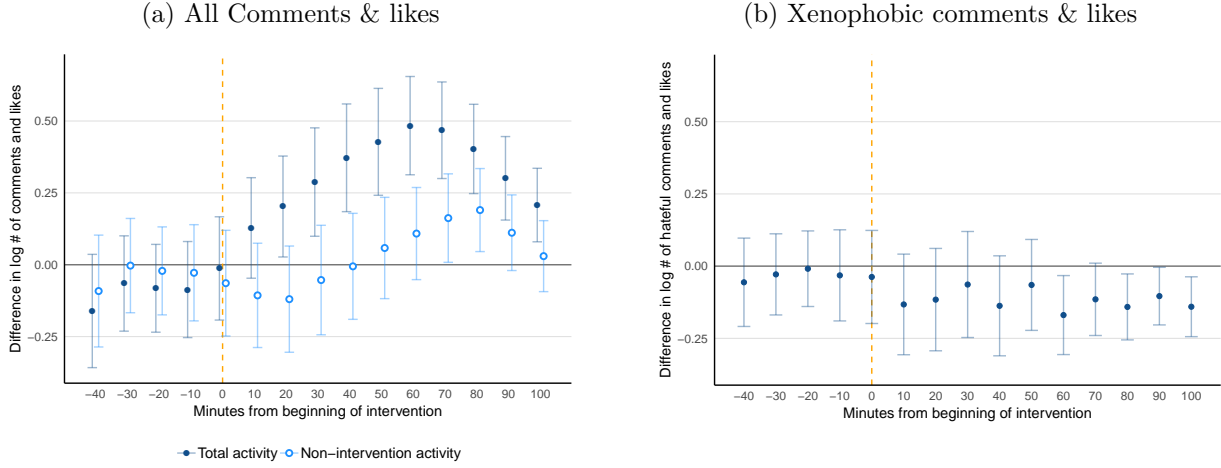
(a) All Comments & likes

(b) Xenophobic comments & likes

Note: Event-study graph corresponding to regression (5) with 95% confidence intervals based on post-clustered standard errors. Regressions include time-trends and post dummies. Non-intervention activity includes all comments and likes which have been written by users who did not write a counterspeech message in the ongoing or a previous intervention.

Figure 10: Classification examples using LSTM
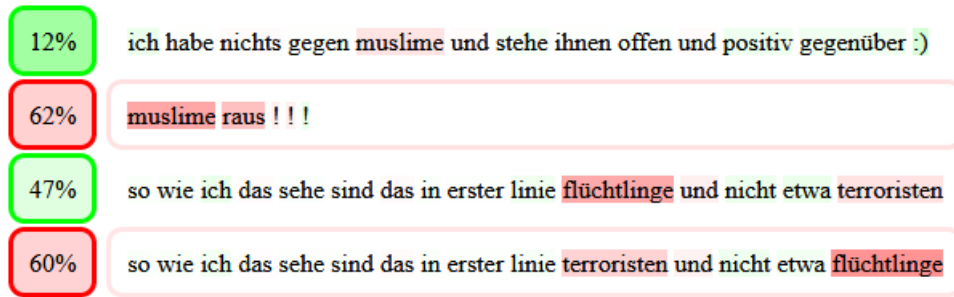
Note: The left column shows the probability that a comment should be classified as xenophobic predicted by the classifier. Each word is colored by the extent to which the predicted probability would be higher (red) or lower (green) if that word was removed from the comment. The comments are inspired by actual comments found in the data but have been modified for illustrative purposes. They translate to : "I have nothing against muslims and I have an open and positive attitude towards them", "Muslims out!!!", "The way I see it these are first and foremost refugees and by no means terrorists", "The way I see it these are first and foremost terrorist and by no means refugees".

Figure 11: Number of user-event tuples by prior treatment exposure

Note: Number of user × treatment/control-event cells broken down by the number of prior treatment and control events that a user has experienced. The top row and rightmost column contain cells with four and more prior treatment events and control events respectively.

Figure 12: Spillover-effect between media pages
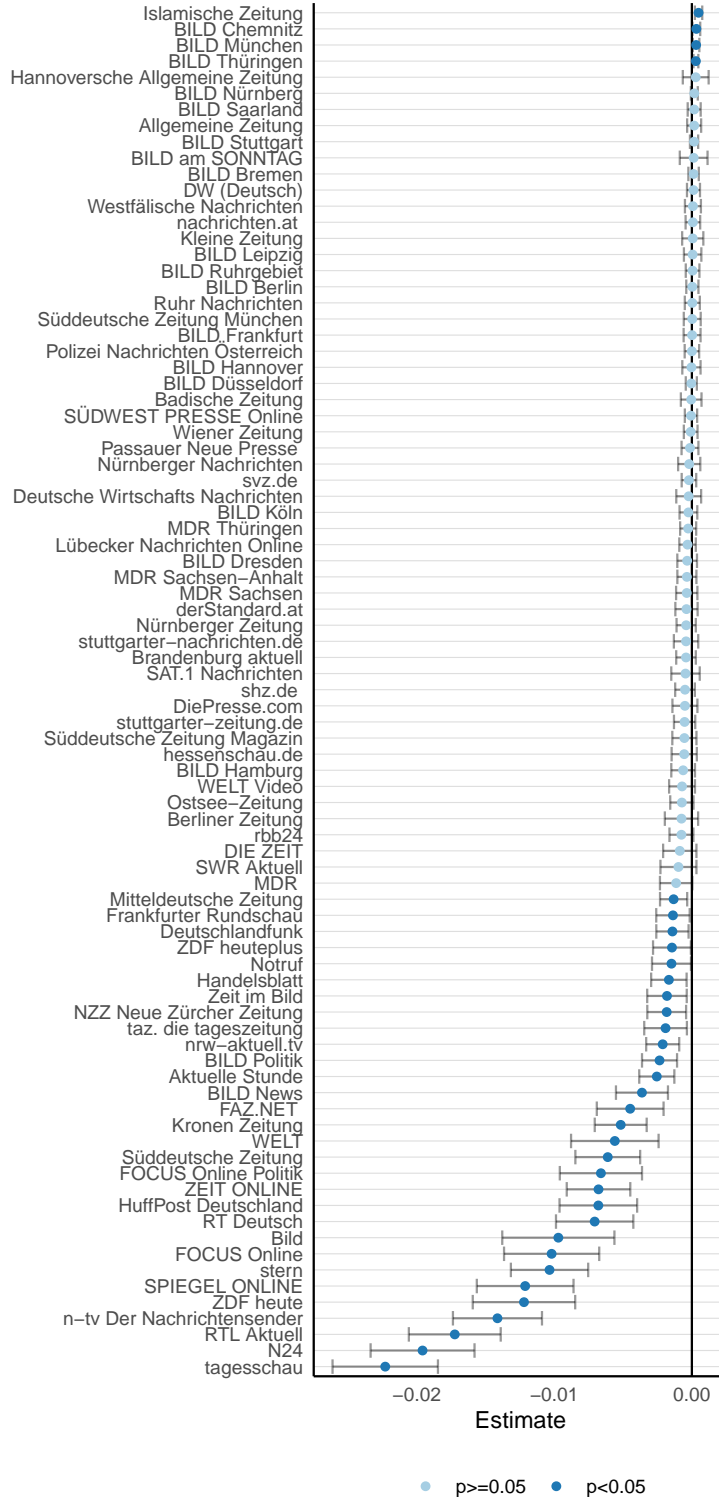
Note: This figure reports the impact a counterspeech intervention on a users probability to write or like a comment on a given Facebook page. It plots the δ coefficient of the differences-in-differences regression $Active_{itp} = \delta_{treat} \times Treat_{itpe} + \delta_{contr} \times Contr_{itpe} + \delta \times Treat_{ite} \times I_p + \alpha_{ie} + \beta_{ip} + \gamma_t + \varepsilon_{itp}$, where I_p is a page dummy. Standard errors are clustered at the user level.

