Andres, Raphaela; Slivko, Olga

Working Paper

Combating online hate speech: The impact of legislation on Twitter

ZEW Discussion Papers, No. 21-103

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Andres, Raphaela; Slivko, Olga (2021) : Combating online hate speech: The impact of legislation on Twitter, ZEW Discussion Papers, No. 21-103, ZEW - Leibniz-Zentrum für Europäische Wirtschaftsforschung, Mannheim

This Version is available at:
https://hdl.handle.net/10419/248857

// NO.21-103 | 12/2021

DISCUSSION PAPER

// RAPHAELA ANDRES UND OLGA SLIVKO

Combating Online Hate Speech: The Impact of Legislation on Twitter

Leibniz Association

ZEW

Combating Online Hate Speech: The Impact of Legislation on Twitter

Raphaela Andres* and Olga Slivko†

December 2021

Abstract

We analyze the impact of the Network Enforcement Act, the first regulation which aims at restraining hate speech on large social media platforms. Using a difference-in-differences framework, we measure the causal impact of the German law on the prevalence of hateful content on German Twitter. We find evidence of a significant and robust decrease in the intensity and volume of hate speech in tweets tackling sensitive migration-related topics. Importantly, tweets tackling other topics as well as the tweeting style of users are not affected by the regulation, which is in line with its aim. Our results highlight that legislation for combating harmful online content can influence the prevalence of hate speech even in the presence of platform governance mechanisms.

Keywords: Social Networks, User-Generated Content, Hate Speech, Policy Evaluation.

JEL Class: H41, J15, K42, L82, L86.

*Raphaela Andres: raphaela.andres@zew.de; Phone: +49 621 1235 – 198. ZEW Mannheim, Digital Economy Department, L7 1, 68161 Mannheim, Germany and i3, Telecom Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France.

†Olga Slivko: slivko@rsm.nl; Phone: +49 176 433 19 409. Department of Technology and Operations Management, Rotterdam School of Management, Erasmus University, Rotterdam, The Netherlands.

We are grateful for helpful comments from Irene Bertschek, Ulrich Laitenberger, Dominik Rehse, Felix Rusche, Anthony Strittmatter, Michael Zhang and conference participants of ITS 2021, EARIE 2021, the 2nd AI Policy Conference by RegHorizon and ETH Zürich, and ICIS 2021.

1 Introduction

Social media have become a primary information channel for many individuals (Pentina and Tarafdar 2014). In 2019, one-in-three people in the world used social media platforms.[1] The far-reaching spread of social media provides new opportunities for marketing and political involvement, but also for the dissemination of extremist thoughts and aggressive or harassing content. Since the 2015 refugee and migration crisis in 2015, the online dissemination of "hate speech" is an omnipresent topic in the public discourse in Germany. Additionally, at the beginning of Covid-19 lockdowns, scholars and media documented the emergence of clustered hateful communities in some countries, for example, in the US and the Philippines on Twitter and Reddit (Uyheng and Carley 2021). These developments are dangerous, since the use of mass media and social media for incitement to hatred can lead to *stochastic terrorism*, i.e. can incite attacks by random extremists.[2]

Hate speech spread online often implies aggressive and derogatory statements towards people belonging to certain groups based on e.g. their gender, religion, race, or political views (Geschke et al. 2019). To counteract the dissemination of online hate, some platforms have introduced community standards and house rules, allowing them to moderate the content distributed on the platforms. However, the incentives of profit-making platforms for content moderation may diverge from the socially desirable level of content moderation. In fact, it might be optimal for platforms to keep extreme content on the platform to extent their user base and, hence, profits from advertising (Liu, Yildirim, and Z. J. Zhang 2021). As a first legal framework to combat hateful speech, the German government implemented the Network Enforcement Act (NetzDG) in January 2018. This law obliges social platforms with more than two million users in Germany to implement mechanisms so that users can easily complain about hateful posts. Furthermore, NetzDG requires that reported posts containing clearly hateful and insulting content must be deleted within 24 hours after it has been reported.

We exploit the implementation of this regulation in a natural experimental framework to causally identify its effect on the user-generated content (UGC) by a target group of German Twitter users. Specifically, we investigate if the introduction of the law mitigates the prevalence and intensity of hateful speech among Twitter posts of German right-wing sympathizers. For measuring hateful speech, we use pre-trained algorithms provided by Jigsaw and Google's Counter Abuse Technology team. The Application Programming Interface (API) "Perspective" can identify dimensions such as toxicity and profanity in short texts. Since the application of NetzDG is restricted to the content on social networks that users on the German territory are exposed to, we apply a difference-in-differences framework comparing the evolution of the language used by comparable subgroups of users

[1] Our World in Data

[2] Wired Article ; Original Quote ; Recent example

in the German and Austrian Twittersphere.

Isolating the causal effect of the law on several measures of hateful speech, we find that the regulation reduces the intensity of hateful speech in Germany by about 2 percentage points, which corresponds to a reduction of 6%-10% of the standard deviation. The volume of original hateful tweets is reduced by 10%, resulting on average in i.e., one original insulting tweet less per user in two months. Although these effects seem modest, they measure the lower bound of the impact of NetzDG on hateful content on Twitter. First, we additionally find spillover effects to users located outside Germany. Second, because hateful tweets receive higher user engagement, e.g., a hateful tweet is on average retweeted four times in our sample, the reduction in one original hateful tweet reduces the exposure of the Twitter audience to hate by five hateful tweets.

Our results contribute to the current discussion whether legal regulation of online content can complement the platform's own guidelines such as the "Twitter hateful conduct policy". The causal effect of NetzDG on tweets tackling migration topics on Twitter suggests that while the platform guidelines apply to both German and Austrian users, hateful content in tweets posted in Germany decreased significantly due to the law. Our findings are especially relevant since other countries[3] and the European Commission[4] are currently working on the design of similar regulations. A better understanding of the effects of such policies is crucial for establishing guidelines for the successful regulation of UGC.

2 Literature

Regulation of UGC on social media has recently become a salient issue in the light of current research findings that draw a direct link between xenophobic attitudes on social media and real-world hate crimes (Müller and Schwarz 2021, Müller and Schwarz 2020, Bursztyn et al. 2019, Olteanu et al. 2018). Aral and Eckles 2019 highlight that due to misinformation campaigns targeting millions of U.S. citizens during the 2016 presidential election campaign, it is important to design platforms and policies so that they cannot be manipulated by social media. Hence, platforms as well as policymakers strongly debate the effectiveness and possible side effects of content moderation. Our paper contributes to this debate by presenting the first empirical evidence on the effects of content moderation imposed by government regulators.

Research on the voluntary content moderation of social media platforms has investigated several contexts, including pro-eating disorder tags on Instagram (Chancellor et al. 2016) and streaming content on Twitch (Seering, Kraut, and Dabbish 2017). Focusing on specific

[3]https://www.rcmediafreedom.eu/Publications/Reports/Countering-online-hate-speech-against-migrants-and-refugees

[4]https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia

communities and content features, such as tags or emoticons in posts, these studies find limited effects of content moderation on the features of the subsequently generated content. More closely related to our context of hateful posts, research on the moderation of hateful discourse suggests that moderation can sometimes be effective in restraining harmful UGC. Chandrasekharan et al. 2017 exploit the ban of two hateful communities on Reddit as a quasi-experiment to analyze if the deletion of hateful subreddits leads to the reduction or reallocation of hate speech. They analyze the usage of hate speech according to their manually constructed lexicon and find the ban to be effective in reducing hate speech. Although many users who had previously posted hate speech left the platform, the ones who stayed on Reddit reduced their usage of hate terms according to their lexicon. Srinivasan et al. 2019 analyze the evolution of swear words and hate terms within a subreddit and exploit the time gaps that moderators need to remove the non-compliant content. The authors find no significant effect of content deletion on the use of swear words and hate terms by the previously non-compliant users. Our paper contributes to this literature by tracking the evolution of hateful speech by more advanced measures than lexika. The analysis of hateful terms cannot capture all aspects of hateful speech, e.g. threats and irony, and is sensitive to (unintended and intended) spelling errors. Additionally, Perspective API provides us with continous scores about several dimensions of hateful speech. These scores also allow us to conduct analyses on the intensive margin, measuring the change in hate intensity.

Theoretical studies on the content moderation of digital platforms suggest that platform incentives to provide the optimal level of content regulation may be weak (Buiten, Streel, and Peitz 2020, Liu, Yildirim, and Z. J. Zhang 2021). In fact, it might be optimal for platforms to keep extreme content on the platform to extend their user base and advertising-driven profits(Liu, Yildirim, and Z. J. Zhang 2021). We contribute to these studies by presenting the first empirical evidence that policy-enforced content moderation, which imposes the risk of high fines in case of non-compliance can significantly reduce the intensity of hate speech among a target group of social media users (right-wing sympathizers) on sensitive topics such as migration and religion.

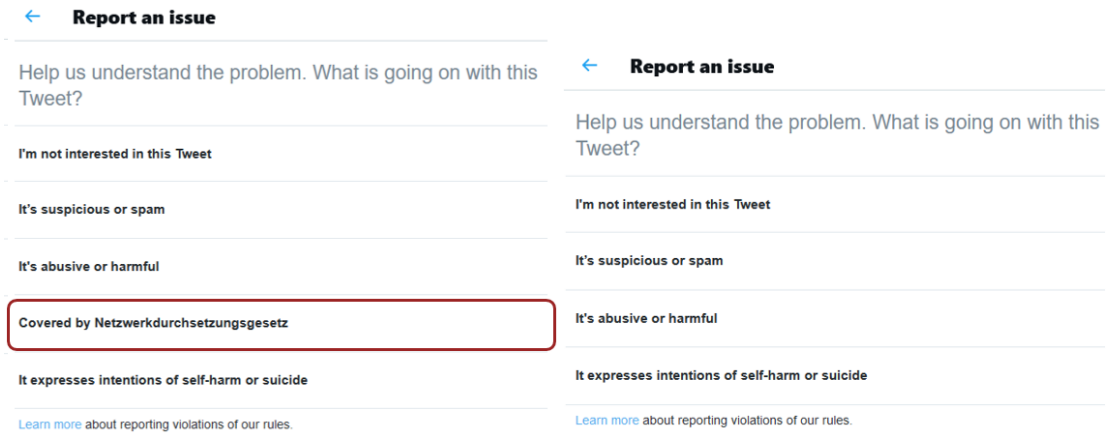
More broadly, we contribute to the literature discussing the social and economic impacts of the exposure to various media sources, including social media (see DellaVigna and La Ferrara 2015 for an overview). The studies focus on the effects of Internet and social media on consumer demand (Chevalier and Mayzlin 2006, Xu and X. Zhang 2013, Hinnosaar et al. 2021) and political participation (Enikolopov, Makarin, and Petrova 2020) and draw a connection between the content on social media and hate crimes (Müller and Schwarz 2021, Müller and Schwarz 2020, Bursztyn et al. 2019). Our findings suggest that the harmful effects of online hate can be weakened if platforms are strongly incentivized to to timely address user-reported hateful content.

3 Theoretical Background

3.1 Overview of the Network Enforcement Act

The NetzDG[5] (NetzDG) was passed by the German Bundestag in October 2017 and came into effect in January 2018. The law aims at increasing legal pressure on platforms to act against hateful content generated by users. Specifically, it obliges social media platforms with more than two million. registered users in Germany[6] to implement mechanisms that provide each user with a transparent and permanently available procedure to report illegal content on the respective platform. After receiving a complaint, the platform is required to review the complaint immediately and act within a reasonable time frame. If the user complaint targets unquestionably illegal content, it must be removed within 24 hours. In more nuanced cases, platforms have seven days to decide whether measures must be taken against the respective content or profile who submitted it. In practice, Twitter decided to add the option "Covered by Netzwerkdurchsetzungsgesetz" if the user accessed Twitter via a German IP address.[7]

Figure 1: Menu Options for Reporting Tweets with a German/Austrian IP Address



(a) German IP address (b) Austrian IP address

Next, the reporting users need to choose the paragraph of the criminal code violated by the post. Finally, they must sign an acknowledgement that the wrongful reporting of a tweet itself is a violation of the Twitter house rules.

Measures of the platforms range from deleting the content, sending warnings to the accounts that submitted the content to permanently blocking the account. Importantly, the law does not require platforms to proactively search and delete hate speech, but only

[5] Netzwerkdurchsetzungsgesetz, for English version of the law see

[6] As of December 2020, this applies to: Facebook, Youtube, Instagram, Twitter, Reddit, TikTok, Change.org, Jodel (BMJV 2020)

[7] If users located in Germany click on the broader option "It's abusive or harmful", he or she can indicate "Covered by Netzwerkdursetzungsgesetz", while other options are to report the usage of private information and incitement of suicide or self-harm.

