// NO.21-103 | 04/2023

DISCUSSION PAPER

// RAPHAELA ANDRES AND 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†

First Version: December 2021
This Version: March 2023

Abstract

We analyze the impact of the first regulation restraining hate speech on large social media platforms. Exploiting the Network Enforcement Act (NetzDG) in a quasi-experimental approach, we measure the causal impact of the law on the prevalence of hateful content in the German-speaking segment of Twitter. We find evidence of a significant and robust decrease in the intensity and volume of hate speech in tweets tackling sensitive migration- and religion-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 significantly reduce 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: +31 620 971295. Department of Technology and Operations Management, Rotterdam School of Management, Erasmus University, Rotterdam, The Netherlands.

We are grateful for helpful comments from Shrabastee Banerjee, Irene Bertschek, Dominik Gutt, Ulrich Laitenberger, Dominik Rehse, Felix Rusche, Anthony Strittmatter, Michael Ward, Michael Zhang, seminar participants at Rotterdam School of Management and ZEW Mannheim 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 cause *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 extend their user base and, hence, profits from advertising (Liu, Yildirim, and Zhang 2021). As the first legal framework to combat hate speech, the German government implemented the Network Enforcement Act (NetzDG) in January 2018. This law obliges large social media platforms in Germany to implement simple procedures for users to report hateful content and requires the social networks to shortly remove hateful content. After NetzDG was implemented in Germany, it has stimulated discussions about regulation of harmful content worldwide and was used as a blueprint for designing similar laws in other countries (Tworek and Leerssen 2019) and the European Digital Services Act.

We exploit the implementation of NetzDG in a quasi-experimental framework and measure its causal impact on the user-generated content (UGC) production focusing on a target group of German Twitter users. Specifically, we investigate if the introduction of the law mitigates the prevalence and intensity of hate speech in tweets of German right-wing sympathizers. For measuring hate speech, we use pre-trained algorithms provided by Jigsaw and Google's Counter Abuse Technology team which have demonstrated quite accurate performance (Mondal, Silva, and Benevenuto 2017, ElSherief et al. 2018, Han and Tsvetkov 2020). 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 in the German and Austrian Twittersphere. Since the two neighbouring countries share many cultural aspects,

[1] Our World in Data

[2] Wired Article ; Original Quote ; Recent example

speak the same mother tongue (German) and potential remaining differences will be differenced out by our identification strategy, we are able to isolate the causal effect of the regulation.

We find that the regulation reduces the intensity of hate speech in Germany by about 2 percentage points, which corresponds to a reduction of 6%-11% in mean hate intensity. The volume of original hateful tweets is reduced by 11%, implying one less attacking tweet by each user in three months. These effects are remarkable and contribute to the current discussion on whether legal regulation of online content can complement the platform's own guidelines such as the "Twitter hateful conduct policy". In a survey about NetzDG sent to the platforms, Twitter claims that the law has not increased the deletions on the platform, as most of the illegal acts defined in NetzDG were already captured by its house rules (Liesching et al. 2021). Yet, the incentive of the platforms are not fully aligned with stronger content moderation. This is supported by our findings, which show that while the platform guidelines apply to both German and Austrian users, hateful content in tweets posted in Germany significantly decreased after implementing the law. Furthermore, in recent months, the acquisition of Twitter by Elon Musk and the subsequent lay-off of thousands of content moderators have led to a lawsuit against Twitter in Germany for failing to timely remove the illegal content.[3] Without this legislation, protecting social media environments from abusive and harmful content would fully depend on the free will of the platform managers.

2 Literature

2.1 The Effects of Social Media

Previous studies document substantial offline impacts of social media. Allcott and Gentzkow 2017, Zhuravskaya, Petrova, and Enikolopov 2020, Enikolopov, Makarin, and Petrova 2020 and Vosoughi, Roy, and Aral 2018 highlight the role of social media in "fake" news dissemination and political polarization and mobilization. Morales 2020 shows how that authoritarian governments can manipulate social media and public opinion enhancing the perceived support of their policies using bots. Zhuravskaya, Petrova, and Enikolopov 2020 review the strategies of authoritarian governments to manipulate public opinions and information access on social media via selective censorship and distracting information (Gorodnichenko, Pham, and Talavera 2021, Bail et al. 2020). Furthermore, engaging with social media strongly affects individual well-being (Allcott, Braghieri, et al. 2020, Braghieri, Levy, and Makarin 2022). Allcott, Braghieri, et al. 2020 draw a link between a temporal social media deactivation and improved subjective well-being, as well as reduction in news consumption and political polarization. Braghieri, Levy, and Makarin 2022 analyze the effect of the staggered rollout of Facebook across US campuses on student mental health and found that the negative effect of social media adoption was stronger for students that might suffer from unfavourable social comparisons.

A recent strand of studies identifies an important link between xenophobic attitudes expressed on social media and offline hate crimes (Jiménez Durán, Müller, and Schwarz 2022, Müller and

[3] Euractive Article

Schwarz 2021, Müller and Schwarz 2020, Bursztyn et al. 2019, Olteanu et al. 2018). Müller and Schwarz 2021 measure the short-run effect of social media on violent crimes. They show that the effect of anti-refugee sentiments posted on Facebook disappears on the days of Internet outages and disruptions to Facebook access in Germany. Bursztyn et al. 2019 measure long-term effects of social media penetration in Russia on anti-immigrant hate crimes. Additionally, Olteanu et al. 2018 show that offline violence (Islamist attacks) causes online hate speech against muslims across social media platforms. Additionally to the strong connection between online and offline hate, Beknazar-Yuzbashev et al. 2022 show that toxic UGC is contagious and users exposed to lower toxicity reduce their own toxicity in posts and comments on Facebook and Twitter.