$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

Figure 13: Robustness check on post-level event-study plots

(a) All comments and likes

(b) Xenophobic comments and likes
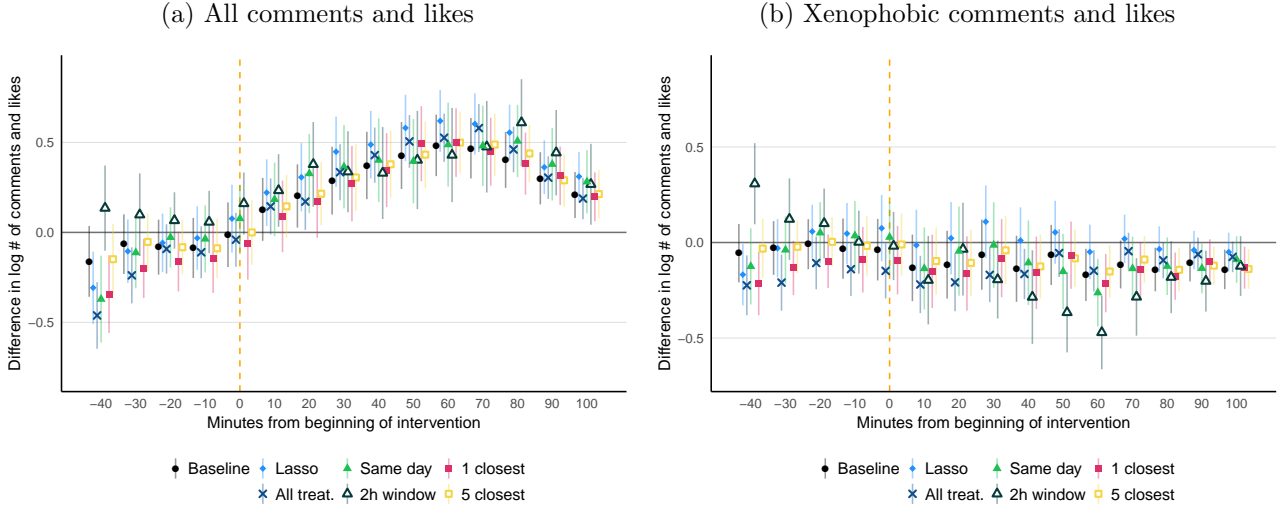
Note: Event-study graph corresponding to Figure 9 (baseline) with alternative ways of defining treatment and control posts. "LASSO" uses the LASSO to compute treatment propensities. "All treat" keeps all treatment articles and the control articles with the closest propensity. "1 closest" and "5 closest" retain only the closest or the five closest matching potential control posts respectively. "Same day" and "2h window" only allow for matching control posts that were discussed the same day as the treatment post or within a two hour window of the treatment post respectively.

Figure 14: Robustness check on user-level event-study plot
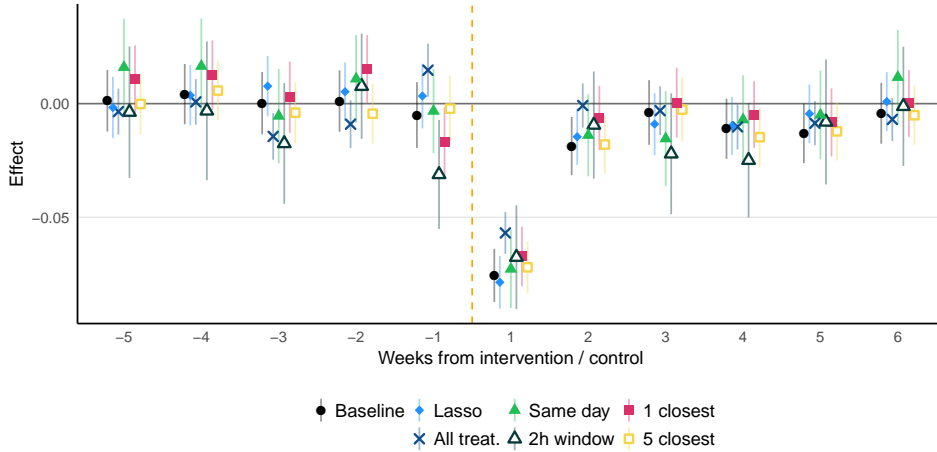
Note: Event-study plot graphs corresponding to Figure 6 (baseline) with alternative ways of defining treatment and control users. "LASSO" uses the LASSO to compute treatment propensities. "All treat" keeps all treatment articles and the control articles with the closest propensity. "1 closest" and "5 closest" retain only the closest or the five closest matching potential control posts respectively. "Same day" and "2h window" only allow for matching control posts that were discussed the same day as the treatment post or within a two hour window of the treatment post respectively.

Figure 15: Event-study plots on drivers of ripple-on effect

(a) Number of users who are new to the article
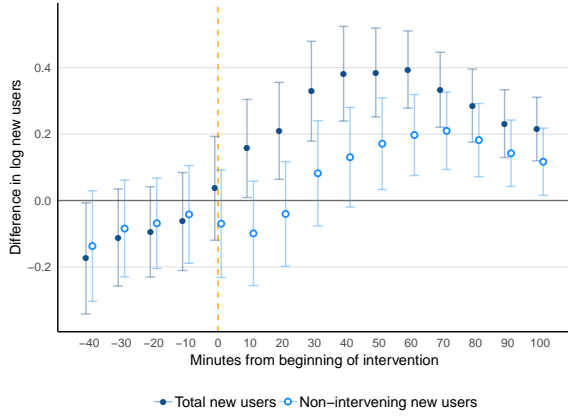
(b) Activity by users active before intervention
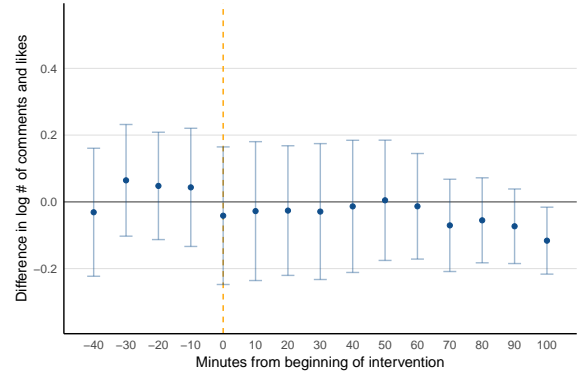
Total new users   Non–intervening new users

Note: Event-study graph corresponding to regression (5) with 95% confidence intervals based on post-clustered standard errors. Regressions include time-trends and post dummies. The dependent variable in the left panel is the number of users who comment on a given post for the firs time. Non-intervention users include all users who did not write a counterspeech message in the ongoing or a previous intervention. The dependent variable in the right panel is the log number of comments and likes written by users who had already written or liked a comment on the article prior to the announcement of the intervention.