Figure 2: Second Step of Reporting a Post on Twitter

to become active after receiving a concrete complaint that indicated the "Covered by Netzwerkdurchsetzungsgesetz"-option. Further NetzDG requirements include the appointment of a domestic authorized contact person for each platform and the semi-annual publication of the compliance report. This report must contain information for users how they can report illegal content and the number of deleted posts by the groups of users (ordinary users or reporting offices), reaction time, and the reason for reporting.

According to §4 of NetzDG, non-compliance can be penalized with a fine of up to five million €. However, due to the risk of content overblocking, the examples for punishable offences only include technicalities about the report and a systematic incorrect execution or monitoring of the complaint management system. To prevent platforms from pursuing the "better to be safe than sorry" strategy and delete any content that might seem questionable at first sight, non-compliance is not determined on individual case decisions.[8]

3.2 Theoretical Mechanisms of the Effect of Regulation

The introduction of the regulation combating hate speech implies that platforms should timely delete hateful content reported by users or bear the financial responsibility in the case of non-compliance. In this subsection, we derive theoretically how the introduction of NetzDG can affect content and user engagement on social media platforms.

User utility on a social media platform is composed of (i) the utility from consuming content and (ii) the utility from the post content. For the purpose of our analysis, we distinguish user-generated content on the platform with respect to the content extremeness

[8]Under the NetzDG, Facebook was fined five million € for an erroneous compliance report. This is the only legally effective fine under the NetzDG as of September 2021. Heise article

varying within the unit interval. If the degree of extremeness of the content on the platform is higher than the user's preference for the extremeness of content, the user's utility from consuming such content decreases. Hence, consuming content of high extremeness increases user's utility with preferences for equal or higher content extremeness , but decreases the utility of users with preferences for lower content extremeness. Moreover, users post content according to their own level of extremeness. Deleting this content imposes a psychological cost and decreases their utility from engaging with the platform. Users will only be active content consumers and producers on the platform if the utility of being active is non-negative.

A platform's advertising revenues increase with the number of users viewing and posting original content. In the absence of regulation, a platform can moderate user-generated content by imposing a threshold of content extremeness, above which content is deleted. While this can decrease the utility of users with stronger preferences for extreme content (above this threshold), the users with preferences for less extreme content derive higher utility from consuming content if the platform engages in content moderation. Hence, the platform's ex ante incentives to moderate content is determined by the tradeoff between the increase in the number of users who prefer content with extremeness below the moderation threshold and the decrease in the number of users with a preference for content extremeness above the threshold. If the threshold for extremeness chosen by the platform is too high, the regulator can impose a lower threshold and reinforce it with a fine that the platform has to pay in the case of non-compliance. In our empirical setting, we focus on the effect of NetzDG, which facilitates the reporting of hateful content on social media platforms and requires the complying platforms its timely deletion or imposes a large fine in case of non-compliance. Hence, as a result of imposed content moderation, we can expect that (i) platforms will moderate content and will allow only content below the threshold to remain on the platform; (ii) users who post high hate intensity content will decrease their content production or the level of hate intensity in the posted content; (iii) users with a preference for high hate intensity will migrate to competing platforms which are not subject to the regulation, for example, platforms that have a low adoption rate in Germany.

Transferring these theoretical considerations into the context of our empirical analysis, we expect that the level of hatred in the content on Twitter will decrease in Germany after the introduction of NetzDG if the threshold for the content to be considered illegal is lower than the platforms' own incentives. The mechanisms driving the decrease in the extremeness of content are then as follows. First, NetzDG facilitates the reporting of illegal content by platform users and requires the platforms to address these reports. Second, users might self-censor and decrease the intensity of hate in their posts, fearing that the platform will remove their tweets or block their profiles. Finally, the users can decide to be active on other platforms which are not subject to the regulation and, hence, engage in no or weaker content moderation.

However, it is also possible that the prevalence of hate speech on Twitter does not change substantially after the introduction of NetzDG. This is because Twitter must only react to posts that were reported by users. Yet, if a hateful post is tweeted inside a filter bubble of like-minded users, these users are unlikely to report the post. Moreover, although NetzDG requires platforms to provide a transparent procedure for reporting illegal content, the implementation of this procedure on Twitter is lengthy. It requires the user to be informed about the title and the specific paragraphs of the regulation and to be able to distinguish them from alternative reporting options. This multi-step procedure might discourage users from completing their reports and lead to less content being reported and, hence, deleted by the platform.

Our empirical setting does not allow us to distinguish between the potential mechanisms driving the effect of the regulation (discussed in Section 8). Rather, we focus on the broader question of the overall impact of NetzDG. A qualitative evaluation report, published by the German government in September 2020, suggests a positive regulation evaluation (BMJV 2020). Yet, the report mentions that the evaluation of the quality of platforms' decisions to remove or keep content is not possible due to the lack of data. To the best of our knowledge, our paper is the first to address this gap by quantitatively measuring the effect of the regulation on the level of hateful speech in tweets focusing on the target group of the law.

4 Data

4.1 Data Extraction

We focus our study on the online microblogging platform Twitter for a variety of reasons, although a study of the US suggests that Twitter users in the US may not be representative for the US population (Wojcik and Hughes 2019). In Germany, Twitter is one of the major social platforms with a user base of about 13% of Germans (Hölig and Hasebrink 2020). Moreover, this platform has recently become a major tool for political communication worldwide, as it provides politicians with a direct channel of communication with their electorate (Nulty et al. 2016, Petrova, Sen, and Yildirim 2021). In the last decade, the amount of hate speech on Twitter has increased steadily, just like on other social media platforms. However, as Twitter posts are public and visible for every internet user – even those without a Twitter profile – Twitter is an especially far-reaching platform.

For our empirical analysis, we selected all national and regional party profiles of the right-wing populist parties of two neighboring and German speaking countries Germany and Austria - the German Alternative für Deutschland (AfD) and the Austrian Freiheitliche Partei Österreichs (FPÖ). Of those 201 AfD and 30 FPÖ profiles, we downloaded all unique 200,000 followers on Twitter in May 2020 and drew a random subset of 5,500 followers for

each of the two parties. By selecting the tweets by followers of the two right-wing populist parties, we analyze the development of hate speech among two comparable subgroups in the German and Austrian Twittersphere. Moreover, hate speech is a prevalent issue in the right-wing populist community, since it often uses anti-elite rhetoric and declares immigration skepticism (Halikiopoulou 2018). In Germany, content posted by AfD politicans and their respective comments were found to be most hateful along the political spectrum. The xenophobic content generated by right-wing users in Germany was directly linked to the incidences of hate crime (Müller and Schwarz 2021). Hence, the impact of the law among this subgroup is especially of high interest. Furthermore, we chose to track the tweets of all followers of those parties and not necessarily the politicians belonging to these parties since we expect politicians to use more careful language. To avoid preselecting the sample based on assumptions, we kept the randomly drawn 23 politicians in our sample and added an indicator for being a politician (defined as members of the German or Austrian parliaments) to our data. Yet, the selection of specific followers - all followers of AfD and FPÖ profiles - implies that our results are not representative of the entire German Twittersphere, but only for this specific target group of the law. Since our estimation strategy hinges on the assumption that the users are either treated if they are located in Germany or not if they are located outside of Germany, we manually assigned the randomly drawn users that also tweeted original posts during our sample period to the country based on their profile information. This necessary step reduced our sample of Twitter users to 1,337 but increased the precision of our estimations.

Having a sample of Twitter users that we could cleanly locate to the treated (Germany) or control group (Austria or any other country than Germany), we downloaded all of their original posts between July, 2016 to June, 2019 (1.5 years before and 1.5 years after the introduction of NetzDG). Throughout the analysis, we only consider original posts and no simple retweets to track the language of individual users in our sample. In the next step, we filtered the 2.3 million retrieved tweets for migration and religion related buzzwords[9] in messages and hashtags. According to the "Political Speech Project" [10], German tweets related to anti-immigrant and anti-muslim topics are likely to be unlawful. Therefore, we further narrowed down our analysis on UGC that is likely to violate NetzDG.

We use this resulting sample of 160,000 tweets about sensitive topics as our sample for the baseline analysis. Importantly, our sample composition implies that the the results are not representative of the entire Twittersphere, but rather for an important target group of the law. Due to the problems associated with hate speech posted by right-wing populists (Caiani and Parenti 2013), the effect of the law on this user segment on Twitter is of especially high interest.

[9]For the filtering, we used the following word stems: reli, migra, islam, terror, flucht, flücht, moslem, koran, ausländ, ausland

[10]`https://rania.shinyapps.io/PoliticalSpeechProject/`

4.2 Outcome Variables

Our dependent variables measure the intensity of hate speech in tweets. Since the regulation does not provide any measurable definition of hate speech, we conducted our analysis on several dimensions of hateful speech to account for differentials in the understanding of hateful speech and learn more about potential channels through which the law might tackle the issue of hate speech. These measures are constructed by the Application Programming Interface (API) "Perspective", which is provided by Jigsaw and Google and allows the employment of pre-trained machine learning models to score the probability that short texts are hateful. The models include dimensions of hatefulness such as, severely toxic, toxic, threatening, an identity attack, profane, and insulting language. These scores range from 0 to 1 and can be interpreted as intensities of hate in tweets. Perspective API is used by well-known entities such as The Financial Times and the discussion website Reddit (Perspective API n.d.). In the natural language processing literature, it is a benchmark prediction algorithm (Fortuna, Soler, and Wanner 2020).[11] The predictions by Perspective API rely on a convolutional neural network trained on large corpora of publisher and user-generated content from multiple domains (such as Wikipedia, the New York Times, The Economist, The Guardian, including user comments on their forums). Each comment within these corpora was labeled by at least ten human classifiers (Fortuna, Soler, and Wanner 2020). Table 1 presents the definitions of the various dimensions of hate speech used in our analysis and their scores for one exemplary tweet from our sample (Table 10 in the Appendix gives further examples of tweets with high scores of hate). In our analysis, we include all sentiment dimensions related to hate speech and available for the German language.[12]

4.3 Summary Statistics

Following the extraction procedure described in Section 4.1, we obtained 735 right-wing sympathizers located in Austria and 602 users in Germany. Several users (187) indicated that they did not live in Germany or Austria in their profile information. Since we are not interested in the user's residency per se but only if they live in German territory and are therefore exposed to the regulation. We therefore assign those users to the control group together with the users from Austria. We kept an indicator for those profiles to account for potential differences in tweets between those living in Austria and those living somewhere else. Table 2 presents measures describing the profiles of users in our sample. Most of the user characteristics in Table 2 are quite dispersed. For example, the number of followers ranges from 0 to almost 550,000. The oldest profile in our sample was created in

[11] Comparing tweets by two German-speaking countries allows us to apply the same algorithms to the treated and control group. Therefore, potential prediction biases distributed randomly across tweets in our sample do not affect our results due to our identification strategy.

[12] `https://support.perspectiveapi.com/s/about-the-api-attributes-and-languages`

Table 1: Outcome Variables for an Exemplary Tweet as Computed by Perspective API

Example tweet, translated to English:
"We have pulled the teeth out of pagan + witch-killing Christianity... Islam is waiting"

Outcome	**Score**	**Definition** [a]
Severe Toxicity	0.5809524	A very hateful, aggressive, disrespectful comment or otherwise very likely to make users leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.
Toxicity	0.8143812	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.
Threat	0.6558015	Describes an intention to inflict pain, injury, or violence against an individual or group.
Identity Attack	0.9192697	Negative or hateful comments targeting someone because of their identity.
Profanity	0.3270008	Swear words, curse words, or other obscene or profane language.
Insult	0.6519685	Insulting, inflammatory, or negative comment towards a person or a group of people.

Note. This table shows the estimated hate intensity scores with regard to all hate dimensions used in this analysis. The last column includes the definitions of the dimensions as defined by Perspective.

[a] `https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages`

2007, whereas other users created their accounts after the introduction of NetzDG. Some users (18%) only tweeted once. This might be due to low account age, inactivity during our sample period, or little interest in migration or religion, since we only include tweets about these sensitive topics in our main sample. Among our randomly chosen accounts, 22 (1.6%) are the user accounts of politicians (i.e., members of the German or Austrian parliament), and 28 accounts belong to a well-known personality ("verified").