Hence, policy makers, platform stakeholders and the civil society intensely debate the necessity of moderating hateful content and the possible side effects (e.g. censorship and limitation to the freedom of speech). Our paper contributes to this debate by presenting sound empirical evidence on the effects of harmful content moderation on social media imposed by the German government regulator. Our findings suggest that the harmful effects of online hate can be weakened if platforms are obliged by law to timely address user-reported hateful content. Moreover, we show that only UGC covering sensitive topics are affected by the regulation, while we find no effect on other topics or the user style of expression on Twitter.

2.2 Content Moderation and Regulation on Platforms

Theoretical studies on the content moderation suggest that the incentives of the social media platforms to provide the optimal level of content regulation may be insufficient (Buiten, Streel, and Peitz 2020, Liu, Yildirim, and 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 Zhang 2021). Additionally, the design of the regulation also matters. Feher 2023 shows that a uniform regulation that treats all platform users in the same way could encourage platforms to punish only users with low overall impact on the platform in order to avoid sanctions while keeping the ones with high impact. He highlights the importance to consider the harm that concrete platform users may cause due to their audience sizes in the regulation design.

In the recent years, social media platforms increasingly undertake efforts to set boundaries on misinformation and harmful content prevalence. For misinformation, platforms experimented with the implementation of nudges as well as peer content moderation mechanisms. Ershov, Morales, et al. 2021 analyze how Twitter users responded to the user interface change nudging users into adding a comment on the content they were going to share. After this change, content sharing was significantly reduced, and while there was no difference for low vs high factualness, left-wing media experienced a very high drop in sharing compared to the right-wing media.
Several studies assessed platform governance mechanisms for content moderation. Chandrasekharan et al. 2017 study an event of banning two hateful communities on Reddit and show that after the ban some users reduced their usage of hate terms, while others left the platform. Srinivasan et al. 2019 analyze the evolution of swear words and hate terms within a subreddit and find

no relationship between content removal and the use of hate terms for non-compliant users. Borwankar, Zheng, and Kannan 2022 assess the impact of the Birdwatch program on Twitter and show that peer content moderation increases cognition in writing and decreases content extremity at the cost of substantially decreased content quantity. Additionally to platforms' own governance, the regulators increasingly engage with social media for content regulation, and in some cases the results of the interventions are not as intended. Ershov and Mitchell 2020 study how content provided by the influencers on Instagram changed in response to the changes in disclosure rules for sponsored content in Germany. They show that after the change in disclosure the amount of sponsored content increased while followers might be worse off.

Our paper adds to these studies providing evidence on the effects of regulation of harmful content on the social media platform. It particularly contributes to the emerging studies discussing and evaluating the consequences of the implementation of NetzDG.

2.3 The Impact of NetzDG

Most previous studies on NetzDG provide descriptive evidence adopting the legal and media perspective (Kasakowskij et al. 2020, Liesching et al. 2021). Several studies rely on the data from the NetzDG transparency reports published by social media platforms (for an overview see Griffin 2021). Kasakowskij et al. 2020 conclude that the vast majority of user reports on Twitter did not lead to deletion or blocking, because most of the content reported by users was, apparently, not unlawful. Also, Liesching et al. 2021 observe a "[..] marginal importance of the Network Enforcement Act in application practice" (translated from Liesching et al. 2021, p. 368). However, the data from transparency reports are not very informative about the causal effect of NetzDG because they only include UGC reported by platform users and do not capture the totality of online hate on the platforms (Griffin 2021). Furthermore, the take down numbers within the transparency reports are likely biased since platforms have an incentive to delete under house rules instead of NetzDG to avoid the legal consequences (Echikson and Knodt 2018). This incentive is mirrored by a survey sent to the platforms, in which the platforms claim that the increased deletion practice documented in the NetzDG transparency reports is due to the platform's house rules and not due to NetzDG (Liesching et al. 2021). Contrary to these studies, our paper uses data directly drawn from one of the largest social media platforms and analyzes the effect of the law on UGC in a quasi-experimental setting. Comparing UGC in the treated and the control groups, we show that there is an additional reduction in online hate due to regulation in the presence of the platform's own governance mechanisms.

For the more formal assessment of the impact of NetzDG, recent studies rely on quasi-experimental and experimental approaches. In a large field experiment, Jiménez Durán 2022 addressed the impact of NetzDG on the likelihood of removing the reported content by Twitter. The author finds that reported tweets are more likely to be deleted (3.5%) while non-reported hateful tweets are only 2.1% likely to be deleted. While Jiménez Durán 2022 suggests no evidence of self-censorship for the users whose tweets were deleted, the decrease in toxicity in our setting is driven by users decreasing hate intensity in their tweets on sensitive topics due to self-censorship.

Building on our paper, Jiménez Durán, Müller, and Schwarz 2022 show results similar to ours for all tweets mentioning asylum seekers over the same time period and combining the data with Müller and Schwarz 2021 suggests that NetzDG also decreased offline violence. Our paper contributes to this literature by presenting a broader picture of the changes in content posted by Twitter users and their engagement with the platform subsequent the NetzDG implementation. We show that even in the presence of the platform's own guidelines for moderation of harmful content, the law can additionally reduce the prevalence of hate speech for the sensitive content, without affecting other topics and user tweeting patterns.