Figure 16: Timeline of first intervention on Aug 8, 2017

8:05 *Focus Online Politik* posts article on police intervention at refugee shelter in Bautzen

8:18 *N24* posts article on social democrat saying "Nazis more frightening than refugees to most people"

8:35 Group moderators notice article on *Focus Online*

8:38 Group moderators notice article on *N24*

9:25 Moderators announce Intervention on article on *N24* on group's facebook page and members start to write counterspeech comments.

Figure 17: Content of counterspeech comments
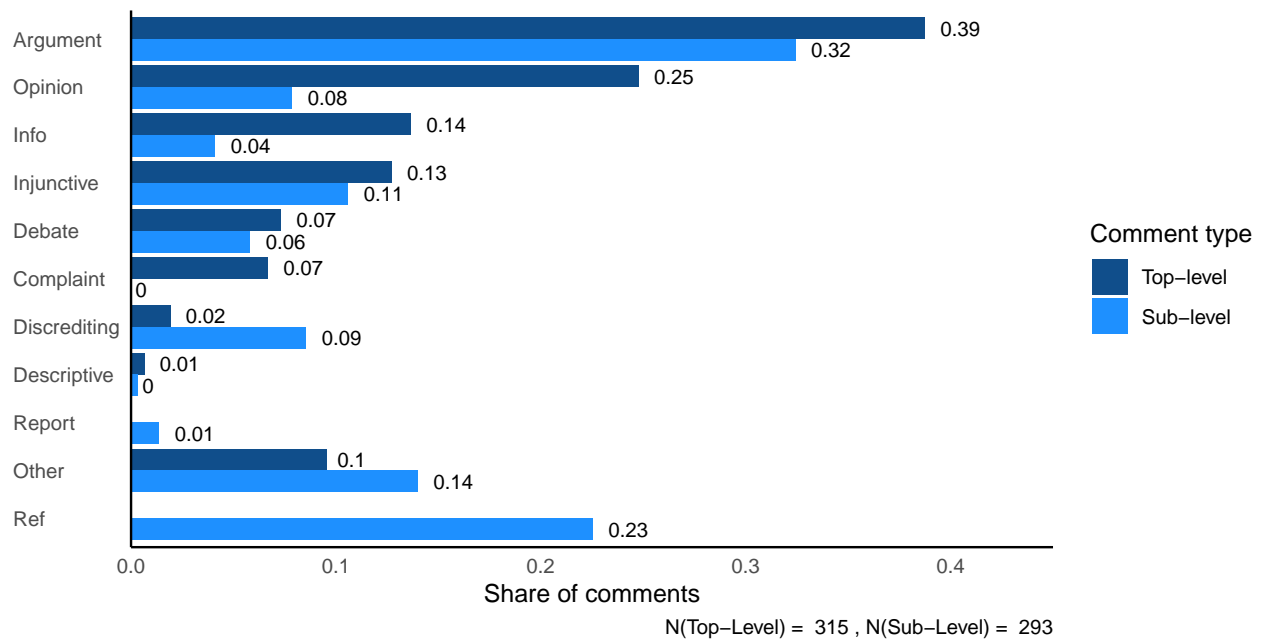
Note: Manually classified comments by participants in counterspeech interventions by type of comment and content of the message.

List of Tables

Table I—: Posts, user activity and hateful comments by article topic

	Articles			Users		
	# Articles	Avg. activity	% Xen.	# Users	Avg. activity	% Xen.
Other	98,657	125.6	2.9	2,696,237	4.6	5.2
Refugees / Foreigners	21,071	270.2	14.8	718,966	7.9	24.1
Politics	51,955	208.3	4.3	995,894	10.9	13.4
Business	16,026	94.1	3.1	502,008	3.0	5.7
Miscellaneous	45,236	117.0	7.7	1,055,183	5.0	12.1
Sports	12,036	81.4	2.1	390,597	2.5	3.9
Weather	4,445	64.6	1.5	186,589	1.5	2.0
Total	249,426	148.2	3.2	3,645,980	10.1	9.6

Note: For each topic category, the columns in the first supercolumn report the total number of articles, the average number of comments and likes per article, the average share of xenophobic comments and likes respectively. The second supercolumn reports the total number of individual users who were active on the topic, users' average number of comments and likes and the share of users who wrote or liked at least one hateful comment. "Miscellaneous" contains articles about accidents, crimes and disasters. "Other" is a catch-all category containing articles about science, human interest stories, celebrity news, cooking, etc.

Table II—: Funnel of potential treatment and control posts

	Treatment	Control	Total
All posts in sample			249,426
All posts in chat log	315	1,727	2,042
Retained posts	178	370	548

Note: Summary of the number of potential treatment and control posts. The first row reports the total number of media posts in the sample. The second row contains all those posts that are mentioned in the chat log of the counterspeech group's leaders. They are divided into posts on which an intervention took place (treatment) and the rest (control). The last row reports the number of posts that are retained using the restrictions explained in Section 3.

Table III—: Posts by post's topic in treatment and control group

	Treatment		Control	
	No.	%	No.	%
Business	2	1.1%	6	1.62%
Miscellaneous	25	14.0%	69	18.65%
Other	21	11.8%	62	16.76%
Politics	33	18.5%	56	15.14%
Refugees / Foreigners	97	54.5%	175	47.30%
Sports	0	0.0%	2	0.54%
Total	178	100.0%	370	100.00%

Note: Number and column percentages of treatment and control articles by topic category. "Miscellaneous" contains articles about accidents, crimes and disasters. "Other" is a catch-all category containing articles about science, human interest stories, celebrity news, cooking, etc. The "weather" category is empty and therefore not reported.