Table 3 presents summary statistics on the tweet level. Besides the tweet's text, we extracted additional meta information such as the number of retweets, likes, and replies. The popularity of tweets can differ greatly. Most are not retweeted or liked, whereas others have more than 1,700 likes. The median tweet length in our sample comprises 15 words, while the number of possible characters of a tweet doubled from 140 to 280 characters within our sample period. As Twitter imposed this rule for both countries simultaneously and we include month fixed effects in every estimation, the increase of allowed characters does not threaten the validity of our identification strategy. Further information we collected on the tweets are indicators if the tweet includes a video, photo, URL, or a link to a media outlet. We also observe the time when the tweet was posted and construct an indicator if a tweet includes a link shortener as often used by automated bots. Finally, we added a country-specific daily indicator if a terrorist attack or an election (European, national or regional elections) took place in Germany or Austria. Within our

Table 2: Summary Table of User Characteristics

	N	Mean	Median	SD	Min	Max
no. Followers	1334	2008.9	235.5	16337.6	0	534819
no. Friends	1334	1601.4	487	16668.3	1	590754
year of account creation	1335	2014.1	2014	3.00	2007	2019
Verified account	1335	0.021	0	0.14	0	1
live in GER	1337	0.45	0	0.50	0	1
live outside GER/AUT	1337	0.14	0	0.34	0	1
only 1 tweet in sample	1337	0.18	0	0.39	0	1
no. tweets in sample	1337	120.0	9	524.0	1	12279
no. sens. tweets user/month	1337	10.1	2.22	33.8	1	848.1
Politician	1337	0.016	0	0.13	0	1

Notes. The table shows summary statistics on the user level. All statistics combined show that the users in our sample are diverse with regard to Twitter activity and connectedness.

sample period, national elections in Germany as well as in Austria took place in the fall of 2017.

In Table 4 we compare the average of all outcome dimensions between the treated and control group. The overall intensity of hateful content is higher in our sample of German compared to Austrian users. However, the descriptive evidence suggests that in Germany, the mean values decreased after NetzDG became effective, whereas they increased in Austria. Table 28 (see Appendix) shows the pairwise correlations among the outcome variables, indicating high correlations between toxicity and insults and between severe toxicity and toxicity.

5 Empirical Strategy

Since the application of NetzDG is restricted to the content on social networks on German territory, we apply a difference-in-differences (DID) framework comparing the evolution of the language used by comparable subgroups of users of the German and Austrian Twittersphere. We estimate the DID by ordinary least squares (OLS) and include fixed effects for users, calendar months, and account age at the time the tweet was posted:

$$Hate\ Intensity_{ijt} = \beta_0 + \beta_1\ AfterT_t\ Treated_{ij} + \beta_2\ X_{it}^{'} + \mu_j + \nu_t + k_{t'} + \varepsilon_{ijt}$$

We estimate seperate regression models for each of the hate speech outcomes, such that the left-hand side of the equation $Hate\ Intensity_{ijt}$ corresponds to the respective hate intensity of a tweet i issued by user j on day t concerning severe toxicity, identity attacks, etc. provided by Perspective API. $X_{it}^{'}$ is a vector of the time variant control variables indicating the day of the week the tweet was posted and if the tweet was posted at

Table 3: Summary Table of Tweet Characteristics

	N	Mean	Median	SD	Min	Max
Toxicity	160474	0.43	0.45	0.22	0	1
Severe Toxicity	160474	0.30	0.29	0.24	0	1
Threat	160474	0.35	0.21	0.25	0	1
Identity Attack	160474	0.57	0.61	0.28	0	1
Profanity	160474	0.20	0.11	0.20	0	1
Insult	160474	0.37	0.36	0.22	0	1
no. Retweets	160474	4.00	0.00	19.35	0	911
no. Likes	160474	7.03	0.00	36.12	0	1711
no. Replies	160474	1.12	0.00	5.62	0	292
Video in tweet	160474	0.00	0.00	0.05	0	1
Photo	160474	0.07	0.00	0.26	0	1
URL	160474	0.68	1.00	0.46	0	1
Link to media outlet	160474	0.06	0.00	0.24	0	1
no. Words	160474	18.19	15.00	9.60	1	57
Tweeted at night	160474	0.08	0.00	0.26	0	1
Terrorist attack in country	160474	0.02	0.00	0.12	0	1
Election in country	160474	0.01	0.00	0.10	0	1

Notes. The table shows summary statistics on the tweet level. The first six rows are the outcome variables of the main analysis. Subsequently listed are tweet characteristics such as the number of retweets and number of words. FInally, we included country specific indicators for days on which an election and/or terroristic attack took place.

Table 4: Outcome Variables by Country and before/after NetzDG

	Austria before	Austria after	Germany before	Germany after
	Mean	Mean	Mean	Mean
Toxicity	0.40	0.43	0.45	0.42
Severe Toxicity	0.28	0.28	0.33	0.29
Threat	0.33	0.32	0.38	0.35
Identity Attack	0.53	0.59	0.60	0.57
Profanity	0.19	0.22	0.21	0.20
Insult	0.35	0.39	0.38	0.37
Observations	33016	30322	47855	49281

Notes. The table shows the average of all hate dimension scores by country and before/after NetzDG became effective.

night. We also added country-specific daily indicators for terrorist attacks, and national or regional elections, as these events could affect the usage of hate speech in a country which would not be captured by country fixed effects. In both of the countries, national elections took place in the fall of 2017. The coefficient of $AfterT_t\ Treated_{ij}$, β_1, is the coefficient of interest, which measures the change in the hate intensity in a tweet in Germany after NetzDG. μ_j represent user fixed effects (FE) to control for user-specific tweeting style and ν_t account for calendar month FE to capture general time trends. We additionally include account age FE $k_{t'}$, as the literature suggests that cohorts of social network users may differ in their writing style (Ershov and Mitchell 2020).

6 Results

6.1 The Effect of NetzDG on Hate Intensity

Table 5 reports the results of our baseline specifications.[13] The results suggest that the probability of tweets to include potentially illegal content significantly decreases after the introduction of NetzDG in Germany. As our dependent variable is measured on a scale between 0 and 1, the coefficients of interest are interpreted as percentage point (pp) changes in the dependent variables. For example, Col. (1) shows that NetzDG significantly reduces the intensity of severe toxicity, toxicity, and insulting remarks by 2 pp and profanity by 1 pp. Noteworthy, the introduction of NetzDG has the highest effect on the tweets related to identity attacks: the probability decreased by 3 pp.

The comparison of the effect sizes to the means of the outcome variables hints at modest effect sizes. For example, the average intensity of an identity attack in all tweets in our sample is 0.57, at the mean, this would decline by 3 pp and result in an average intensity of 0.54. In percentage terms, these numbers indicate a reduction in the hate intensity measured by the intensity of identity attack of 5% of the mean value or 9% of the standard deviation. Similarly, for severe toxicity the decline is 2 pp, which implies a reduction in hate intensity by 6% of the mean or 8% of the standard deviation. The semielasticities of the changes in hate intensity are approximately 1% - 3% for most of the dependent variables, except for threat intensity which is insignificant throughout our analysis.[14] This can be explained by the fact that threats have already been actionable and illegal before the NetzDG. These results remain strong and robust to sample composition tests. Using a balanced sample consisting of accounts that tweeted before and after NetzDG and excluding users living outside of Germany and Austria does not alter the results (see Table 18 and Table 19 in the Appendix). Furthermore, omitting the transition period (i.e., six months before the introduction of NetzDG which ellapsed between the moment when the law was

[13] Table 15 in Appendix presents the full list of control variables with the respective coefficients.
[14] See Table 16 for the regression outputs.

approved by the Bundestag and actually came into force, and which also includes the national elections in both countries) also does not change the results (Table 20). Moreover, our preferred specification only controls for the weekday and indicators for night time, terrorist attacks and elections, since we consider the tweet characteristics shown in the summary statistics (Table 3) rather as potential outcomes of the treatment effect. However, including these tweet characteristics as a robustness check does also not alter our results (see Table 17) and further confirms the robustness of our main specification. Lastly, our results are robust to the placebo treatment. If we set NetzDG to January 2017, the year before the actual implementation of the law, the treatment effect vanishes (see Table 21).

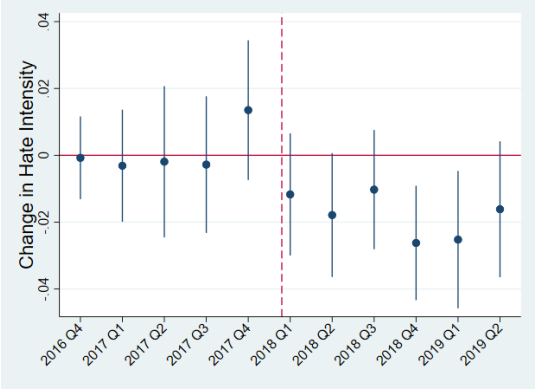
Table 5: Baseline Analysis: The Effect of NetzDG on the Intensity of Hate in Tweets

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID Attack	(5) Profanity	(6) Insult
Treated after T.	-0.02***	-0.02***	-0.01	-0.03***	-0.01***	-0.02***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.16	0.18	0.07	0.20	0.12	0.18
Observations	160161	160161	160161	160161	160161	160161
Mean of Outcome	0.30	0.43	0.35	0.57	0.20	0.37
SD of Outcome	0.24	0.22	0.25	0.28	0.20	0.22

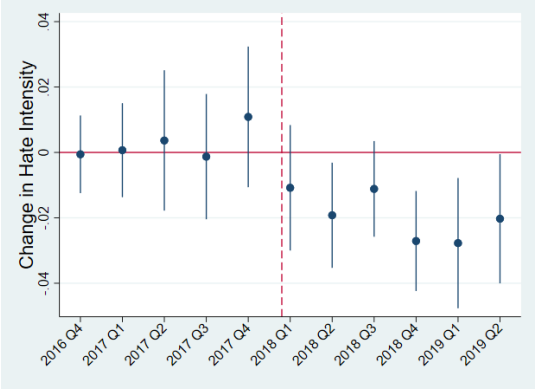
Notes. The table shows the main coefficients of the difference-in-differences estimations comparing the hate intensity in tweets by users affected and unaffected by the law (NetzDG). The columns contain the outcome measures discussed in the data section: Continuous scores ranging from 0 to 1 with regard to severe toxicity, toxicity etc. as calculated by Perspective API. The coefficient *Treated after T.* shows the change in hate intensity in terms of percentage points for users located in Germany after NetzDG became effective. Besides the treatment effect, all estimations control for country-specific events of regional/national elections and terroristic attacks, the day of the week the tweet was sent and an indicator if the tweet was sent at night. All estimations include a constant and user fixed effects, year-month fixed effects, and fixed effects for the account age in months at which the respective tweet was posted. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

Our results are based on the assumption that the measures related to hateful speech followed comparable trends in the treated and control group before NetzDG was introduced in Germany. We test the parallel trend assumption by decomposing the treatment effect by quarters before and after the regulation was introduced. Figure 3 presents the results for our six dependent variables of interest, corresponding to Col. (1)-(6) in Table 5. The standard errors of coefficients plotted in the figures correspond to the 90% significance level. The graph shows that the treatment and control group do not systematically differ *ex ante* the law was implemented in Germany, but they do differ in the quarters subsequent to the treatment (except for panel (c) (threat), for which we do not find any effect of the regulation).

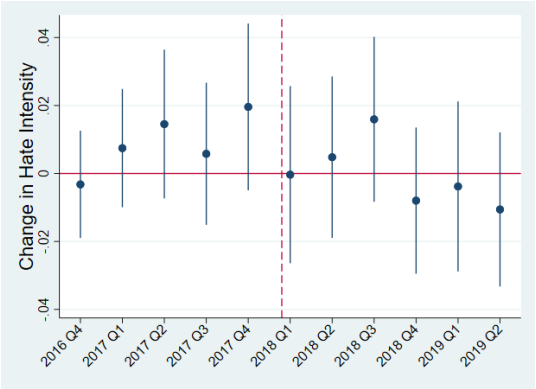
Figure 3: Quarterly Treatment Effects with Pre-trends