3 Overview of NetzDG

NetzDG[4] 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[5] 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 the user account which submitted it. In practice, Twitter decided to add the option "Covered by Netzwerkdurchsetzungsgesetz" if the user accessed Twitter via a German IP address.[6] 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. Further, the platforms can choose which steps to undertake to address the complaints: they can remove the content in case it is clearly hateful, send a warning to the user account that posted it, or temporarily or permanently block this user account.

Importantly, the law does not require platforms to proactively search and delete hate speech, but only 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 should include information on how to report illegal content and the total numbers of content removal requests by the groups of users (private users or organizations), reaction time, and the reason for reporting.

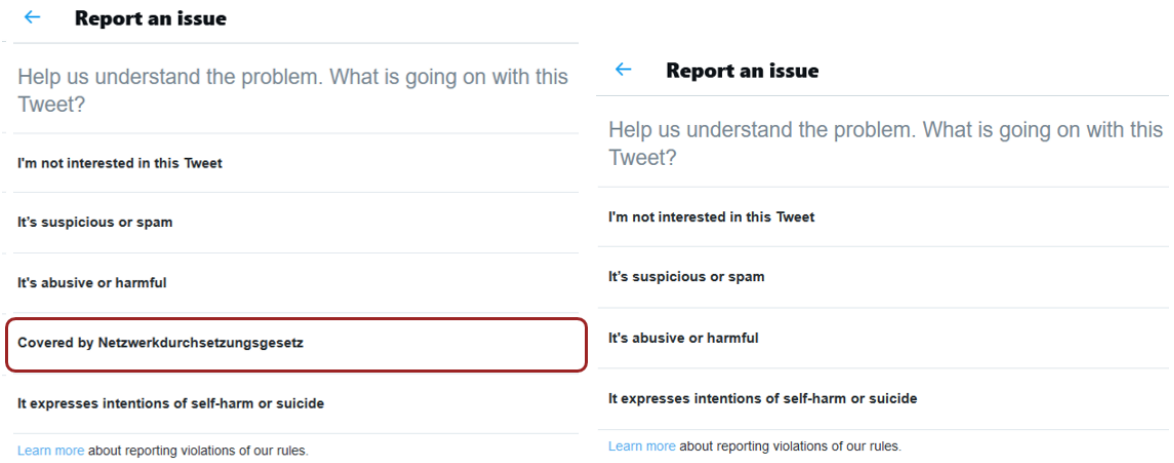
According to §4 of NetzDG, non-compliance can be penalized with a fine of up to five million €. However, due to the risk of content overblocking, the examples for punishable offences only

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

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

[6] 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 1: Menu Options for Reporting Tweets with a German/Austrian IP Address



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

Notes. Screenshots of the menu options when reporting a tweet with a German or an Austrian IP address.

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, social media platforms are deemed non-compliant only if they fail systematically to meet the requirements of the regulation.[7]

4 Data and Empirical Strategy

4.1 Data

We measure the effect of NetzDG on the tweets posted by followers of the right-wing populist party Alternative für Deutschland (AfD), which is represented in the German Bundestag. This segment of the Twittersphere is particularly relevant for our research question because hate speech has a higher prevalence among these users. As studies suggest, individuals with populist views are more likely to use strongly negative rhetoric towards different social groups, e.g. migrants (Halikiopoulou 2018), hence, fuelling hate speech. The xenophobic content generated by right-wing users in Germany directly connects to the incidences of hate crime (Müller and Schwarz 2021). Therefore, the impact of the law on this part of the Twitter community is of particularly high interest.

For our analysis, we manually selected 201 national and regional profiles of the AfD party on Twitter. We further downloaded all followers of those party profiles who are located in Germany based on their profile information.[8] For these followers, we observe all their original

[7] Under 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

[8] In our full sample, 63% of the users provided some information about their location in their profile such that they could cleanly be assigned to the treatment group (being located in Germany) versus the control group (being located outside Germany). The resulting sample is a bit more active and better connected on Twitter compared to those users not indicating any location. Hence, our sample is composed of the majority of the users and measuring the performance of NetzDG is even more important on this subgroup.

tweets between July, 2016 to June, 2019 (1.5 years before and 1.5 years after the introduction of NetzDG) which were still present on the platform at the moment of data collection in May 2020. Throughout the analysis, we only consider original tweets, as opposed to retweets, because we focus on the language that users in our sample choose. Finally, out of 2.3 million retrieved tweets, we selected tweets tackling the topics of migration and religion[9] in messages and hashtags as German-language tweets related to anti-immigrant and anti-muslim topics are the most likely to contain hatred according to the "Political Speech Project" [10].

As a control group, we chose a similar German-language segment of Twitter which was not affected by NetzDG: we manually identified about 30 profiles of the right-wing populist party in Austria, Freiheitliche Partei Österreichs (FPÖ). Out of all followers of this party, we selected all users who are *not* located in Germany and are therefore not treated by the NetzDG. Importantly, the language used by German and Austrian users is similar, as German is the mother tongue in both countries. Although the spoken Austrian German sounds like a dialect to German users, written German is the same in both countries. Furthermore, even if there were slight differences between the language of German and Austrian users, these level differences would be differenced out in our estimation approach. Hence, our control group allows us to analyze the development of hate speech in two comparable segments of Twitter before and after the implementation of the NetzDG.