Table IV—: Posts by post's topic in treatment and control group

	Treatment	Control	Δ	Full sample
Comments	99.2	85.8	-13.4*	16.4
Reactions	399.2	395.9	-3.3	79.7
Users	253.9	255.6	1.7	57.9
Xen. comments (%)	28.1	27.6	-0.5	3.3
Comments with tags only (%)	0.9	0.8	-0.1	2.8
Observations	178	370	548	249,425

Note: The columns report pre-intervention information on the posts in the treatment group, the control group, the difference between the two and the full set of all posts in the sample. The first row reports the average number of comments written on a post prior to intervention, the second one row contains all likes and other reactions to user comments. The third row corresponds to the average number of users. Row four reports the share of comments and likes that are xenophobic. Row five reports the share of comments who only tag or reference another user. This is often done to attract attention of specific users to an article. All values are computed just before an intervention is announced, except for the full sample, where this time is undefined. Instead the values in the last column are computed 52 minutes after a post's publication, which corresponds to the median time between publication and intervention.

Table V—: Observables on treatment and control users

	Treatment	Control	Δ	Full sample
Avg. weekly comments, likes	5.06	4.94	-0.12	1.26
Avg. weekly hateful comments, likes	0.40	0.40	0.00	0.07
Share of weeks w. activity	0.70	0.69	-0.01	0.3
Share of weeks w. hateful activity	0.25	0.23	-0.02	0.01
# of commented media outlets	4.30	4.23	-0.07**	2.27
Observations	20,342	61,508	81,850	1,654,729

Note: The columns report the average pre-intervention activity of users in the treatment and control group, as well as the difference between the two. For comparison, the last contains averages for the full sample of users with activity on at least two days.

Table VI—: Differences-in-differences regression results at the user level

	$\mathbb{1}\{Xen.\ Comment/like\}$				
Intervention	-0.053^{***}	-0.052^{***}	0.010	-0.049^{***}	-0.049^{***}
	(0.005)	(0.005)	(0.006)	(0.006)	(0.005)
× hates ≤ weekly			-0.139^{***}		
			(0.009)		
× hates > weekly			-0.031^{**}		
			(0.011)		
× small Intervention				-0.007	
				(0.010)	
× large Intervention				-0.008	
				(0.010)	
× share xen. comments on article					-0.559^{***}
					(0.031)
Controls		Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes	Yes
Users	101,274	101,274	101,274	101,274	101,274
Observations	248,219	248,219	248,219	248,219	248,219
R^2	0.620	0.620	0.621	0.620	0.622

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. "Hates ≤ weekly" and "Hates > weekly" are dummy variables indicating if an individual wrote or liked less or more than one hateful comment a week prior to intervention respectively. The excluded category is users who have not written or liked a hateful comment before. The excluded category in the forth column is medium sized interventions. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention. $^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

Table VII—: Differences-in-differences regression with narrow treatment definition

	$\mathbb{1}\{Xen.\ Comment/like\}$			
Intervention	-0.016^*	-0.016^*	-0.016^*	-0.016^*
	(0.007)	(0.007)	(0.007)	(0.007)
Intervention × direct reply (SLC)	-0.071^{***}	-0.070^{***}	-0.019	-0.078^{***}
	(0.009)	(0.009)	(0.011)	(0.011)
× hates ≤ weekly			-0.120^{***}	
			(0.012)	
× hates > weekly			-0.001	
			(0.015)	
× log(SLC)				0.006
				(0.005)
Controls		Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes
Users	101,274	101,274	101,274	101,274
Observations	248,219	248,219	248,219	248,219
R^2	0.620	0.620	0.621	0.620

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. "Hates ≤ weekly" and "Hates > weekly" are dummy variables indicating if an individual wrote or liked less or more than one hateful comment a week prior to intervention respectively. The excluded category is users who have not written or liked a hateful comment before. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.

$^*\ p < 0.05,\ ^{**}\ p < 0.01,\ ^{***}\ p < 0.001$

Table VIII—: Differences-in-differences on user activity

			Measure of activity			
	$activity > 0$	$log(activity)$		$log(1 + activity)$		
Intervention	−0.057***	−0.018	−0.068***	−0.015	−0.059***	−0.067***
	(0.003)	(0.009)	(0.008)	(0.011)	(0.010)	(0.008)
× hates ≤ weekly				−0.115***		
				(0.015)		
× hates > weekly				−0.035		
				(0.022)		
× small Intervention					0.006	
					(0.016)	
× large Intervention					−0.056**	
					(0.019)	
× share xen. comms.						−0.092
						(0.054)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
Users	101,274	85,994	101,274	101,274	101,274	101,274
Observations	248,219	213,937	248,219	248,219	248,219	248,219
R^2	0.501	0.801	0.810	0.810	0.810	0.810

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. Activity includes all comments and likes. "Hates ≤ weekly" and "Hates > weekly" are dummy variables indicating if an individual wrote or liked less or more than one hateful comment a week prior to intervention respectively. The excluded category is users who have not written or liked a hateful comment before. The excluded category in the fifth column is medium sized interventions. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table IX—: Substitution patterns

	Measure of activity		
	$\mathbb{1}\{act > 0\}$	log(act)	log(1+act)
Intervention on topic	-0.279^{***}	-0.174^{***}	-0.409^{***}
	(0.005)	(0.012)	(0.007)
Control on topic	-0.175^{***}	-0.065^{***}	-0.301^{***}
	(0.002)	(0.005)	(0.003)
Int. on other topic $\times$ biz. econ.	0.003	0.059^{***}	0.029^{***}
	(0.003)	(0.015)	(0.004)
Int. on other topic $\times$ miscellaneous	-0.028^{***}	-0.0003	-0.042^{***}
	(0.004)	(0.011)	(0.005)
Int. on other topic $\times$ other	-0.021^{***}	-0.060^{***}	-0.089^{***}
	(0.003)	(0.009)	(0.005)
Int. on other topic $\times$ politics	-0.008^{*}	-0.024^{*}	-0.065^{***}
	(0.003)	(0.010)	(0.006)
Int. on other topic $\times$ refugees	-0.080^{***}	-0.001	-0.105^{***}
	(0.004)	(0.013)	(0.007)
Int. on other topic $\times$ sports	0.019^{***}	0.078^{***}	0.054^{***}
	(0.003)	(0.017)	(0.004)
Int. on other topic $\times$ weather	0.029^{***}	0.129^{***}	0.066^{***}
	(0.002)	(0.028)	(0.003)
Controls	Yes	Yes	Yes
User $\times$ event FE	Yes	Yes	Yes
User $\times$ topic FE	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Observations	1,737,533	740,247	1,737,533
R^2	0.779	0.852	0.854

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. The excluded topic category is "other" articles, which contain celebrity news, lifestyle articles, movie reviews, etc. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table X—: Impact of interventions on activity by pages