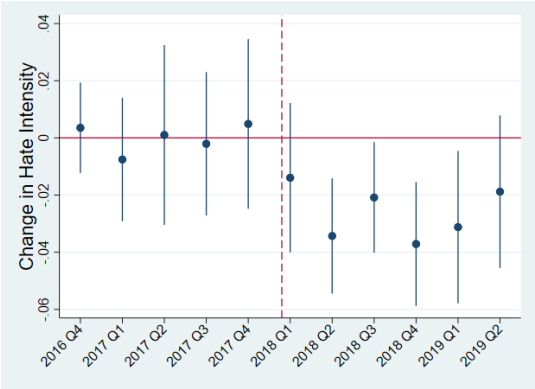
(a) Severe Toxicity



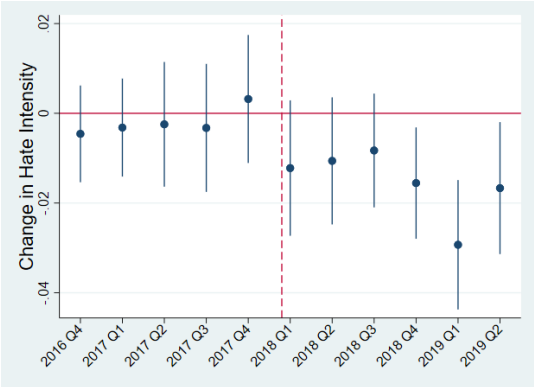
(b) Toxicity



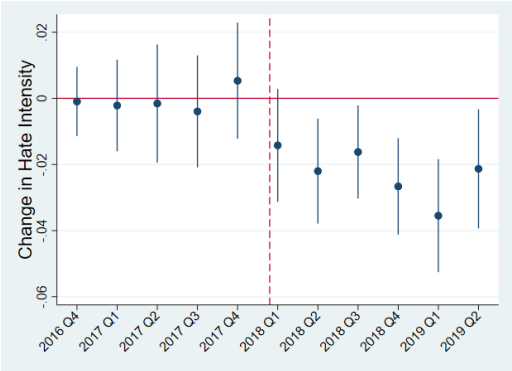
(c) Threat



(d) Identity Attack



(e) Profanity

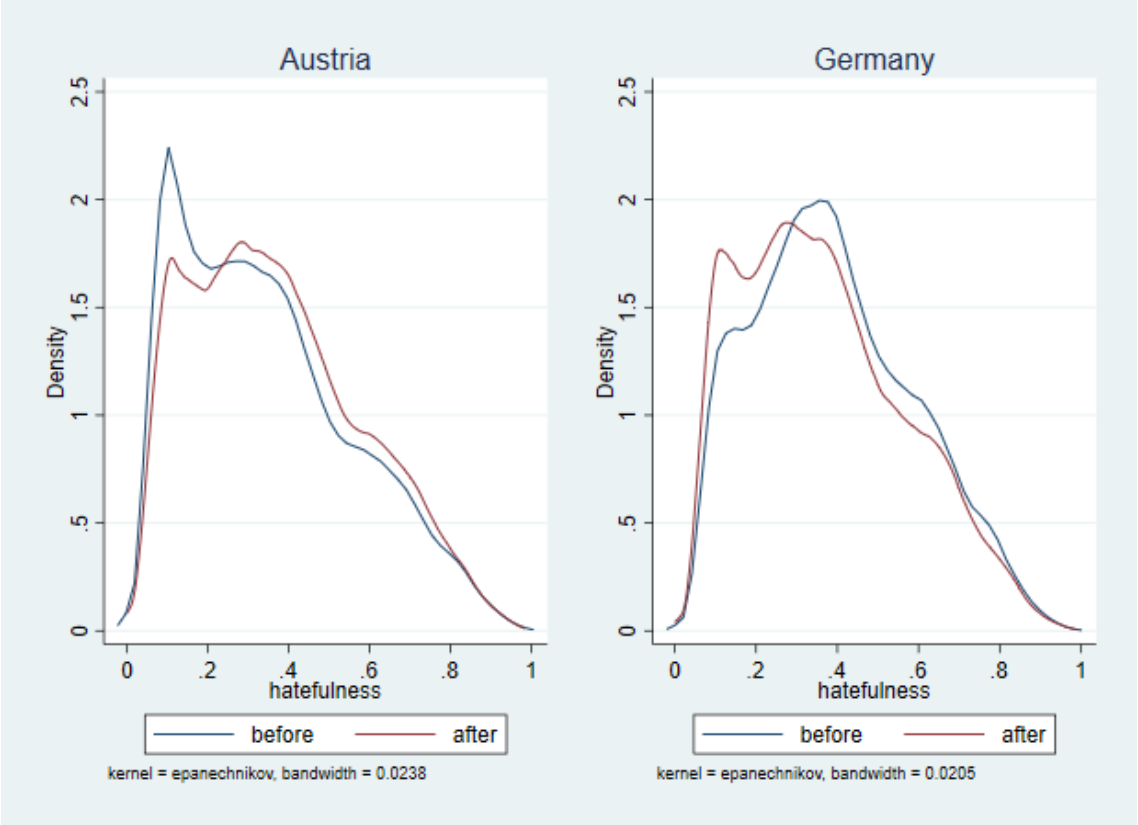


(f) Insult

Note. The plot shows the treatment effects for Q4 2016 - Q2 2019 for the six hate dimensions. Shown is the coefficient of the interaction of a treated tweet (posted by a user located in Germany) with different timings for NetzDG and the 90% confidence interval, while controlling for country specific events of regional/national elections and terrorist attacks, the day of the week the tweet was sent and an indicator if the tweet was sent at night. All estimations include a constant and user fixed effects, year-month fixed effects, and fixed effects for the account age in months when the respective tweet was posted. Standard errors are clustered at the user level. The vertical line indicates the date NetzDG became effective.

Comparing the distribution of the average value of hate intensity measures before and after the introduction of NetzDG for the treated (Germany) and untreated (Austria) group suggests interesting patterns. The graphs in Figure 4 display the average hatefulness score across tweets, calculated for each tweet as the mean value of our six hate dimensions. Figure 4 shows that the distribution of the hatefulness score shifts towards higher scores in Austria after January 2018, the shift in German scores is trending towards lower scores. The change in the distribution of these scores mainly happens in the middle part of the distribution, whereas the density at the tails does not change much. This indicates that the treatment effect is mainly driven by a reduction in the average hate intensities in the main body of the distribution and not by a reduction in the number of very hateful tweets.

Figure 4: Distribution of Average Hate Intensity by Time Period and Treatment Status



Note. The above plots show the distribution of the hate intensity score of each tweet by untreated (Austria) and treated (Germany) users before and after NetzDG became effective. Observations range from 1.5 years before to 1.5 years after NetzDG. We used the average of the scores of all six dimensions to calculate the hate intensity score.

6.2 The Effect of NetzDG on the Volume of Hateful Tweets

In addition to the hate intensity in tweets, we address the effect of NetzDG on the volume of original hateful content posted by the followers of right-wing parties located in Germany.

We set up a panel at the user-month level and aggregate the number of tweets containing hateful speech according to Perspective API. Since the outcome variables are measured in intensities of the six hate dimensions, we constructed an indicator for each tweet and defined a tweet as belonging to the category, for example, "severely toxic" if the probability of being severely toxic is above 0.5 - i.e., it is more likely than unlikely that the tweet is severely toxic. Due to the nature of our data, the data is very unbalanced as very few users tweet frequently about migration and/or religion. Therefore, the following estimations only include users who tweeted at least twice before and after the introduction of NetzDG to properly account for user fixed effects.

Our fixed effects estimations in Table 6 yield a similar picture to the tweet-level estimations. The coefficients with respect to all of the measures related to hateful speech are negative, yet we lose precision in the estimations (Table 22 in the Appendix presents the list of control variables). Hence, the volume of potentially unlawful tweets also declined in Germany as a consequence of NetzDG. Since we estimate the impact of the law on the logarithmic outcomes, the coefficients are interpreted as semielasticities of the change in the number of potentially unlawful tweets. According to Table 6, the number of severely toxic tweets fell by 8% in Germany due to the introduction of NetzDG. Comparing this effect to the average number of severely toxic tweets by user and month (3) implies that on average, there is one severely toxic tweet less per user in four months in Germany. The highest effect is found for identity attacks. On average, each user in the sample posts nine identity attacks per month. If this is reduced by 11%, there is one identity attack less per user per month in Germany due to the law.

Table 6: Panel: Volume of Outcome Variables by UserMonth in Logs

	(1)	(2)	(3)	(4)	(5)	(6)
	sev. Toxicity	Toxicity	Threat	ID Attack	Profanity	Insult
Treated after T.	-0.08*	-0.11**	-0.08	-0.11*	-0.08*	-0.10**
	(0.05)	(0.05)	(0.05)	(0.06)	(0.04)	(0.05)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.03	0.03	0.04	0.03	0.02	0.02
Observations	9546	9546	9546	9546	9546	9546
Groups	492	492	492	492	492	492
Mean of Outcome	3.08	4.64	3.41	9.45	1.66	3.93
SD of Outcome	8.51	11.25	9.92	23.16	4.29	9.08

Notes. The table shows the main coefficients of the panel difference-in-difference estimations on the user-month level. For each user and each month, the number hateful tweets is contructed by the number of tweets with hate intensity >0.5 with regard to the outcome variable indicated at the top of the column. The sample is restricted to users who posted at least twice before and after NetzDG. Besides the treatment effect, all estimations control for the country specific share of tweets posted during night times and on days of regional/national elections and terroristic attacks. All estimations include a constant and user and year-month fixed effects. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

7 Mechanisms and Implications of NetzDG

7.1 The Effect of NetzDG on All Tweets by AfD and FPÖ Followers

We now conduct the same analysis as in Table 5, but use all the tweets posted by the users in our sample. The results show that NetzDG specifically affects sensitive UGC (see Table 7). While we find no effect of NetzDG on our hate speech measures in the unrestricted sample of tweets, the treatment effect is driven by tweets on topics such as religion and migration.

Table 7 allows us to better assess the effect size of NetzDG. While the severe toxicity score is on average 11 pp higher in sensitive topics than across all topics, NetzDG reduces the average score of severe toxicity by 2 pp in those tweets.

Table 7: Baseline with All Tweets: Effect of NetzDG on Hate Intensity in All Tweets and Tweets on Sensitive Topics

	(1)	(2)	(3)	(4)	(5)	(6)
	sev. Toxicity	Toxicity	Threat	ID attack	Profanity	Insult
Treated after T.	-0.00	-0.00	0.01	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Sensitive topic	0.11***	0.13***	0.07***	0.28***	0.03***	0.09***
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Treated after T.						
× Sensitive topic	-0.02*	-0.02**	-0.01**	-0.02	-0.01	-0.01*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.13	0.18	0.10	0.25	0.10	0.16
Observations	2270652	2270652	2270652	2270652	2270652	2270652
Mean of Outcome	0.18	0.28	0.26	0.27	0.16	0.27
SD of Outcome	0.20	0.23	0.19	0.24	0.19	0.22

Notes. The table shows the main coefficients of the difference-in-differences estimations comparing the hate intensity in all tweets by users affected and unaffected by the law (NetzDG). The columns contain the different outcome measures discussed in the data section: Continuous scores ranging from 0 to 1 with regard to severe toxicity, toxicity etc. as calculated by Perspective API. The coefficient *Treated after T.* shows the change in hate intensity in terms of percentage point changes for users located in Germany after NetzDG became effective; The interaction with *sensitive topic* shows the additional effect on tweets containing migration and religion specific buzzwords. Besides the treatment effects, all estimations control for country specific events of regional/national elections and terroristic attacks, the day of the week the tweet was sent and an indicator if the tweet was sent during night times. All estimations include a constant and user fixed effects, year-month fixed effects and fixed effects for the account age in months at which the respective tweet was posted. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

7.2 The Effect of NetzDG on Tweeting Style

The evidence in the baseline analysis shows that hate intensity in tweets decreased in tweets after NetzDG was implemented in Germany. However, when the use of severely toxic language is bounded by the law, online users may substitute hateful words by other ways to express hate such as posting hateful images, adding links to special kinds of media, or increasing the overall negativity. Therefore, we additionally test how other aspects of tweets changed in response to NetzDG estimating regressions similar to our baseline regression as in Section 5. However, instead of continuous scores ranging from 0 to 1, the tweeting style measures are now dummy variables for columns 1-3 and count data for columns 4-8. Table 8 suggests that contrary to the measures related to hateful speech, the tweeting style among German users only changed marginally as compared to Austrian users. The only marginally significant increase can be observed in the propensity of including a media link in tweets. Overall, these results strengthen our findings that NetzDG is effective in targeting hateful speech. At the same time, we do not observe potential substitution patterns, i.e. posting hateful memes, which could not be captured by our hate speech measures.