Our resulting data set comprises more than 160,000 tweets for German and Austrian right-wing followers about sensitive topics, like migration and religion, as our sample for the baseline analysis. Importantly, our sample composition implies that we can measure the effect of the law not for the entire Twittersphere, but rather for an important target group of the law. Due to the negative real-world consequences of hate speech posted by right-wing populists (Caiani and Parenti 2013), the effect of the law on this segment on Twitter is of particularly high interest.

4.2 Outcome Variables

We measure the intensity of hate in tweets using Jigsaw and Google's Perspective API, which employs pre-trained machine learning models to score the probability that short texts are hateful. In the natural language processing literature, Perspective API is considered a benchmark prediction algorithm (Fortuna, Soler, and Wanner 2020). It relies 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).

Since the NetzDG text does not include any measurable definition of hate speech, we use several dimensions of hate speech available in Perspective API for the German language, namely, severely toxic, toxic, threatening, an identity attack, profane, and insulting language. Exploring these different dimensions allows us to learn more about potential channels through which the law

[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/`

might tackle the issue of hate speech. As our outcomes for both treatment and control group are tweets written in the German language, we evaluate hate intensity in both groups using the *same* algorithm. Therefore, potential prediction biases are distributed randomly across tweets in our sample and do *not* affect our results due to our identification strategy.

Perspective API algorithm evaluates the probability scores of each tweet to contain hate for each of the six dimensions in the range $[0, 1]$, so that the probabilities can be interpreted as intensities of hate in tweets. In our analysis, we multiply these scores by 100 to improve the interpretation of the estimation coefficients.

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 1 presents measures describing the profiles of users in our sample. Most of the user characteristics in Table 1 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 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 1: Summary Table of User Characteristics

	N	Mean	Median	SD	Min	Max
No. of Followers	1334	2008.91	236	16337.57	0	534819
No. of Friends	1334	1601.41	487	16668.30	1	590754
Year of account creation	1335	2014.07	2014	3.00	2007	2019
Verified account	1335	0.02	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. of tweets in sample	1337	120.03	9	524.00	1	12279
No. sens. tweets user/month	1337	10.09	2	33.80	1	848
Politician	1337	0.02	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.

Table 2 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 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 sample period, national elections in Germany as well as in Austria took place in the fall of 2017.

Table 2: Summary Table of Tweet Characteristics

	N	Mean	Median	SD	Min	Max
Severe Toxicity	160474	29.98	29	24.13	0	100
Toxicity	160474	42.81	45	22.47	0	100
Threat	160474	34.73	21	24.90	0	100
Identity Attack	160474	57.42	61	27.59	0	100
Profanity	160474	20.48	11	20.47	0	100
Insult	160474	37.24	36	22.06	0	100
No. of Retweets	160474	4.00	0	19.35	0	911
No. of Likes	160474	7.03	0	36.12	0	1711
No. of Replies	160474	1.12	0	5.62	0	292
Video in tweet	160474	0.00	0	0.05	0	1
Photo	160474	0.07	0	0.26	0	1
URL	160474	0.68	1	0.46	0	1
Link to media outlet	160474	0.06	0	0.24	0	1
No. of Words	160474	18.19	15	9.60	1	57
Tweeted at night	160474	0.08	0	0.26	0	1
Terrorist attack in country	160474	0.02	0	0.12	0	1
Election in country	160474	0.01	0	0.10	0	1

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

In Table 3 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 16 (see Appendix) shows the pairwise correlations among the outcome variables, indicating high correlations between toxicity and insults and between severe toxicity and toxicity.

Table 3: Outcome Variables by Country and before/after

	Germany before Mean	Germany after Mean	Austria before Mean	Austria after Mean
Severe Toxicity	33.4	28.7	28.4	28.3
Toxicity	45.4	42.0	40.3	42.8
Threat	37.6	34.8	33.0	31.9
Identity Attack	59.9	57.4	52.8	58.6
Profanity	20.8	20.2	19.1	21.8
Insult	38.1	37.0	34.6	39.2
Observations	47855	49281	33016	30322

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

4.4 Empirical Model

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:

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

We estimate separate 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 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 $After T_t\ Treated_{ij}$, 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). ε_{ijt} indicates the stochastic error term.

5 Results

5.1 The Effect of NetzDG on Hate Intensity

Table 4 reports the results of our baseline specifications.[11] The results suggest that the intensity of hate speech significantly decreases after the introduction of NetzDG in Germany. As our dependent variable is measured on a scale between 0 and 100, 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 tweets related to identity attacks: the probability significantly decreased by 3 pp.

The comparison of the effect sizes to the means of the outcome variables hints at modest effect sizes. The average intensity of an identity attack in all tweets in our sample is 57. At the mean, this would decline by 3 pp and result in an average intensity of 54. In percentage terms, these numbers indicate a reduction in hate intensity 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 changes in hate intensity are approximately 1% - 3% for most of the dependent variables, except for threat intensity which is insignificant throughout our analysis. This can be explained by the fact that threats have already been actionable and illegal before the NetzDG. Moreover, the evaluation of Perspective API in Fortuna, Soler, and Wanner 2020 and also in our manual check (see Figure 6 in the Appendix) suggest that the classifier performs worse at identifying threats.