	$\mathbb{1}\{act > 0\}$			
Intervention on page	-0.259^{***}	-0.254^{***}	-0.249^{***}	-0.254^{***}
	(0.004)	(0.004)	(0.004)	(0.004)
Control on page	-0.157^{***}	-0.157^{***}	-0.157^{***}	-0.157^{***}
	(0.002)	(0.002)	(0.002)	(0.002)
Intervention on other page	-0.0004^{***}	0.029^{***}	0.010^{***}	0.029^{***}
	(0.0001)	(0.001)	(0.001)	(0.001)
$\times$ log(# page followers)		-0.003^{***}	-0.001^{***}	-0.002^{***}
		(0.0001)	(0.0001)	(0.0001)
$\times$ 1-30 treatments			-0.007^{***}	
			(0.0005)	
$\times$ >30 treatments			-0.013^{***}	
			(0.001)	
$\times$ page share of xen. comments				-0.047^{***}
				(0.005)
User $\times$ event FE	Yes	Yes	Yes	Yes
User $\times$ page group FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Observations	21,098,615	21,098,615	21,098,615	21,098,615
R^2	0.844	0.844	0.844	0.844

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. "Page followers" is the number of users who liked a news media page. "1-30 treatments" and ">30 treatments" are dummy variables indicating that a media page was targeted by the corresponding number of interventions during the observation period. The omitted category is pages that were never targeted by an intervention during the observation period. "Page share of xen. comments" is the average share of xenophobic comments and likes on a given media page.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table XI—: Diff-in-diff regression results at the post level

	log(1+# xen. coms)		Share xen. (excl.CS)		User share xen.	
Intervention	-0.106	-0.136^{*}	-0.028^{**}	-0.030^{**}	-0.010^{***}	-0.010^{***}
	(0.066)	(0.064)	(0.011)	(0.011)	(0.002)	(0.002)
Post FE	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes		Yes		Yes	
Time $\times$ page FE		Yes		Yes		Yes
Posts	548	548	548	548	548	548
Observations	20,824	20,824	14,490	14,490	14,349	14,349
R^2	0.594	0.616	0.236	0.274	0.193	0.226

Note: Post-clustered standard errors in parentheses. Share xen. is the number of xenophobic comments and likes which have been written by users who did not write a counterspeech message in the ongoing or a previous intervention divided by the total number of comments and likes by these users. User share xen. is the average share of xenophobic comments made in all previous comments by users who did not write a counterspeech message in the ongoing or a previous intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table XII—: Examples of tokens with similar vector representations

flüchtling	flüchling	merkel	:)		goldstücke
asylant	asylant	murksel	;)		goldjungs
migrant	flüchtling	merkels	:-)		bereicherer
wirtschaftsflüchtling	wirtschaftsflüchtling	mekel	:p		neubürger
kriegsflüchtling	kriegsflüchtling	merkl	;-)		kulturbereicherer
schutzsuchender	wirtschaftsmigrant	kanzlerin	:slightly_smiling_face:	goldstückchen

Note: This table reports the five closest tokens of five example tokens base on the cosine distance between their vector representations obtained by Word2Vec. The English translations are "refugee","refgee" (typo), "merkel" (the last name of the German chancelor Angela Merkel), an emoticon and "piece of gold" - a term that is used as a slur by some users.

Table XIII—: Classifier's average performance in ten fold cross-validation

Performance measure	Classifier performance
Accuracy	94.4%
Area under ROC	93.4%
Specificity	97.3%
Sensitivity	56.9%

Note: All performance measures are computed on the hold-out set in 10 fold cross-validation. Accuracy is the share of correct predictions. The area under the receiver operating characteristic curve is the integral over the curve plotting the true-positive rate against the false-positive rate for each probability threshold of predicting positive outcome. Specificity is defined as the share of correctly predicted negatives in all negatives in the sample. Sensitivity is the share of correctly predicted positives in all positives in the sample.

Table XIV—: Full list of news media included in the data with number of followers and media posts.

Facebook Page	Followers	Posts
Bild	2394020	1188
SPIEGEL ONLINE	1439586	4248
tagesschau	1348413	2167
RTL Aktuell	1110008	1250
N24	1063394	3174
WELT	920983	2271
n-tv Der Nachrichtensender	841519	1699
ZEIT ONLINE	823817	5441
stern	726010	6319
ZDF heute	703057	4879
Süddeutsche Zeitung	689118	820
FOCUS Online	683579	1439
HuffPost Deutschland	644071	5252
FOCUS Online Politik	531325	989
FAZ.NET - Frankfurter Allgemeine Zeitung	486204	1107
Süddeutsche Zeitung Magazin	475061	185
DIE ZEIT	423062	530
DW (Deutsch)	389498	5369
BILD News	388444	3641
Zeit im Bild	365929	399
derStandard.at	296121	3453
RT Deutsch	288129	3641
Kronen Zeitung	269631	5014
taz. die tageszeitung	268264	801
WELT Video	253376	2587
Handelsblatt	213681	1029
Berliner Zeitung	179554	1244
nrw-aktuell.tv	171853	1371
Aktuelle Stunde	170882	1284
DiePresse.com	163588	2210
Deutschlandfunk	161246	1805
ZDF heuteplus	154607	2351
NZZ Neue Zürcher Zeitung	152981	2268
Kleine Zeitung	137012	2515
BILD am SONNTAG	134502	2304
MDR - Mitteldeutscher Rundfunk	103755	2619
Deutsche Wirtschafts Nachrichten	100588	1083
BILD Hamburg	100507	1940
Nürnberger Nachrichten	100460	1186
shz.de - Nachrichten aus Schleswig-Holstein	95570	1416
SWR Aktuell	94627	1469
Mitteldeutsche Zeitung	94426	676
MDR Sachsen-Anhalt	92871	3319
Notruf	88805	612
Passauer Neue Presse - PNP	86417	5257
Ostsee-Zeitung	86009	841
stuttgarter-nachrichten.de	85695	2460
Frankfurter Rundschau	85325	5719
Islamische Zeitung	79899	415
Hannoversche Allgemeine Zeitung / HAZ	79684	2078
stuttgarter-zeitung.de	79536	6090
Badische Zeitung	77900	1369
hessenschau.de	77751	3970
SAT.1 Nachrichten	75797	1705
Nürnberger Zeitung	75601	8662
nachrichten.at – Oberösterreichische Nachrichten	73767	4418
Süddeutsche Zeitung München	72771	2300
Lübecker Nachrichten Online	68862	4266
Westfälische Nachrichten	67752	2244
Polizei Nachrichten Österreich	65345	2578
Ruhr Nachrichten	62447	882
svz.de - Nachrichten aus Mecklenburg-Vorpommern	58809	7647
rbb24	55648	6437
MDR Sachsen	50694	9458
Allgemeine Zeitung	50018	3944
BILD Dresden	48887	3241
Brandenburg aktuell	43692	7176
MDR Thüringen	42277	6463
BILD Leipzig	40655	563
SÜDWEST PRESSE Online	39273	575
BILD Köln	38512	9103
FINANCIAL TIMES DEUTSCHLAND	36811	1372
BILD Frankfurt	35793	5993
Wiener Zeitung	35165	613
BILD Saarland	29010	1798
BILD Bremen	26550	188
BILD Politik	22890	473
BILD Hannover	22809	476
BILD Ruhrgebiet	20777	1328
BILD Düsseldorf	18064	4703
BILD Berlin	16034	5001
BILD München	14763	7401
BILD Thüringen	11234	5319
BILD Chemnitz	9348	4027
BILD Stuttgart	8907	996
Total		250113

Note: This table lists all 85 Facebook pages that were monitored from August 2017 through January 2018. The number of followers was recorded in July 2017. It includes most major German news outlets along with some smaller regional outlets. A few fringe pages were included as well.