Table 8: Substitution Patterns: Effect of NetzDG on Non-language Outcomes

	(1) Photos	(2) URL	(3) Media Link	(4) N. Hashtags	(5) N. Words	(6) Positivity	(7) Negativity	(8) Tweet. Freq.
Treated after T.	0.00	0.03	0.02	0.13	1.20	0.00	-0.00	-0.41
	(0.02)	(0.02)	(0.01)	(0.24)	(1.46)	(0.00)	(0.00)	(2.70)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.25	0.47	0.14	0.50	0.44	0.03	0.04	0.71
Observations	160161	160161	160161	160161	160161	160161	160161	12002
Mean of Outcome	0.07	0.68	0.06	1.02	18.19	0.04	0.09	13.29
SD of Outcome	0.26	0.46	0.24	2.05	9.60	0.06	0.08	34.27

Notes. The table shows the main coefficients of the difference-in-differences estimations comparing tweet characteristics by users affected and unaffected by the law (NetzDG). The columns contain the different outcome types: Col (1)-(3) are indicators for Photos, an URL or Media Link in the tweet. Col. (4) and (5) are counts for the number of hashtags and words, while Col. (6) and (7) are normalised measures for positivity and Negativity (share of positive/negative words in tweet). Col. (8) analyses the change in monthly tweeting frequency per user on a monthly basis. Besides the treatment effect, all estimations control for country specific events of regional/national elections and terroristic attacks, the day of the week the tweet was sent and an indicator if the tweet was sent during night times. All estimations include a constant and user fixed effects, year-month fixed effects and fixed effects for the account age in months at which the respective tweet was posted. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

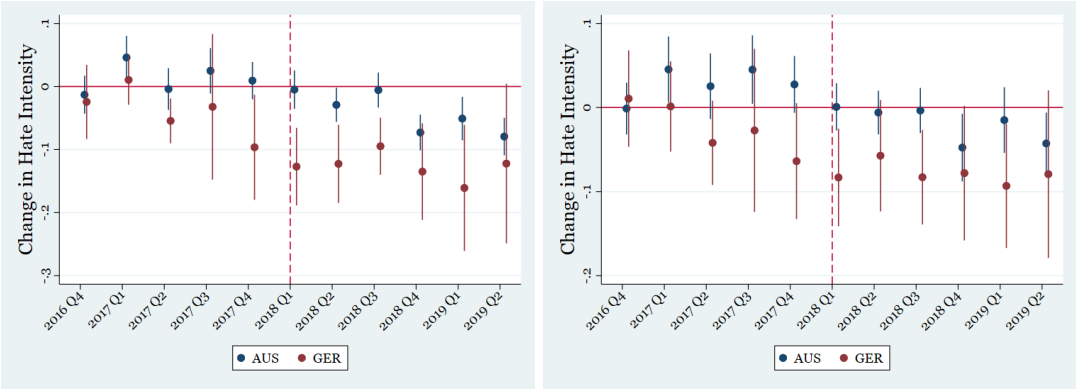
7.3 Spillover Effects to Austria

Although NetzDG imposes stricter deleting rules for tweets visible on German territory, network effects might also affect tweets sent by users located in Austria. For example, Seering, Kraut, and Dabbish 2017 find that imitation behavior of online users of the platform Twitch exist and that the likelihood of a messaged question increases after the event of a question in a previous message. Translating this finding to our context, a reduced number and intensity of hateful posts in Germany could also reduce the hate intensity in Austria. To investigate this effect, we exploit the fact that the average level of hatefulness

of users greatly varies and that users who employ "hateful" language prior to NetzDG are more affected by the law than users employing more objective or soft language. If there are spillover effects of NetzDG to Austria, then these mechanisms should be the same in both countries. Therefore, we conduct country-specific DiD analyses, in which those users with an average hatefulness score above 0.5 compose the treatment group, just as we did in Section **??**. More specifically, we conduct a separate DiD analysis for Austria and for Germany and compare the evolution of all outcome measures by users we defined as hateful (treated group) to tweets by other users (control group).

Figure 5 shows exemplarily the development of treatment effects of severe toxicity and identity attacks within Germany and Austria in the same graph for an easier comparison (the figures of all other outcome measures are qualitatively similar, see Figure 7).

Figure 5: Quarterly Treatment Effects in Germany and Austria



(a) Severe Toxicity (b) Identity Attack

Note. The above coefficients plot shows the treatment effect for different quarters with regard to the respective hate dimension written below the subgraph. The treatment group is defined as tweets by ex ante hateful users, since users who post very hatefully should be more affected by the law than users employing a softer language. Each graph comprises two estimation plots, one for Germany and one for Austria, for an easier comparison of the treatment effects in Germany and Austria. Shown are the coefficients of the interaction of a treated tweet (posted by an ex ante hateful users) with different timings for NetzDG and the 90 percent confidence interval, while controlling for specific events of regional/national elections and terrorist attacks, the day of the week the tweet was sent and an indicator if the tweet was sent at night. All estimations include a constant and user fixed effects, year-month fixed effects and fixed effects for the account age in months when the respective tweet was posted. Standard errors are clustered at the user level. The vertical line indicates the real date on which NetzDG became effective.

In all figures, the red coefficients indicating the quarterly treatment effects for German hateful users are below the zero line after the introduction of the law, strengthening our assumption that more hateful users were affected to a higher extend by the law. The blue coefficients indicating the treatment effects for Austrian toxic users are smaller in absolute size. However, the intensity of severe toxicity also decreased significantly during the last three quarters in Austria. This hints at spillover effects to Austria, as some of the coefficients are significantly lower than zero. In fact, this finding does not threaten but

strengthen our baseline results, as the main findings i) hold in a plausible different setting and ii) seem to measure the lower bound of the effect, as NetzDG also seems to impact Austrian users to some extent. This indicates that the overall effect of the regulation might be higher due to spillover effects.

7.4 User Composition

Anecdotal evidence suggests that hateful users migrate to platforms with less or no content moderation in response to the efforts of large and established social media platforms such as Twitter and Facebook to remove hateful content. A salient example of such migration behaviour is the messenger Telegram with public channels, which was not subject to NetzDG until spring 2021 as it was considered a messenger service and not a social platform.[15] Due to the lax rules regarding any kind of UGC, Telegram attracts conspiracy theorists, right-wing extremists, and terrorists.[16] Telegram reportedly received 25 million new users worldwide in a couple of days after the closure of Parler and media campaigns by Facebook and Twitter stating to increase their moderation efforts.[17]

To investigate potential migration patterns, Table 9 shows the share and the number of users in our sample i) only before NetzDG was introduced, ii) only after NetzDG was introduced, and iii) who stayed in the sample before and after the introduction of NetzDG. Out of the users in our estimation sample (i.e., posting about sensitive topics 1.5 years before and after the introduction of NetzDG), less than half were observed before and after NetzDG. According to the general growth of social media platforms (Hölig and Hasebrink 2020), more users joined than left our sample. This pattern is stronger in Germany than in Austria. Remarkably, the share of users leaving the sample is lower in Germany than in Austria. This suggests that NetzDG did not induce additional platform migration patterns and hence that platform migration alone cannot drive our results.

Table 9: User Composition

	Austria		Germany		Total	
	Share	Count	Share	Count	Share	Count
Stayed in Sample	0.47	351	0.46	272	0.47	623
Joined Sample	0.29	215	0.34	205	0.31	420
Left Sample	0.24	174	0.20	120	0.22	294
Observations		740		597		1337

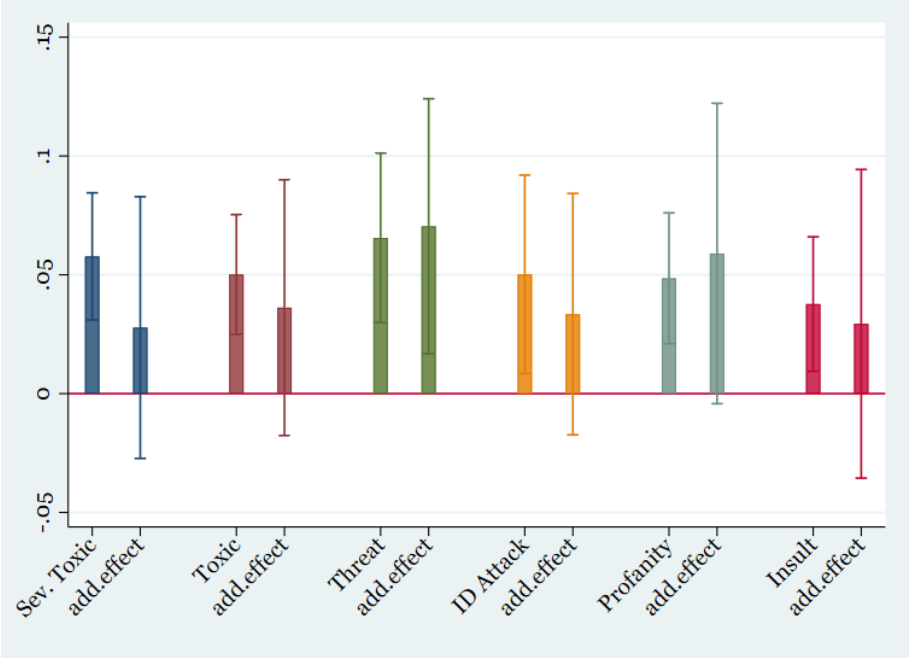
Notes. The table shows the share and absolute number of users who are observed in either both sample periods (before and after NetzDG) or only before or only after NetzDG.

[15] Politico Article
[16] Spiegel Article
[17] Politico Article

7.5 User Engagement with Hateful Tweets

Next, we examine how user engagement with tweets changes after the introduction of NetzDG. User engagement can be measured by the number of likes, retweets, and replies a tweet receives and greatly differs in the tweets in our sample (see Table 3). As in previous sections, we define a tweet as e.g., an identity attack if the probability of containing an identity attack exceeds 0.5. To causally analyze if the user engagement with these sorts of posts changed in response to the law, we apply a difference-in-difference-in-differences (DiDiD) approach. Since there was a general increase in the number of Twitter users in both countries, it is important to account for these time trends by comparing the user engagement with German and Austrian hateful tweets and other tweets before and after NetzDG.

Figure 6: Coefficients Plot: Hateful Tweets are More Often Retweeted



Note: Coefficients plot of the DiDiD estimation comparing the number of retweets (in logs) of hateful and non hateful tweets before and after NetzDG by treated and untreated users. The first colored bars show the coefficient for a hateful (i.e., severely toxic) tweet while the second bars show the additional treatment effect for those hateful tweets due to NetzDG. All estimations include interaction terms "AfterT X Germany", "AfterT X Hateful", and "Germany X Hateful" and control for country-specific events such as elections and terrorist attacks, the day of the week the tweet was posted and an indicator if the tweet was posted at night. All estimations include a constant and user fixed effects, year-month fixed effects and fixed effects for the account age in month. Standard errors are clustered at the user level.

Figure 6 presents the coefficients of the DiDiD estimation analysing the impact of NetzDG on the log number of "retweets" of individual tweets, while the regression tables for all indicators of user engagement can be found in Tables 23a - 25b in the Appendix.

The first bar of each color shows the coefficients of the indicator if a tweet was classified as hateful (toxic, insulting, etc.). The second bar illustrates the treatment effect for hateful tweets. Further interaction coefficients of the DiDiD analysis are shown in Tables 23a and 23b in the Appendix. This analysis shows that hateful tweets receive significantly higher user engagement, collecting significantly more likes (7%-9%) and replies (3%-5%) and are more often retweeted (3%-8%) than the non-hateful ones. At the same time, the treatment effect of NetzDG on the user engagement with hateful tweets is not statistically significant at the 10% confidence level. This indicates that NetzDG neither increases nor decreases the spread of hateful tweets.

Since Twitter displays popular tweets on other users' timelines,[18] the significantly higher user engagememt with potentially illegal tweets has implications for the overall treatment effect. As hateful tweets are retweeted more often, a decrease in the number of hateful original posts decreases the total number of hateful tweets overproportionaly. This is due to an indirect treatment effect, as one prevented hateful tweet prevents even more retweets than a non-hateful tweet. Hence, our baseline specification measuring the hate intensity and volume of original hateful posts documents the lower bound of the policy effect.

8 Discussion

Since our data was drawn ex post, our analyses only include tweets that were still visible on the platform in May 2020. This does not allow us to analyze overdeletion by the platform because we cannot measure hate speech in the deleted tweets and can therefore not comment on the rightfulness of the deletion. Furthermore, we cannot distinguish which of the three mechanisms explained in Section 1 drives the reduction in hateful speech to what extent - is it mostly that users adapt their language and/or migrate to other platforms or is the reduction due to tweets that were deleted by Twitter? However, the above analyses allow to comment on the likelihood of the different mechanisms.
First, our results do not suggest overdeletion issues. As shown in 7.1, we only find significant effects of NetzDG in tweets about sensitive topics. Since the probability of encountering unlawful posts in these tweets is higher, we expect a higher effect of NetzDG in these tweets if Twitter does not randomly delete posts. This indicates that Twitter carefully moderates content and is successfull in targeting prone tweets without changing less targeted tweets. Additionally, the fact that NetzDG does not affect tweeting styles (see Section 7.2), indicates that Twitter targets in fact hate speech and not other forms of communication such as videos, photos, or URLs.

Second, our results show that the effect of NetzDG has not been entirely driven by user migration to other platforms. While this potential mechanism of hateful users migrating

[18] Twitter Help

to other platforms would not tackle the problem of online hate speech but transfer it to other platforms, both other channels (deletion by Twitter or user adaptation to employ a softer language) are targeted policy effects. We can show this by the fixed effect estimation in Section 6.2, which only includes users that tweeted at least twice before and after NetzDG and which shows that the treatment effect is apparent for users who stayed in the sample. In fact, our results become even stronger when we limit the sample of users for the panel who tweeted several times before and after the law came into effect. Also the cross section estimations using the sample of users that tweeted before and after NetzDG yields similar results as the full sample estimations. Furthermore, the share of users of our sample that are only present before but not after NetzDG is even higher in Austria than in Germany. This indicates that although platform migration might play a role, it is not solely responsible for the treatment effect.