Table 4: 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.	-1.89***	-2.15***	-0.76	-2.63***	-1.33***	-2.18***
	(0.68)	(0.54)	(0.71)	(0.81)	(0.50)	(0.55)
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	160165	160165	160165	160165	160165	160165
Mean of Outcome	29.97	42.81	34.73	57.42	20.48	37.24
SD of Outcome	24.13	22.46	24.90	27.59	20.47	22.06

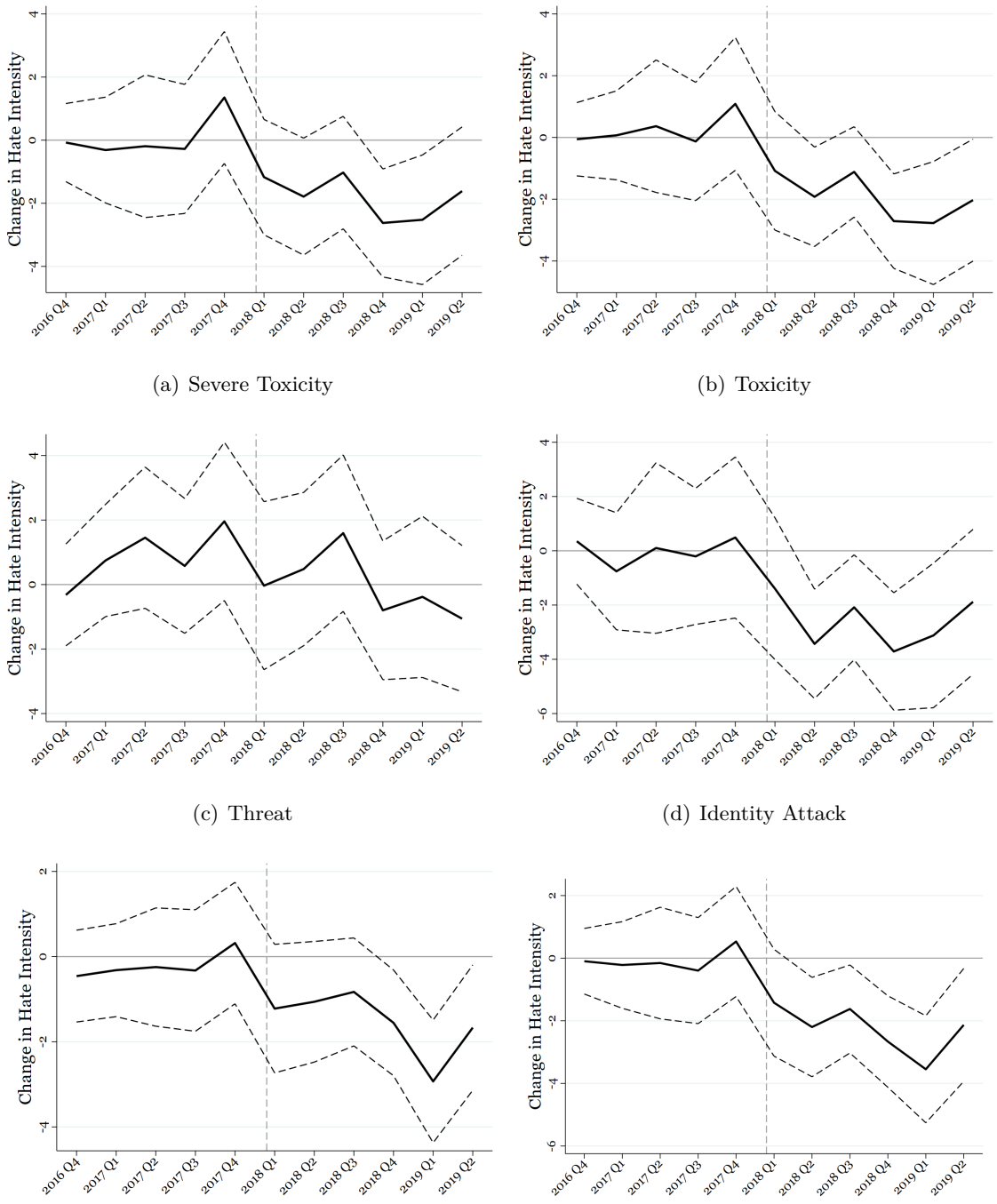
Notes. The table shows the main coefficients of the difference-in-differences estimations comparing the hate intensity in tweets by users affected and unaffected by the law (NetzDG). The columns contain the outcome measures discussed in the data section: Continuous scores ranging from 0 to 100 with regard 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 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 year-month fixed effects, user 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

[11] Table 18 in Appendix presents the full list of control variables with the respective coefficients.

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 28 and Table 19 in the Appendix). Furthermore, omitting the transition period (i.e., six months before the introduction of NetzDG which elapsed between the moment when the law was 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 2) rather as potential outcomes of the treatment effect. However, including these tweet characteristics as controls in a robustness check does also not affect our results and further confirms the robustness of our baseline specification. Lastly, our results are robust to the placebo treatment. If we set the moment of the law implementation to January 2017, the year before the actual implementation, the treatment effect vanishes as expected (see Table 21).

Our results are based on the assumption that the measures of hate 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 2 (similar to BLINDED 2021) presents the results for our six dependent variables of interest, corresponding to Col. (1) - (6) in Table 4. The standard errors of coefficients plotted in the figures correspond to the 90% significance level. The graph shows that the treatment and control groups do not systematically differ *ex ante*, 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 2: Quarterly Treatment Effects with Pre-trends



(a) Severe Toxicity

(b) Toxicity

(c) Threat

(d) Identity Attack

(e) Profanity

(f) Insult

Notes. 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.