Table XV—: Posts, user activity and hateful comments by news media

		Articles		Users	
	Total activity	# Articles	% Xen.	# Users	% Xen.
Bild	5,322,821	8,662	4.4	1,243,642	5.4
Spiegel Online	2,411,631	6,464	2.6	353,357	6.3
N24	2,195,287	5,318	8.8	283,451	12.4
Focus Online	2,102,473	9,447	7.8	168,066	18.5
Tagesschau	2,062,436	3,316	3.6	311,122	8.3
Focus Online Politik	1,974,441	5,994	10.0	97,311	28.2
Rt Deutsch	1,823,520	9,104	9.7	140,440	27.3
Welt	1,590,224	4,027	5.0	224,288	9.0
Zdf Heute	1,386,921	3,174	9.2	161,187	14.0
Rtl Aktuell	1,342,266	5,437	7.0	250,383	11.1
Huffpost Deutschland	1,263,817	6,439	10.2	95,350	22.4
Kronen Zeitung	1,114,944	4,988	10.5	78,828	21.9
Süddeutsche Zeitung	1,020,689	5,721	2.7	154,490	6.2
Frankfurter Allgemeine Zeitung	966,820	7,648	5.5	86,767	12.8
N-TV	888,942	4,877	7.2	115,299	16.4
Other	9,508,964	158,810	3.8	2,058,838	7.1

Note: Number of posts, user reactions and the xenophobic share thereof for the top 15 media pages by user activity. The columns report the total number of comments and likes per media outlet, the total number of articles, the average share of xenophobic activity, the total number of users and the share of users who wrote or condoned a xenophobic comment at least once. Note that the fact that the tabloid Bild has such a low share of xenophobic content is driven by the fact that many of its posts are celebrity news. Where they do post topics that are more prone to hate comments, this share is higher.

Table XVI—: Posts by media page in treatment and control group

	Treatment		Control	
	No.	%	No.	%
Zdf Heute	39	21.9%	54	14.6%
N24	22	12.4%	51	13.8%
Bild	18	10.1%	41	11.1%
Tagesschau	20	11.2%	37	10.0%
Huffpost Deutschland	11	6.2%	26	7.0%
Spiegel Online	10	5.6%	25	6.8%
Focus Online	10	5.6%	24	6.5%
Focus Online Politik	7	3.9%	24	6.5%
Kronen Zeitung	9	5.1%	22	5.9%
Rtl Aktuell	11	6.2%	17	4.6%
Other	21	11.8%	49	13.2%
Total	178	100.0%	370	100.0%

Note: Number and column percentages of treatment and control articles by media outlet. Only the 10 most frequently targeted media outlets are reported.

Table XVII—: Intervention impact on individuals' propensity to write or like xenophobic comments

δ_τ	OLS		Logit		Poisson	
$\tau = -5$	0.001	0.001	0.033	0.068	0.076	0.079
	(0.007)	(0.007)	(0.038)	(0.067)	(0.055)	(0.055)
$\tau = -4$	0.004	0.004	0.068	0.112	0.051	0.053
	(0.007)	(0.007)	(0.036)	(0.062)	(0.051)	(0.050)
$\tau = -3$	0.0003	0.0002	0.041	0.085	-0.030	-0.027
	(0.007)	(0.007)	(0.037)	(0.064)	(0.049)	(0.048)
$\tau = -2$	0.002	0.001	0.039	0.062	-0.036	-0.034
	(0.007)	(0.007)	(0.035)	(0.061)	(0.044)	(0.044)
$\tau = -1$	-0.005	-0.005	-0.057	-0.099	-0.048	-0.045
	(0.007)	(0.007)	(0.034)	(0.061)	(0.032)	(0.031)
$\tau = 1$	-0.075^{***}	-0.076^{***}	-0.466^{***}	-0.818^{***}	-0.488^{***}	-0.487^{***}
	(0.006)	(0.006)	(0.029)	(0.051)	(0.039)	(0.039)
$\tau = 2$	-0.018^{**}	-0.019^{**}	-0.094^{**}	-0.162^{**}	-0.117^{**}	-0.118^{**}
	(0.007)	(0.007)	(0.031)	(0.053)	(0.039)	(0.039)
$\tau = 3$	-0.004	-0.004	0.008	0.004	-0.004	-0.009
	(0.007)	(0.007)	(0.037)	(0.063)	(0.044)	(0.045)
$\tau = 4$	-0.011	-0.011	-0.006	-0.005	0.011	0.010
	(0.007)	(0.007)	(0.035)	(0.061)	(0.048)	(0.048)
$\tau = 5$	-0.013	-0.013^{*}	-0.037	-0.060	-0.045	-0.041
	(0.007)	(0.007)	(0.035)	(0.061)	(0.051)	(0.051)
$\tau = 6$	-0.004	-0.004	0.030	0.051	-0.004	-0.002
	(0.007)	(0.007)	(0.035)	(0.060)	(0.048)	(0.048)
Controls		Yes		Yes		Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Time to event FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	426,979	426,979	221,325	221,325	227,204	227,20

Note: This table reports the results of different specifications for the event-study regression described in Section 4. The coefficients in the second column are plotted in Figure 6. The first two columns contain the linear probability model described in equation 1. The middle two columns report the results of logistic regressions with the same dependent variable. The last two columns contain the results of Poisson regressions where the dependent variable is the number of xenophobic comments written or liked by an individual in a given week. User-clustered standard errors in parentheses. The number of observations excludes singletons and, for the GLMs, perfect classifications. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table XVIII—: Diff-in-Diff estimate on non-xenophobic comments & likes

	Measure of non-xenophobic activity					
	$activity > 0$	$log(activity)$	$log(1+activity)$			
Intervention	−0.052***	0.002	−0.046***	0.004	−0.036***	−0.047***
	(0.003)	(0.010)	(0.008)	(0.011)	(0.010)	(0.008)
× hates ≤ weekly				−0.107***		
				(0.015)		
× hates > weekly				−0.041		
				(0.022)		
× small Intervention					0.009	
					(0.016)	
× large Intervention					−0.065***	
					(0.019)	
× share xen. comms						0.157**
						(0.054)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes	Yes	Yes
Users	101,281	99,953	101,281	101,283	101,283	101,282
Observations	248,219	224,415	248,219	248,219	248,219	248,219
R^2	0.534	0.811	0.807	0.808	0.807	0.807