In addition, the comparison of the distribution of hate intensity before and after NetzDG (Figure 4) shows that the treatment effect is mainly driven by a reduction in hate intensity in the main body of the distribution rather than a drop in the absolute number of very hateful posts. This indicates that the most prominent mechanism at play is the adaptation of users rather than a high deletion rate of very hateful posts by Twitter.

9 Conclusion

Although in the past few years, social media platforms have increasingly attempted to combat hateful speech, policymakers have debated the need for a legal framework to address this problem. Our paper contributes to the ongoing discussion on whether laws are needed in addition to internal governance mechanisms of platforms (e.g., hateful conduct policy on Twitter) and finds that policy regulation can contribute to restraining harmful content even when platforms already have governance rules for the same purpose.

We empirically investigate the effect of the German Network Enforcement Act (NetzDG), a legal regulation aimed at reducing the prevalence of hateful content on social media platforms. We use a natural experiment setting to tease out the causal effect of the regulatory policy on the language used by right-wing sympathizers in the German compared to the Austrian Twittersphere. Throughout the study, we approximate the language employed by users with a variety of measures related to hateful speech as estimated by Perspective API.

We find robust effects of the regulation on decreasing the probability that a tweet is unlawful - i.e., (severely) toxic, profane, insulting, or an identity attack in Germany as opposed to Austria by 2 pp. This reduction implies an average treatement effect of about 6%-10% of the standard deviation. In terms of the absolute volume of potentially unlawful tweets, our estimations yield a decrease of 10% in the number of original hateful tweets. Yet, these effect sizes constitute the lower bound of the overall reduction in hateful speech

due to NetzDG, as we document spillover effects to users outside of Germany and find that hateful tweets generally recieve higher user engagement. Hence, the reduction in the number of original hateful tweets decreases the exposure to hate even more due to prevented retweeting and liking of hateful tweets. Furthermore, we analyze the underlying mechanisms and show that the treatment effect is only present in tweets on sensitive topics such as religion and migration and not in other tweeting style characteristics such as the number of words or uploading images. This suggests that the law is successful in targeting relevant topics without significantly changing less targeted content.

These results are of high policy relevance as they suggest that NetzDG, as one of the few attempts to lawfully moderate UGC, can influence the prevalence of hateful speech. How large the effect sizes should be and whether the regulation has achieved its goal should be evaluated in the dialog between the civil society, policymakers and online platforms. One way to increase the effect size can be to simplify the reporting mechanisms. Currently, users who report a tweet must choose among specific paragraphs of the criminal code and sign that the wrongful reporting of posts is itself a violation of platform rules. This might decrease the incentive to complete the reporting and, hence, the subsequent deletion of the post by the platform.

Our results should stimulate the discussions and guide further regulation designs, such as the European Digital Services Act, which also aims at reducing online hate and fake news. Future research should analyze the long term effects of anti-hate legislation on social media and the content of deleted posts to evaluate potential overdeletion issues.

References

Aral, Sinan and Dean Eckles (2019). "Protecting elections from social media manipulation". In: *Science* 365.6456, pp. 858–861.

BMJV (2020). *Bericht der Bundesregierung zur Evaluierung des Gesetzes zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz – NetzDG)*. Tech. rep. Bundesministerium der Justiz und für Verbraucherschutz.

Buiten, Miriam C, Alexandre de Streel, and Martin Peitz (2020). "Rethinking liability rules for online hosting platforms". In: *International Journal of Law and Information Technology* 28.2, pp. 139–166.

Bursztyn, Leonardo, Georgy Egorov, Ruben Enikolopov, and Maria Petrova (2019). "Social Media and Xenophobia: Evidence from Russia". In: *Communication & Identity eJournal*.

Caiani, Manuela and Linda Parenti (2013). "Extreme right groups and the Internet: Construction of identity and source of mobilization". In: *European and American Extreme Right Groups and the Internet, Londres, Ashgate*, pp. 83–112.

Chancellor, Stevie, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury (2016). "# thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities". In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pp. 1201–1213.

Chandrasekharan, Eshwar, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert (2017). "You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech". In: *Proceedings of the ACM on Human-Computer Interaction* 1. CSCW, pp. 1–22.

Chevalier, Judith A and Dina Mayzlin (2006). "The effect of word of mouth on sales: Online book reviews". In: *Journal of Marketing Research* 43.3, pp. 345–354.

DellaVigna, Stefano and Eliana La Ferrara (2015). "Economic and social impacts of the media". In: *Handbook of media economics*. Vol. 1. Elsevier, pp. 723–768.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova (2020). "Social media and protest participation: Evidence from Russia". In: *Econometrica* 88.4, pp. 1479–1514.

Ershov, Daniel and Matthew Mitchell (2020). "The effects of influencer advertising disclosure regulations: Evidence from Instagram". In: *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 73–74.

Fortuna, Paula, Juan Soler, and Leo Wanner (2020). "Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets". In: *Proceedings of the 12th language resources and evaluation conference*, pp. 6786–6794.

Geschke, Daniel, Anja Klaßen, Matthias Quent, and Christoph Richter (2019). "Hass im Netz: Der schleichende Angriff auf unsere Demokratie". In: *Eine Bundesweite Repräsentative Untersuchung*.

Halikiopoulou, Daphne (2018). "A right-wing populist momentum? A review of 2017 elections across Europe". In: *JCMS: Journal of Common Market Studies* 56.S1, pp. 63–73.

Hinnosaar, Marit, Marit Hinnosaar, Michael Kummer, and Olga Slivko (2021). "Wikipedia matters". In: *Journal of Economics and Management Strategy*.

Hölig, Sascha and Uwe Hasebrink (2020). *Reuters Institute Digital News Report 2020: Ergebnisse für Deutschland*. Hans-Bredow-Institut für Medienforschung an der Universität Hamburg.

Liu, Yi, Pinar Yildirim, and Z. John Zhang (2021). *Social Media, Content Moderation, and Technology*. arXiv: `2101.04618 [econ.GN]`.

Müller, Karsten and Carlo Schwarz (2020). "From hashtag to hate crime: Twitter and anti-minority sentiment". In: *SSRN 3149103*.

— (2021). "Fanning the flames of hate: Social media and hate crime". In: *Journal of the European Economic Association*.

Nulty, Paul, Yannis Theocharis, Sebastian Adrian Popa, Olivier Parnet, and Kenneth Benoit (2016). "Social media and political communication in the 2014 elections to the European Parliament". In: *Electoral Studies* 44, pp. 429–444.

Olteanu, Alexandra, Carlos Castillo, Jeremy Boy, and Kush Varshney (2018). "The effect of extremist violence on hateful speech online". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 1.

Pentina, Iryna and Monideepa Tarafdar (2014). "From "information" to "knowing": Exploring the role of social media in contemporary news consumption". In: *Computers in Human Behavior* 35, pp. 211–223.

Perspective API (n.d.). *Jigsaw and Google's Counter Abuse Technology*. `https://www.perspectiveapi.com/`. Accessed: 2021-02-27.

Petrova, Maria, Ananya Sen, and Pinar Yildirim (2021). "Social media and political contributions: The impact of new technology on political competition". In: *Management Science* 67.5, pp. 2997–3021.

Seering, Joseph, Robert Kraut, and Laura Dabbish (2017). "Shaping pro and anti-social behavior on twitch through moderation and example-setting". In: *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pp. 111–125.

Srinivasan, Kumar Bhargav, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan (2019). "Content removal as a moderation strategy: Compliance and other outcomes in the ChangeMyView community". In: *Proceedings of the ACM on Human-Computer Interaction* 3. CSCW, pp. 1–21.

Uyheng, Joshua and Kathleen M Carley (2021). "Characterizing network dynamics of online hate communities around the COVID-19 pandemic". In: *Applied Network Science* 6.1, pp. 1–21.

Wojcik, Stefan and Adam Hughes (2019). “Sizing up Twitter users”. In: *PEW research center* 24.

Xu, Sean Xin and Xiaoquan Zhang (2013). “Impact of Wikipedia on market information environment: Evidence on management disclosure and investor reaction”. In: *MIS Quarterly*, pp. 1043–1068.

Appendices

A Further Information on Data

Table 10: Original Example Tweets in our Sample

Outcome	Value	Example Tweet
Severe Toxicity	1	eckelhafter drecksack...dann verpisst euch hier,ihr hurensöhne,fuck islam
Toxicity	0.99	wie dumm bist du eigentlich? bei dir ist gleich jeder ein pkkler terrorist.du gehörst zurück gepudert und abgetrieben.
Threat	0.99	diesem typ wünsche ich den tod durch einen dieser krimigranten.
Identity Attack	1	jepp, katholiken ficken kinder, moslems schlagen ihnen die fresse ein und schneiden mädchen die klitoris ab. juden und moslems lassen tiere liebevoll ausbluten. religion ist ein hurensohn.
Profanity	0.99	dieses arschkriechen vor dem scheiß islam ist echt nur noch zum kotzen
Insult	0.99	[...] diese deppen kapieren nie wie völkisch moslems sind

Table 11: Translated Example Tweets in our Sample

Outcome	Value	Example Tweet
Severe Toxicity	1	disgusting scumbag...then fuck off here, you sons of bitches, fuck islam
Toxicity	0.99	how stupid are you? for you every pkkler is a terrorist. you belong back powdered and aborted.
Threat	0.99	i wish this guy death by one of these criminals.
Identity Attack	1	yeah, catholics fuck children, muslims smash their faces and cut off girls' clitorises. jews and muslims lovingly bleed animals. religion is a son of a bitch.
Profanity	0.99	this ass-kissing of the fucking islam is really just to vomit
Insult	0.99	[...] these morons never get how nationalistic muslims are

Table 12: Summary Table of Tweet Characteristics

	Total					Austria					Germany				
	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
Toxicity	0.43	0.45	0.22	0.00	1.00	0.41	0.42	0.23	0.00	1.00	0.44	0.46	0.22	0.00	1.00
Severe Toxicity	0.30	0.29	0.24	0.00	1.00	0.28	0.18	0.25	0.00	1.00	0.31	0.29	0.24	0.00	1.00
Threat	0.35	0.21	0.25	0.00	1.00	0.32	0.20	0.24	0.00	0.99	0.36	0.22	0.25	0.00	1.00
Identity Attack	0.57	0.61	0.28	0.00	1.00	0.56	0.60	0.29	0.00	1.00	0.59	0.63	0.27	0.00	1.00
Profanity	0.20	0.11	0.20	0.00	1.00	0.20	0.11	0.21	0.00	1.00	0.21	0.11	0.20	0.00	1.00
Insult	0.37	0.36	0.22	0.00	1.00	0.37	0.36	0.23	0.00	1.00	0.38	0.36	0.21	0.00	1.00
no. Retweets	4.00	0.00	19.35	0.00	911.00	2.62	0.00	13.84	0.00	779.00	4.89	0.00	22.17	0.00	911.00
no. Likes	7.03	0.00	36.12	0.00	1711.00	5.12	0.00	26.19	0.00	1711.00	8.28	0.00	41.28	0.00	1398.00
no. Replies	1.12	0.00	5.62	0.00	292.00	0.86	0.00	3.30	0.00	275.00	1.28	0.00	6.71	0.00	292.00
Video in tweet	0.00	0.00	0.05	0.00	1.00	0.00	0.00	0.06	0.00	1.00	0.00	0.00	0.04	0.00	1.00
Photo	0.07	0.00	0.26	0.00	1.00	0.05	0.00	0.23	0.00	1.00	0.08	0.00	0.28	0.00	1.00
URL	0.68	1.00	0.46	0.00	1.00	0.59	1.00	0.49	0.00	1.00	0.74	1.00	0.44	0.00	1.00
Link to media outlet	0.06	0.00	0.24	0.00	1.00	0.03	0.00	0.18	0.00	1.00	0.08	0.00	0.27	0.00	1.00
no. Words	18.19	15.00	9.60	1.00	57.00	17.96	15.00	9.56	1.00	52.00	18.33	15.00	9.63	1.00	57.00
Tweeted at night	0.08	0.00	0.26	0.00	1.00	0.06	0.00	0.24	0.00	1.00	0.08	0.00	0.28	0.00	1.00
Terrorist attack in country	0.02	0.00	0.12	0.00	1.00	0.00	0.00	0.06	0.00	1.00	0.02	0.00	0.15	0.00	1.00
Election in country	0.01	0.00	0.10	0.00	1.00	0.01	0.00	0.08	0.00	1.00	0.01	0.00	0.11	0.00	1.00