5.2 Identification

Our identification strategy allows us to measure the causal effect of the regulation on the intensity of hate in tweets. As the parallel trends assumption suggests, there were no significant differences in the trends of hate before the regulation between the German and the Austrian Twitter segments. Even if there would be differences in the levels of hate, due to, for example, API measuring Austrian language specificities differently from the German language, differencing out the levels allows us to focus on the changes.

We could further be concerned that there is contamination between our treated and control Twitter segments. However, our design compares followers of right-wing parties located in Germany with those who are located outside Germany. Moreover, we can exclude followers who are following both German and Austrian parties and our results are unchanged. Hence, relatively isolated segments without interaction between each other are driving our results. This releases our worry about the potential contamination between the users in the treatment and control groups.

We further address a concern about the potential heterogeneity across Twitter users in our treated and control groups using coarsened exact matching (CEM). Compared to the widely used propensity score matching, CEM does not require assumptions on the model connecting covariates and potential outcomes and helps to control the potential imbalances in the covariates (King and Nielsen 2019). CEM coarsens a set of the observed covariates, and then matches the coarsened data. For matching the Twitter users located in Germany with those located outside Germany, we use the set of covariates that describe the average patterns in the user activity in the period between July and December of 2016, a year before the discussion of the regulation went public (see Table 5).

Based on these covariates, 462 followers from our sample were matched with each other. For these matched followers from the treated and control groups, we again compare our hate intensity measures in tweets before and after the implementation of NetzDG. Our matched sample of tweets contains more than 50 thousand observations. The results in Table 6 again suggest that due to NetzDG, the hate intensity in German tweets decreased by 2-3 pp. Similarly to our baseline specification, severe toxicity decreases by about 2 pp and the intensity of identity attacks decreases by 3 pp. Again, the intensity of threat in tweets is insignificant. Hence, our baseline findings show robustness to any potential differences in the composition of users in treated and control groups.

Table 5: Covariates for Coarsened Exact Matching

Variable	Description
Toxicity	Average level of toxicity across all the tweets that each user posted in the period between July and December of 2016.
Insult	Average level of insult across all the tweets that each user posted in the period between July and December of 2016.
Tweeting frequency	Average monthly number of tweets in the period between July and December of 2016.
Word count	Average word count across all tweets that each user posted in the period between July and December of 2016.
Night	Share of tweets posted in the nigh time between 22pm and 7am.
Video	Share of tweets containing videos.
Retweets count	Average number of retweets per tweet for each user in the period between July and December of 2016.
Likes count	Average number of likes per tweet for each user in the period between July and December of 2016.
Verified	Indicator whether the user's Twitter account is verified. It takes value 1 if the account is verified, and 0 otherwise.
Politician	Indicator whether the owner of the Twitter account is a politician. It takes value 1 if the user is a politician, and 0 otherwise.

5.3 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 hate 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 80 - i.e., it is very likely that the tweet is severely toxic. This threshold has been recommended by Perspective API and supported by computer science research (Mondal, Silva, and Benevenuto 2017, ElSherief et al. 2018, Han and Tsvetkov 2020). Our resulting panel is very unbalanced as very few users tweet frequently about migration and/or religion. Therefore, in the following estimations we 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 7 yield a similar picture to the tweet-level estimations. The coefficients with respect to all of the measures of hate speech are negative but less precisely estimated. The effects are significant for severe toxicity, toxicity, and identity attacks, which are the most discussed measures in the literature addressing automated hate speech detection

Table 6: Baseline Analysis on the Coarsened Exact Matched Sample: 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.	-1.96**	-2.57***	-1.59*	-2.99**	-1.84***	-2.49***
	(0.77)	(0.84)	(0.94)	(1.24)	(0.61)	(0.82)
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	0.09	0.11	0.06	0.13	0.08	0.12
Observations	47768	47768	47768	47768	47768	47768
Mean of Outcome	27.76	42.20	32.62	57.63	20.12	37.44
SD of Outcome	23.16	22.30	23.71	27.59	19.99	21.94

Notes. The table shows the coefficients of interest in the difference-in-differences estimations comparing the hate intensity in tweets by CEM-matched 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 100 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 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 year-month fixed effects, user 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

(ElSherief et al. 2018, Han and Tsvetkov 2020) and are the ones for which the Perspective API demonstrates better performance (Fortuna, Soler, and Wanner 2020). 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 7, the number of identity attacks fell by 11% in Germany due to the introduction of NetzDG. Comparing this effect to the average number of identity attacks by user and month throughout the sample (3) implies that on average, there is one identity attack less per user in three months in Germany.