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table XIX—: Diff-in-Diff estimate on xenophobic comments & likes by treatment history

	$\mathbb{1}\{Xen.\ Comment/like\}$	
No previous treatment	−0.051***	−0.052***
	(0.005)	(0.005)
1 previous treatment	−0.035**	−0.030*
	(0.012)	(0.012)
2 previous treatments	−0.045	−0.042
	(0.024)	(0.024)
≥3 previous treatments	−0.003	0.004
	(0.040)	(0.040)
Controls		Yes
Control event FE	Yes	Yes
User FE	Yes	Yes
Period FE	Yes	Yes
Users	101,274	101,274
Observations	248,219	248,219
R^2	0.620	0.620

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. The coefficients in the second column are plotted in Figure 7. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table XX—: User response to articles by topic category

	Share hate coms.		log(activity)		log(#users)	
Business / Economics	0.001	−0.001	0.053	0.075*	0.034	0.049
	(0.001)	(0.001)	(0.029)	(0.038)	(0.027)	(0.035)
Miscellaneous	0.020***	0.027***	0.106***	0.131*	0.082**	0.096
	(0.002)	(0.002)	(0.029)	(0.054)	(0.028)	(0.051)
Politics	0.009***	0.011***	0.709***	0.817***	0.570***	0.654***
	(0.002)	(0.002)	(0.064)	(0.101)	(0.056)	(0.089)
Foreigners / Refugees	0.082***	0.093***	0.881***	1.033***	0.728***	0.855***
	(0.004)	(0.004)	(0.068)	(0.106)	(0.060)	(0.093)
Sports	−0.007***	−0.009***	−0.363***	−0.418***	−0.351***	−0.399***
	(0.001)	(0.002)	(0.046)	(0.083)	(0.043)	(0.077)
Weather	−0.007***	−0.010***	−0.480***	−0.700***	−0.424***	−0.618***
	(0.001)	(0.002)	(0.083)	(0.135)	(0.078)	(0.130)
Page × date FE	Yes	Yes	Yes	Yes	Yes	Yes
Pages	All	Core	All	Core	All	Core
Observations	248,913	121,377	248,913	121,377	248,913	121,377
R^2	0.144	0.147	0.412	0.318	0.419	0.328

Note: Page-clustered standard errors in parentheses. The number of observations excludes singletons. The excluded topic category is "other" articles, which contain celebrity news, lifestyle articles, movie reviews, etc. The first, third and fifth column include all pages in the regression. The second, fourth and sixth column restrict the sample to core pages which have at least one article that was considered or targeted by the counterspeech group.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table XXI—: User response to articles by level of hate speech

	log(activity)		log(#users)	
Share of hateful comments	1.927***	2.511***	1.583***	2.044***
	(0.170)	(0.237)	(0.147)	(0.216)
Page × date FE	Yes	Yes	Yes	Yes
Topic × date FE	Yes	Yes	Yes	Yes
Pages	All	Core	All	Core
Observations	248,909	121,377	248,909	121,377
R^2	0.432	0.348	0.436	0.354

Note: Page-clustered standard errors in parentheses. The number of observations excludes singletons. The first and third column include all pages in the regression. The second and fourth column restrict the sample to core pages which have at least one article that was considered or targeted by the counterspeech group.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table XXII—: Differences-in-differences on user activity (Poisson regression)

	Activity (# comments + # likes)				
Intervention	-0.039^{***}	-0.081^{***}	-0.012	-0.054^{***}	-0.080^{***}
	(0.010)	(0.010)	(0.019)	(0.013)	(0.010)
× hates < weekly			-0.093^{***}		
			(0.023)		
× hates > weekly			-0.094^{***}		
			(0.024)		
× small Intervention				-0.095^{***}	
				(0.021)	
× large Intervention				-0.021	
				(0.024)	
× Share xen. comments					-0.345^{***}
					(0.069)
Controls		Yes	Yes	Yes	
User FE	Yes	Yes	Yes	Yes	
Period FE	Yes	Yes	Yes	Yes	
Observations	251,355	251,355	251,355	251,355	251,355

Note: The table reports the parameter estimates of Poisson regressions along with user-clustered standard errors in parentheses. The number of observations excludes singletons. Activity includes all comments and likes. "Hates $\leq$ weekly" and "Hates > weekly" are dummy variables indicating if an individual wrote or liked less or more than one hateful comment a week prior to intervention respectively. The excluded category is users who have not written or liked a hateful comment before. The excluded category in the fourth column is medium sized interventions. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table XXIII—: Differences-in-differences result on users remaining active

	$\mathbb{1}\{Xen.\ Comment/like\}$	
Intervention	-0.050^{***}	-0.049^{***}
	(0.005)	(0.005)
Controls		Yes
User FE	Yes	Yes
Period FE	Yes	Yes
Users	85,502	85,502
Observations	210,415	210,415
R^2	0.621	0.622

Note: User-clustered standard errors in parentheses. The number of observations excludes singletons. Activity includes all comments and likes. The controls include the average weekly pretreatment activity and xenophobic comments and likes, as well as the share of xenophobic comments on the treatment or control article at the time of intervention.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table XXIV—: Intervention propensity prediction

	Logit	Lasso
Constant	-3.427***	-3.904
	(0.821)	
Share xen. act. 5 min bef. intervention	-7.769***	-1.505
	(1.933)	
Share xen. act. 100 min bef. intervention	–	-4.329
Share xen. act. 30 min bef. intervention	2.934	2.600
	(1.539)	
log(cum. act. at 100 min bef. int.)	-0.416	-0.218
	(0.264)	
log(cum. act. at 30 min bef. int.)	-0.375***	-0.382
	(0.078)	
log(act over 100-5 minutes bef. int.)	–	0.202
log(act over 30-5 minutes bef. int.)	-0.707***	-0.462
	(0.207)	
log(cum. act. at 5 min bef. int.)	–	–
log(cum. xen. act. at 5 min bef. int.)	1.612**	1.360
	(0.554)	
log(cum. xen. act. at 100 min bef. int.)	0.329	0.108
	(0.365)	
log(cum. xen. act. at 30 min bef. int.)	–	0.057
log(xen. act over 100-5 minutes bef. int.)	0.135	–
	(0.547)	
log(xen. act over 30-5 minutes bef. int.)	–	-0.147
Observations	1,081	1,081
$\log(\lambda)$		-7.15
AUC	0.686	0.686

Note: The dependent variable is a dummy variable indicating whether a post from the chat log was subject to an intervention by the counterspeech group. For the LASSO regression, the regularization parameter λ is chosen in five-fold cross validation to maximize the area under the receiver operating characteristic curve (AUC).
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table XXV—: Sensitivity analysis: descriptive statistics