Table 13: Raw Correlation among Outcome Variables

	Severe Toxicity	Toxicity	Threat	Identity Attack	Profanity	Insult
Severe Toxicity	1.00					
Toxicity	0.90***	1.00				
Threat	0.57***	0.53***	1.00			
Identity Attack	0.75***	0.85***	0.38***	1.00		
Profanity	0.86***	0.81***	0.45***	0.59***	1.00	
Insult	0.85***	0.93***	0.39***	0.80***	0.86***	1.00
Observations	160474					

B Baseline and Robustness Checks

Table 14: OLS

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Germany	0.03	0.03*	0.04***	0.04*	0.01	0.01
	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)
Treated after T.=1	-0.02*	-0.03***	-0.01	-0.05***	-0.02*	-0.03***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)
Tweeted at night	0.03***	0.03***	0.00	0.03***	0.02***	0.03***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
verified=1	-0.05***	-0.04***	-0.03***	-0.04**	-0.05***	-0.05***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)
no. Followers	-0.00**	-0.00***	-0.00***	-0.00***	-0.00***	-0.00***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Tuesday	-0.00	-0.00	-0.00	-0.00	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Wednesday	-0.01***	-0.00**	-0.01***	-0.01***	-0.01**	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Thursday	-0.01**	-0.01**	-0.01***	-0.00*	-0.00*	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Friday	0.00	0.00	0.00	-0.00	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Saturday	0.00	0.01**	-0.00	0.01	0.00	0.01**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Sunday	0.00	0.00	-0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Terrorist attack in country	0.00	0.00	0.01**	-0.00	0.00	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Election in country	-0.00	-0.01	-0.00	-0.01	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.39***	0.50***	0.45***	0.61***	0.26***	0.42***
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
month indicators	Yes	Yes	Yes	Yes	Yes	Yes
account age	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.05	0.07	0.02	0.08	0.04	0.06
Observations	160403	160403	160403	160403	160403	160403
Mean of Outcome	0.30	0.43	0.35	0.57	0.20	0.37
SD of Outcome	0.24	0.22	0.25	0.28	0.20	0.22

Clustered standard errors in parentheses, clustered at user_id level, * p<0.10, ** p<0.05, *** p<0.01.
All models include an intercept.

Table 15: Baseline Analysis: The Effect of NetzDG on the Intensity of Hate in Tweets (OLS with FE, all coefficients)

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Treated after T.	-0.02***	-0.02***	-0.01	-0.03***	-0.01***	-0.02***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Tweeted at night	0.01	0.01	0.00	0.00	0.01	0.01
	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)
Tuesday	-0.00	-0.00	-0.00	-0.00	-0.00*	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Wednesday	-0.01**	-0.00*	-0.01***	-0.01***	-0.00**	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Thursday	-0.01**	-0.01**	-0.01***	-0.01**	-0.00**	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Friday	-0.00	0.00	0.00	-0.00*	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Saturday	0.00	0.00	-0.00	0.00	0.00	0.01*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Sunday	0.00	0.00	-0.00	0.00	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Terrorist attack in country	0.00	0.00	0.01*	-0.00	-0.00	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Election in country	-0.00	-0.01	-0.01	-0.00	-0.01	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.31***	0.44***	0.35***	0.58***	0.21***	0.38***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.163	0.177	0.072	0.199	0.123	0.178
Observations	160161	160161	160161	160161	160161	160161
Mean of Outcome	0.300	0.428	0.347	0.574	0.205	0.372
SD of Outcome	0.241	0.225	0.249	0.276	0.205	0.221

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 16: Baseline Analysis in Logs: The Effect of NetzDG on the Logarithmic Intensity of Hate in Tweets

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Treated after T.	-0.01***	-0.02***	-0.01	-0.02***	-0.01***	-0.02***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)
Tweeted at night	0.00	0.00	0.00	0.00	0.01	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Tuesday	-0.00	-0.00	-0.00	-0.00	-0.00*	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Wednesday	-0.00**	-0.00*	-0.00**	-0.00***	-0.00**	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Thursday	-0.00**	-0.00**	-0.01***	-0.00**	-0.00**	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Friday	-0.00	0.00	0.00	-0.00	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Saturday	0.00	0.00	-0.00	0.00	0.00	0.00*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Sunday	0.00	0.00	-0.00	0.00	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Terrorist attack in country	0.00	0.00	0.01*	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Election in country	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Constant	0.25***	0.35***	0.29***	0.44***	0.18***	0.31***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.17	0.18	0.07	0.20	0.13	0.18
Observations	160161	160161	160161	160161	160161	160161
Mean of Outcome	0.25	0.34	0.28	0.44	0.17	0.30
SD of Outcome	0.18	0.16	0.17	0.19	0.15	0.16

Clustered standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 17: Baseline Analysis: The Effect of NetzDG on the Intensity of Hate in Tweets (Including the Full List of Tweet Characteristics as Control Variables)

	(1)	(2)	(3)	(4)	(5)	(6)
	sev. Toxicity	Toxicity	Threat	ID attack	Profanity	Insult
Treated after T.	-0.02**	-0.02***	-0.01**	-0.03**	-0.02***	-0.03***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
no. Retweets	0.00***	0.00***	0.00***	0.00***	0.00**	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
no. Likes	-0.00*	-0.00	-0.00***	-0.00***	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
no. Replies	-0.00	0.00	-0.00***	0.00**	0.00	0.00**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Video in tweet	0.00	-0.01	0.00	-0.03	-0.02	-0.02
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Photo	0.01	-0.00	0.01*	-0.01*	-0.00	-0.01
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
URL	-0.02***	-0.04***	0.02***	-0.07***	-0.03***	-0.04***
	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)
Link to media outlet	-0.02***	-0.01***	0.03***	-0.05***	0.01	-0.01**
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)
no. Words	0.00***	0.00***	0.00***	0.01***	0.00***	0.00***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Tweeted at night	0.01	0.01	0.00	-0.00	0.01*	0.01
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Tuesday	-0.00	-0.00	-0.00	-0.00	-0.00*	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Wednesday	-0.01**	-0.00*	-0.01***	-0.01***	-0.00**	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Thursday	-0.01**	-0.01***	-0.01***	-0.01**	-0.00**	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Friday	-0.00	0.00	0.00	-0.00	-0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Saturday	0.00	0.00	-0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Sunday	-0.00	0.00	-0.00	0.00	-0.00	-0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Terrorist attack in country	0.00	0.00	0.01**	-0.00	-0.00	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Election in country	-0.00	-0.01	-0.00	-0.00	-0.01	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.170	0.196	0.086	0.234	0.141	0.209
Observations	160161	160161	160161	160161	160161	160161
Mean of Outcome	0.300	0.428	0.347	0.574	0.205	0.372
SD of Outcome	0.241	0.225	0.249	0.276	0.205	0.221

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 18: Robustness Check: Sample Restricted to Users Tweeting Before and After NetzDG

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Treated after T.	-0.02**	-0.02***	-0.01	-0.03***	-0.01**	-0.02***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.17	0.18	0.07	0.20	0.12	0.18
Observations	110612	110612	110612	110612	110612	110612
Mean of Outcome	0.29	0.42	0.34	0.56	0.19	0.36
SD of Outcome	0.24	0.22	0.25	0.28	0.20	0.22

Clustered standard errors in parentheses. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 19: Robustness Check: Sample without users living outside Germany/Austria

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Treated after T.	-0.02***	-0.02***	-0.01	-0.03***	-0.01**	-0.02***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.30***	0.43***	0.35***	0.58***	0.20***	0.37***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.14	0.17	0.07	0.19	0.11	0.17
Observations	140872	140872	140872	140872	140872	140872
Mean of Outcome	0.29	0.42	0.35	0.56	0.20	0.36
SD of Outcome	0.24	0.22	0.25	0.28	0.20	0.22

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 20: Robustness Check: Baseline Analysis Excluding Transition Period (July'17-Dec'17)

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Treated after T.	-0.02**	-0.02***	-0.01	-0.03***	-0.01**	-0.02***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.16	0.18	0.07	0.20	0.12	0.18
Observations	135881	135881	135881	135881	135881	135881
Mean of Outcome	0.30	0.43	0.35	0.57	0.20	0.37
SD of Outcome	0.24	0.22	0.25	0.28	0.20	0.22

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 21: Robustness Check: Setting NetzDG to Jan2017

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Treated before T.	-0.01	-0.01	0.01	-0.02	-0.01	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.30***	0.43***	0.34***	0.58***	0.21***	0.38***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.16	0.18	0.07	0.20	0.12	0.18
Observations	160161	160161	160161	160161	160161	160161
Mean of Outcome	0.30	0.43	0.35	0.57	0.20	0.37
SD of Outcome	0.24	0.22	0.25	0.28	0.20	0.22

Clustered standard errors in parentheses. All models include an intercept.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 22: Panel: Volume of Outcome Variables by User-Month in Logs

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Treated after T.	-0.08*	-0.11**	-0.08	-0.11*	-0.08*	-0.10**
	(0.05)	(0.05)	(0.05)	(0.06)	(0.04)	(0.05)
first month in sample	-0.25***	-0.29***	-0.25***	-0.37***	-0.18***	-0.28***
	(0.04)	(0.05)	(0.04)	(0.05)	(0.04)	(0.05)
night	0.11**	0.14**	0.09*	0.08	0.12***	0.15***
	(0.05)	(0.06)	(0.05)	(0.06)	(0.04)	(0.06)
attack in country	0.10	0.13	0.04	0.19*	0.03	0.14
	(0.08)	(0.09)	(0.08)	(0.11)	(0.06)	(0.09)
election in country	0.08	0.10	0.15*	0.11	0.09	0.15*
	(0.07)	(0.08)	(0.08)	(0.11)	(0.06)	(0.08)
Constant	1.26***	1.56***	1.42***	2.12***	0.89***	1.42***
	(0.07)	(0.07)	(0.07)	(0.08)	(0.06)	(0.07)
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.025	0.027	0.036	0.027	0.015	0.020
Observations	9546	9546	9546	9546	9546	9546
Groups	492	492	492	492	492	492
Mean of Outcome	3.078	4.635	3.410	9.454	1.660	3.932
SD of Outcome	8.512	11.254	9.915	23.163	4.293	9.077

Clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$. All models include an intercept.

C User Engagement

Table 23a: Cross Section with FE: User Engagement with Potentially Unlawful Tweets - Log Retweets

	(1) sev. Toxicity	(2) Toxicity	(3) Threat
Germany			
× AfterT	0.09	0.08	0.07
	(0.06)	(0.06)	(0.05)
Severely toxic	0.06***		
	(0.01)		
Germany			
× Severely toxic	-0.02		
	(0.02)		
AfterT			
× Severely toxic	0.01		
	(0.02)		
Germany			
× AfterT			
× Severely toxic	0.03		
	(0.03)		
Toxic		0.05***	
		(0.01)	
Germany			
× Toxic		-0.02	
		(0.02)	
AfterT			
× Toxic		-0.00	
		(0.02)	
Germany			
× AfterT			
× Toxic		0.04	
		(0.03)	
Threat			0.07***
			(0.02)
Germany			
× Threat			-0.03
			(0.02)
AfterT			
× Threat			-0.01
			(0.02)
Germany			
× AfterT			
× Threat			0.07**
			(0.03)
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes
R^2	0.54	0.54	0.54
Observations	160165	160165	160165
Mean of Outcome	0.56	0.56	0.56
SD of Outcome	1.00	1.00	1.00

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 23b: Cross Section with FE: User Engagement with Potentially Unlawful Tweets - Log Retweets

	(1) ID attack	(2) Profanity	(3) Insult
Germany			
× AfterT	0.07	0.09	0.08
	(0.05)	(0.06)	(0.06)
ID Attack	0.05**		
	(0.02)		
Germany			
× ID Attack	0.01		
	(0.03)		
AfterT			
× ID Attack	0.01		
	(0.02)		
Germany			
× AfterT			
× ID Attack	0.03		
	(0.03)		
Profanity		0.05***	
		(0.01)	
Germany			
× Profanity		-0.03	
		(0.02)	
AfterT			
× Profanity		-0.01	
		(0.02)	
Germany			
× AfterT			
× Profanity		0.06*	
		(0.03)	
Insult			0.04***
			(0.01)
Germany			
× Insult			-0.01
			(0.02)
AfterT			
× Insult			0.01
			(0.02)
Germany			
× AfterT			
× Insult			0.03
			(0.03)
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes
R^2	0.54	0.54	0.54
Observations	160165	160165	160165
Mean of Outcome	0.56	0.56	0.56
SD of Outcome	1.00	1.00	1.00

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 24a: Cross Section with FE: User Engagement with Potentially Unlawful Tweets - Log Likes