6 Implications of NetzDG

The Effect Size A decrease of 2-3 pp in our baseline specifications corresponds to a decrease of 5-6% in the mean hate intensity in tweets and 6-10% of a standard deviation. These numbers measure the lower bound of the effect, as our data are drawn ex post and do not include tweets by users who have been banned from the platform due to violating NetzDG. Hence, the regulation achieved a remarkable decrease even in the presence of the working platform governance infrastructure for content moderation which employed automated tools and thousands of human content moderators. Anecdotal evidence suggests that around the year 2019, users of the social network who wanted to escape hate reportedly switched to the German segment of Twitter.[12]

[12]CNBC Article

Table 7: Panel: The Volume of Hateful Tweets by User and Month in Logs

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID Attack	(5) Profanity	(6) Insult
Treated after T.	-0.07**	-0.05*	-0.05	-0.11**	-0.03	-0.03
	(0.03)	(0.03)	(0.04)	(0.05)	(0.02)	(0.02)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.014	0.009	0.033	0.021	0.006	0.006
Observations	9546	9546	9546	9546	9546	9546
Groups	492	492	492	492	492	492
Mean of Outcome	0.743	0.476	1.665	3.380	0.358	0.361
SD of Outcome	2.703	1.812	5.222	8.594	1.448	1.516

Notes. The table shows the coefficients of interest of the panel difference-in-difference estimations at the user-month level. For each user and each month, the number of hateful tweets is the number of tweets with hate intensity $> 80\%$. 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 terrorist attacks. All estimations include a constant and user and year-month fixed effects. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Importantly, we measure this effect before the recent event that changed the face of the platform. In the fall of 2022, after the acquisition of Twitter by Elon Musk, media reported lay-offs of many thousands of content moderators who monitored the prevalence of abusive content and misinformation on the platform.[13] This was an alarming event for the public and the expert community, and while lay-offs affected many countries, the German Twittersphere was deemed one of the most legally protected due to NetzDG. Two months later, Twitter faced a lawsuit in Germany for failing to timely remove illegal content.[14] The recent developments highlight that the regulation is of paramount importance for protecting individuals from offline (psychological) harm caused by online presence.

Content Targeting Our baseline model (Table 4) shows that hate intensity decreased in Germany due to the implementation of NetzDG for tweets that tackle sensitive topics related to migration and religion. However, it is important to understand the broader effect of the law on the content posted by social media users in Germany. To assess this broader effect, we replicate our baseline analysis using *all* the tweets posted by the observed right-wing followers in the period from July 2016 to June 2019.

Table 8 shows that only tweets related to sensitive topics of migration and religion experience a decrease in hate intensity after the regulation was implemented. The effect sizes vary from 1 to 2 pp, which, in the case of severe toxicity, implies a decrease in hate by 11% of the mean. Hence, NetzDG reduces hate by 11% in the topics which are traditionally used as targets for hate. The indicator that a tweet tackles a sensitive topic is positive and strongly significant, suggesting that hate intensity in all dimensions is significantly higher in these "sensitive" tweets. These results confirm that tweets with lower average hate intensity are not affected by NetzDG

[13] Deutsche Welle Article

[14] Euractive Article

suggesting that at the time of our data collection, Twitter carefully moderated content and targeted quite well tweets prone to hate speech without affecting other tweets.

Table 8: OLS with FE; Sample: All Tweets, Interaction with Sensitive Topic

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID attack	(5) Profanity	(6) Insult
Treated after T.	-0.21	-0.19	0.63	-0.00	-0.27	-0.32
	(0.27)	(0.37)	(0.39)	(0.54)	(0.28)	(0.44)
Sensitive topic	10.72***	13.00***	7.01***	28.10***	3.01***	8.73***
	(0.75)	(0.63)	(0.46)	(1.30)	(0.36)	(0.50)
Treated after T.						
× Sensitive topic	-1.87*	-2.08**	-1.47**	-1.90	-0.54	-1.19*
	(0.80)	(0.72)	(0.51)	(1.42)	(0.40)	(0.58)
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	0.129	0.178	0.099	0.249	0.095	0.164
Observations	2270652	2270652	2270652	2270652	2270652	2270652
Mean of Outcome	17.822	27.527	25.704	26.906	16.324	26.563
SD of Outcome	19.841	23.011	18.863	24.356	19.285	22.260

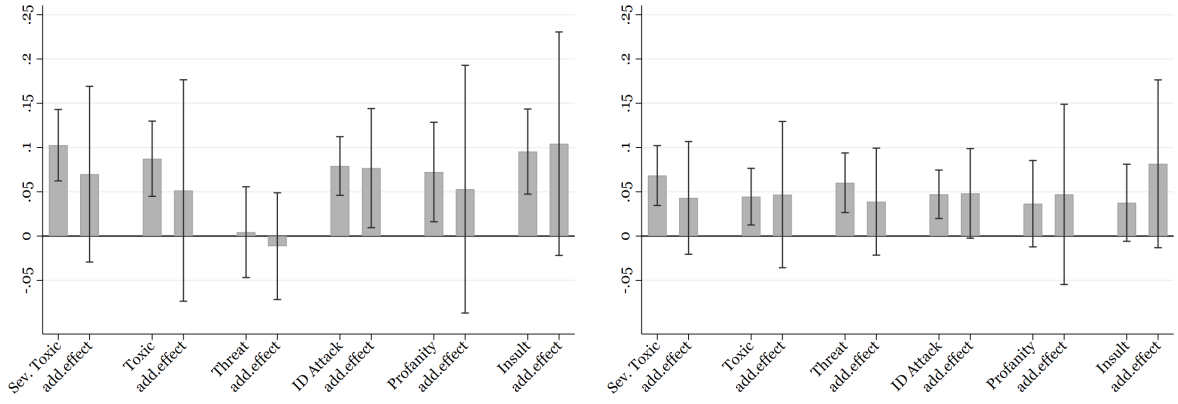
Notes. The table shows the main coefficients of the difference-in-difference estimations comparing the hate intensity in all tweets by users affected and unaffected by the law (NetzDG). The columns contain the different outcome measures discussed in the data section: Continuous scores ranging from 0 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

Spillovers from User Engagement Although NetzDG concerns only the deletion of hateful content, the overall impact from removing hateful tweets might be much larger. This is because user feeds are shaped by the algorithms which select tweets maximizing the potential attractiveness for users, based on impressions and user engagement metrics (i.e. likes, retweets and comments). Due to these algorithms, user engagement with hateful tweets increases the further exposure and subsequent user engagement with these tweets. Therefore, the law may additionally decrease hate speech on Twitter via decreased user engagement.