	Baseline			Lasso			All treat.			Same day			2h window			1 closest			5 closest		
	Treat.	Contr.	Δ	Treat.	Contr.	Δ	Treat.	Contr.	Δ	Treat.	Contr.	Δ	Treat.	Contr.	Δ	Treat.	Contr.	Δ	Treat.	Contr.	Δ
Post-level comparison																					
Comments	99.2	85.8	-13.4*	103.8	88.6	-15.2*	112.8	96.7	-16.2*	90	74.5	-15.5	92.8	80.4	-12.4	98.9	88.1	-10.8	98.3	80.7	-17.6**
Reactions	399.2	395.9	-3.3	401.5	399.3	-2.2	441.7	475.2	33.5	299.4	284.8	-14.5	378.6	302.8	-75.8	424.9	424.8	0	397	368.2	-28.8
Comments & reactions	498.4	481.7	-16.6	505.5	488	-17.3	554.5	571.8	17.3	389.4	359.4	-30	471.4	383.2	-88.2	523.7	512.9	-10.8	495.3	448.9	-46.3
Users	253.9	255.6	1.7	262.5	258	-4.5	285.6	292.1	6.5	211.9	198.1	-13.8	246.8	205.9	-40.9	266.5	271.1	4.6	252.1	239.7	-12.4
Xen. comments (%)	28.1	27.6	-0.5	27.9	26.6	-1.3	20.6	21.9	1.3	24.8	23.5	-1.4	23.7	23.9	0.2	28.2	28.6	0.5	28	27.2	-0.7
Comments with tags only (%)	0.9	0.8	-0.1	0.9	0.8	0	0.8	0.9	0.1	0.9	0.7	-0.2	0.8	0.6	-0.2	0.7	0.7	0	0.9	0.9	0
Observations	178	370	548	176	379	555	312	312	624	96	170	266	53	76	129	170	170	340	176	494	670
User-level comparison																					
Avg. weekly activity	5.06	4.94	-0.12	4.91	4.77	-0.14	4.63	4.18	-0.45***	8.22	8.24	0.02	11.59	11.90	0.31	6.77	6.20	-0.57***	4.66	4.54	-0.12
Avg. weekly xen. activity	0.40	0.40	0.00	0.39	0.39	0.00	0.36	0.33	-0.03***	0.72	0.78	0.06**	1.07	1.22	0.15***	0.59	0.55	-0.04**	0.37	0.36	-0.01
Share of weeks w. activity	0.70	0.69	-0.01	0.70	0.69	-0.01	0.67	0.66	-0.01	0.77	0.77	0.00	0.79	0.79	-0.01	0.73	0.73	0.01	0.69	0.69	0.00
Share of weeks w. xen. activity	0.25	0.23	-0.02	0.25	0.23	-0.02	0.22	0.20	-0.02	0.32	0.32	0.00	0.36	0.38	0.01	0.29	0.30	0.01	0.23	0.22	-0.01
# of commented media outlets	4.30	4.23	-0.07**	4.26	4.21	-0.05	4.10	4.12	0.02	4.80	4.81	0.01	5.29	5.27	-0.03	4.60	4.52	-0.09**	4.18	4.15	-0.03
Observations	20,342	61,508	81,850	21,373	61,234	82,607	38,238	53,992	92,230	14,458	32,459	46,917	10,855	18,847	29,702	26,099	34,524	60,623	18,640	71,629	90,269

Note: Descriptive statistics for various alternative definitions of treatment and control group. The top panel of the table corresponds to Table IV, the bottom part to Table V. The first supercolumn reproduces the baseline results from the main body of the paper. In the second supercolumn, treatment propensities were computed using the LASSO. In the third supercolumn, all treatment articles were retained. The fourth supercolumn decreases this time window to two hours. The fifth supercolumn, all treatment posts that were considered on the same day. The fifth supercolumn decreases this time window to two hours, i.e. only the exact alternatives that the counterspeech group considered for intervention. In the sixth supercolumn treatment and control groups have been restricted to only retain the closest match in terms of predicted intervention probability (compared to the closest three in the baseline). The last supercolumn restricts them to the closest five matches.

Table XXVI—: Sensitivity analysis: regressions

Dependent var.	Coefficient	Baseline		Lasso		All treat.		Same day		2h window		1 closest		5 closest	
		Est	Se	Est	Se	Est	Se	Est	Se	Est	Se	Est	Se	Est	Se
User-level regressions															
$\mathbb{1}\{xen.\ activity\}$	Intervention	−0.052***	0.005	−0.052***	0.005	−0.028***	0.004	−0.044***	0.006	−0.025***	0.008	−0.031***	0.005	−0.056***	0.005
$\mathbb{1}\{xen.\ activity\}$	Intervention	−0.016*	0.007	−0.015*	0.006	−0.008	0.005	−0.016	0.009	−0.003	0.011	0.002	0.007	−0.018**	0.007
$\mathbb{1}\{xen.\ activity\}$	Int. × SLC	−0.070***	0.009	−0.071***	0.008	−0.040***	0.006	−0.050***	0.011	−0.037**	0.013	−0.060***	0.008	−0.073***	0.009
log(activity)	Intervention	−0.068***	0.008	−0.077***	0.008	−0.030***	0.006	−0.066***	0.010	−0.061***	0.013	−0.018*	0.008	−0.081***	0.008
Post-level regressions															
log(xen. activity)	Intervention	−0.136*	0.064	−0.074	0.067	−0.044	0.058	−0.103	0.087	−0.237	0.138	−0.089	0.078	−0.144*	0.062
Share xen.	Intervention	−0.030**	0.011	−0.033**	0.011	−0.021*	0.009	−0.043**	0.014	−0.043**	0.016	−0.032*	0.013	−0.036***	0.010
User share xen.	Intervention	−0.010***	0.002	−0.010***	0.002	−0.012***	0.002	−0.009**	0.003	−0.010*	0.005	−0.010***	0.002	−0.010***	0.002

Note: Key regression results for various alternative definitions of treatment and control group. The rows in the top panel correspond to the user-level regression reported in Table VI column 2, Table VII column 2 and Table VIII column 3. The bottom panel corresponds to the post-level regression results contained in Table XI. The first supercolumn reproduces the baseline results from the main body of the paper. In the second supercolumn, treatment propensities were computed using the LASSO. In the third supercolumn, all treatment articles were retained. The fourth supercolumns only allows treatment posts to be matched to potential control posts that were considered for intervention on the same day. The fifth supercolumn decreases this time window to two hours, i.e. only the exact alternatives that the counterspeech group considered for intervention. In the sixth supercolumn treatment and control groups have been restricted to only retain the closest match in terms of predicted intervention probability (compared to the closest three in the baseline). The last supercolumn restricts them to the closest five matches. The dependent variables in the bottom two rows are computed by excluding users who already participated in counterspeech interventions previously, as explained in the main body of the paper.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$