	(1) sev. Toxicity	(2) Toxicity	(3) Threat
Germany			
× AfterT	-0.00	-0.01	0.00
	(0.07)	(0.07)	(0.07)
Severely toxic	0.07***		
	(0.02)		
Germany			
× Severely toxic	-0.02		
	(0.02)		
AfterT			
× Severely toxic	0.01		
	(0.03)		
Germany			
× AfterT			
× Severely toxic	0.03		
	(0.04)		
Toxic		0.08***	
		(0.02)	
Germany			
× Toxic		-0.02	
		(0.02)	
AfterT			
× Toxic		-0.01	
		(0.03)	
Germany			
× AfterT			
× Toxic		0.06*	
		(0.03)	
Threat			0.02
			(0.03)
Germany			
× Threat			0.01
			(0.03)
AfterT			
× Threat			0.03
			(0.02)
Germany			
× AfterT			
× Threat			0.00
			(0.03)
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes
R^2	0.55	0.55	0.55
Observations	160165	160165	160165
Mean of Outcome	0.75	0.75	0.75
SD of Outcome	1.14	1.14	1.14

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 24b: Cross Section with FE: User Engagement with Potentially Unlawful Tweets - Log Likes

	(1) ID Attack	(2) Profanity	(3) Insult
Germany			
× AfterT	-0.02	-0.00	-0.01
	(0.07)	(0.07)	(0.07)
ID Attack	0.07***		
	(0.02)		
Germany			
× ID Attack	0.00		
	(0.03)		
AfterT			
× ID Attack	0.02		
	(0.03)		
Germany			
× AfterT			
× ID Attack	0.05		
	(0.03)		
Profanity		0.07***	
		(0.02)	
Germany			
× Profanity		-0.03	
		(0.02)	
AfterT			
× Profanity		-0.01	
		(0.03)	
Germany			
× AfterT			
× Profanity		0.07*	
		(0.04)	
Insult			0.08***
			(0.02)
Germany			
× Insult			-0.02
			(0.02)
AfterT			
× Insult			0.01
			(0.03)
Germany			
× AfterT			
× Insult			0.05
			(0.04)
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes
R^2	0.55	0.55	0.55
Observations	160165	160165	160165
Mean of Outcome	0.75	0.75	0.75
SD of Outcome	1.14	1.14	1.14

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 25a: Cross Section with FE: User Engagement with Potentially Unlawful Tweets - Log Replies

	(1) sev. Toxicity	(2) Toxicity	(3) Threat
Germany			
× AfterT	-0.02	-0.03	-0.03
	(0.05)	(0.05)	(0.05)
Severely toxic	0.03***		
	(0.01)		
Germany			
× Severely toxic	-0.02		
	(0.02)		
AfterT			
× Severely toxic	-0.00		
	(0.02)		
Germany			
× AfterT			
× Severely toxic	0.01		
	(0.03)		
Toxic		0.04***	
		(0.01)	
Germany			
× Toxic		-0.02	
		(0.02)	
AfterT			
× Toxic		-0.01	
		(0.02)	
Germany			
× AfterT			
× Toxic		0.02	
		(0.02)	
Threat			0.01
			(0.01)
Germany			
× Threat			0.00
			(0.01)
AfterT			
× Threat			-0.00
			(0.01)
Germany			
× AfterT			
× Threat			0.01
			(0.01)
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes
R^2	0.43	0.43	0.43
Observations	160165	160165	160165
Mean of Outcome	0.32	0.32	0.32
SD of Outcome	0.65	0.65	0.65

Clustered standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

Table 25b: Cross Section with FE: User Engagement with Potentially Unlawful Tweets - Log Replies

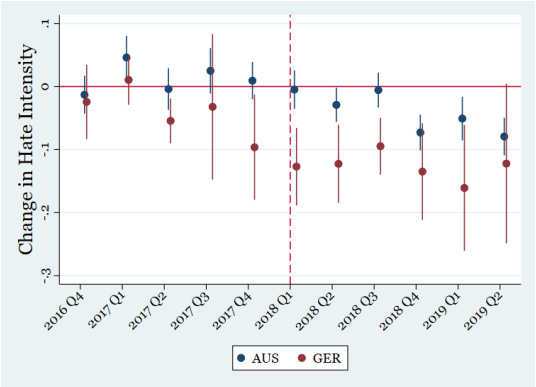
	(1) ID attack	(2) Profanity	(3) Insult
Germany			
× AfterT	-0.04	-0.02	-0.03
	(0.05)	(0.05)	(0.05)
ID Attack	0.05**		
	(0.02)		
Germany			
× ID Attack	-0.01		
	(0.03)		
AfterT			
× ID Attack	-0.01		
	(0.02)		
Germany			
× AfterT			
× ID Attack	0.02		
	(0.02)		
Profanity		0.03**	
		(0.01)	
Germany			
× Profanity		-0.01	
		(0.02)	
AfterT			
× Profanity		-0.01	
		(0.02)	
Germany			
× AfterT			
× Profanity		0.01	
		(0.03)	
Insult			0.05***
			(0.01)
Germany			
× Insult			-0.03
			(0.02)
AfterT			
× Insult			-0.02
			(0.02)
Germany			
× AfterT			
× Insult			0.03
			(0.02)
Month FE	Yes	Yes	Yes
User FE	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes
Day of the Week	Yes	Yes	Yes
R^2	0.43	0.43	0.43
Observations	160165	160165	160165
Mean of Outcome	0.32	0.32	0.32
SD of Outcome	0.65	0.65	0.65

Clustered standard errors in parentheses

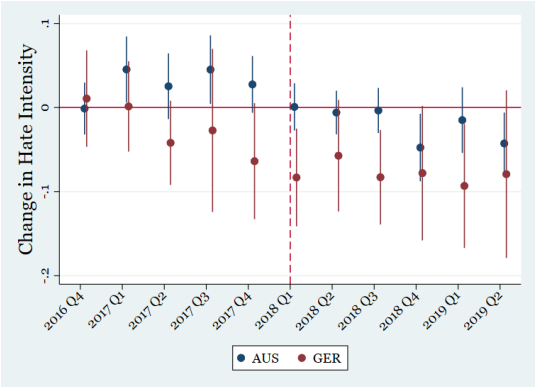
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$

D Spillovers to Austria

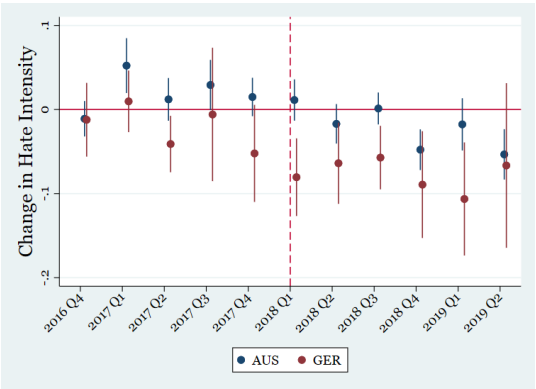
Figure 7: Quarterly Treatment Effects within Germany and Austria



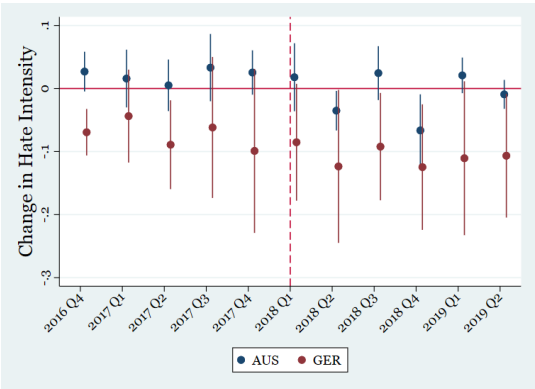
(a) Severe Toxicity



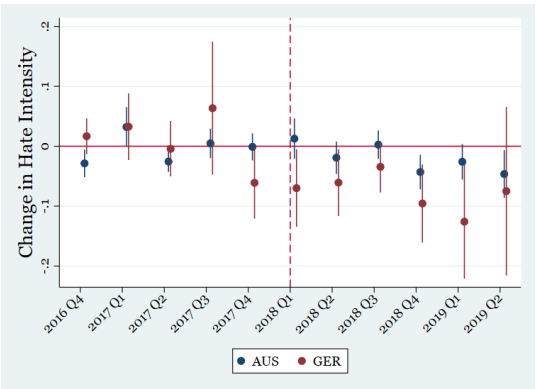
(b) Identity Attack



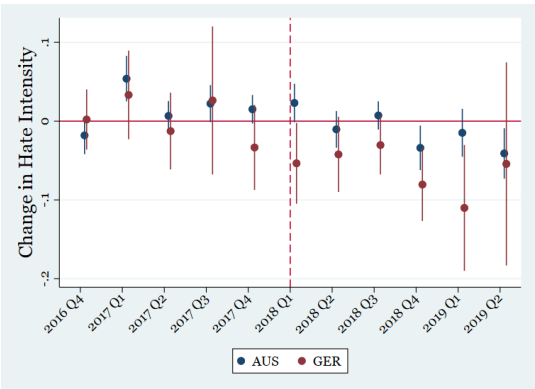
(c) Toxicity



(d) Threat



(e) Profanity



(f) Insult

E Tables Using All Tweets

Table 26: Summary Table of Tweet Characteristics

	N	Mean	Median	SD	Min	Max
no. Retweets	2271745	1.33	0.00	10.25	0.0	3891.0
no. Likes	2271745	3.25	0.00	22.63	0.0	8984.0
no. Replies	2271745	0.58	0.00	3.28	0.0	646.0
Video in tweet	2271745	0.01	0.00	0.07	0.0	1.0
Photo	2271745	0.07	0.00	0.26	0.0	1.0
URL	2271745	0.50	1.00	0.50	0.0	1.0
Link to media outlet	2271745	0.02	0.00	0.13	0.0	1.0
no. Words	2271745	15.04	13.00	9.26	1.0	88.0
Tweeted at night	2271745	0.07	0.00	0.26	0.0	1.0
Terrorist attack in country	2271745	0.01	0.00	0.10	0.0	1.0
Election in country	2271745	0.01	0.00	0.12	0.0	1.0

Table 27: Outcome variables by country and before/after

	Austria before	Austria after	Germany before	Germany after
	Mean	Mean	Mean	Mean
Toxicity	0.24	0.26	0.31	0.30
Severe Toxicity	0.16	0.16	0.21	0.20
Threat	0.23	0.23	0.29	0.28
Identity Attack	0.23	0.24	0.31	0.31
Profanity	0.14	0.16	0.18	0.18
Insult	0.23	0.26	0.29	0.29
Observations	531417	709716	463603	567009

Table 28: Raw Correlation among Outcome Variables

	Severe Toxicity	Toxicity	Threat	Identity Attack	Profanity	Insult
Severe Toxicity	1.00					
Toxicity	0.88***	1.00				
Threat	0.55***	0.51***	1.00			
Identity Attack	0.72***	0.77***	0.48***	1.00		
Profanity	0.89***	0.85***	0.38***	0.57***	1.00	
Insult	0.82***	0.95***	0.38***	0.73***	0.85***	1.00
Observations	2271745					

Table 29: Panel: Volume of Outcome Variables by User-Month

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Treated after T.	-2.02*	-3.05*	-0.83	-2.62	-1.54*	-2.90**
	(1.09)	(1.56)	(1.19)	(2.06)	(0.85)	(1.45)
month indicators	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.003	0.003	0.001	0.002	0.004	0.006
Observations	30290	30290	30290	30290	30290	30290
Mean of Outcome	6.753	11.146	7.917	14.781	6.310	11.552
SD of Outcome	22.844	34.795	36.022	49.967	19.548	34.333

Robust standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

All models include an intercept.

Table 30: Panel: Volume of Outcome Variables by UserMonth in Logs

	(1)	(2)	(3)	(4)	(5)	(6)
	sev. Toxicity	Toxicity	Threat	ID attack	Profanity	Insult
Treated after T.	-0.02	-0.01	-0.03	-0.02	-0.02	-0.02
	(0.04)	(0.05)	(0.04)	(0.05)	(0.04)	(0.05)
month indicators	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.009	0.008	0.011	0.008	0.009	0.009
Observations	30290	30290	30290	30290	30290	30290
Mean of Outcome	0.908	1.153	0.962	1.262	0.898	1.185
SD of Outcome	1.220	1.385	1.247	1.475	1.203	1.399

Outcomes in logs and clustered standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

Download ZEW Discussion Papers from our ftp server:

https://www.zew.de/en/publications/zew-discussion-papers

or see:

https://www.ssrn.com/link/ZEW-Ctr-Euro-Econ-Research.html
https://ideas.repec.org/s/zbw/zewdip.html

IMPRINT

ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim
ZEW – Leibniz Centre for European Economic Research

L 7,1 · 68161 Mannheim · Germany
Phone +49 621 1235-01
info@zew.de · zew.de