We examine how user engagement with tweets changes after the introduction of NetzDG. On Twitter, 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 2). As in previous sections, we define a tweet as e.g., an identity attack if the score of identity attacks estimated by the Perspective API exceeds 80. To causally analyze if the user engagement with these 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 the time trends by comparing the user engagement with German and Austrian

hateful tweets and other tweets before and after NetzDG.

Figure 3: Coefficients Plot: Hateful Tweets Receive more User Engagement



(a) Log Likes

(b) Log Retweets

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 coloured 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 3 presents the coefficients of interest for the estimation of 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 22 - 27 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 analyses are shown in Tables 22 and 23 in the Appendix. This analysis shows that hateful tweets receive higher user engagement, collecting significantly more likes (7%-10%) and replies (3%-5%) and are more often retweeted (4%-7%) than the non-hateful ones. This evidence is consistent with Mallipeddi et al. 2021, who show that negative sentiments in tweets are associated with higher user engagement. However, we find no treatment effects on user engagement with hateful tweets. This suggests users do not compensate less hateful tweets by granting stronger promotion for these tweets due to NetzDG.

Since Twitter displays popular tweets on other users' feeds,[15] the significantly higher user engagement with potentially illegal tweets implies that a decrease in the number of hateful original tweets decreases the total exposure to hateful tweets overproportionally. Moreover, Beknazar-Yuzbashev et al. 2022 show in an experimental setting that toxicity is contagious. This implies that when users are exposed to less toxicity, they also reduce their own toxicity in posts and comments. Hence, the actual decrease in hate exposure due to NetzDG is higher than our baseline finding and documents the lower bound of the policy effect.

[15] Twitter Help

User Tweeting Style Our baseline findings suggest that after the implementation of NetzDG, the average hate intensity of tweets as well as the volume of tweets with high hate intensity decreased in the German Twittersphere. However, social network users might have adopted other ways to express hate, while reducing hate in the texts. For example, when the use of severely toxic language is bounded by the law, online users may express hate via hateful images or videos, or by adding links to specific media with biased articles. If social media users adjust to the regulation substituting texts with hate by hateful images or videos, we could expect an increase in the volumes of images and videos after NetzDG.

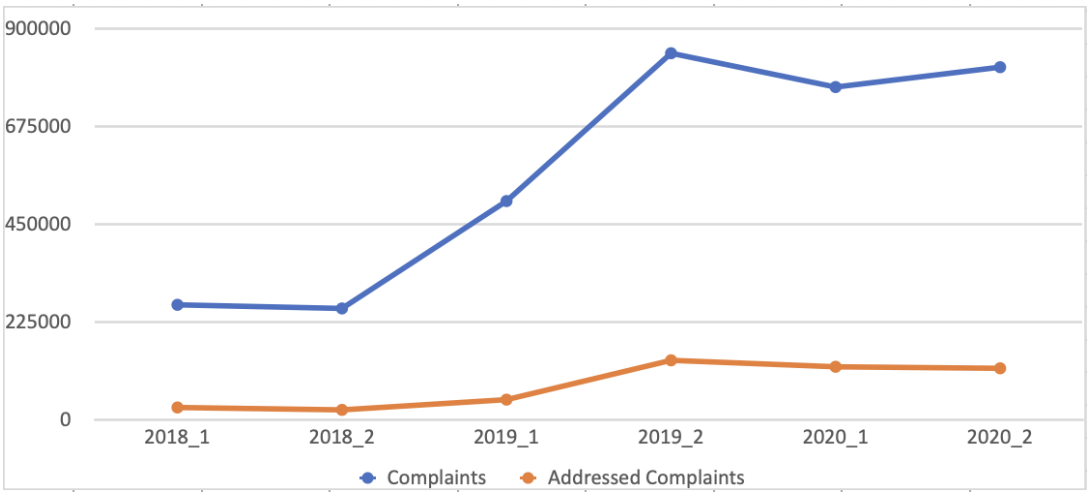
We analyse the effect of the law on the other potential ways of hate expression estimating regressions which are similar to our baseline model with a set of tweet characteristics as dependent variables. Instead of continuous scores ranging from 0 to 100, we use tweeting style measures which are indicators of whether a tweet contains an image, video, any URL or a URL to the media from the top-25 media outlets in Germany. Additionally, we measure the change in the number of hashtags and words in a tweet and in the daily tweeting frequency. Table 9 suggests that, contrary to the hate speech intensity in tweets' texts, the tweeting style among German users did not change as compared to Austrian users. Our data, however, do not allow us to assess the content of images, videos, or links. Acknowledging our data limitations, we do not find any potential substitution patterns in tweeting due to the implementation of NetzDG, which could be overlooked by our hate speech measures.

Table 9: Substitution Patterns: Effect of NetzDG on Tweet Characteristics

	(1) Videos	(2) Images	(3) URL	(4) Media Link	(5) No. Hashtags	(6) No. Words	(7) Tweet. Freq.
Treated after T.	-0.00	0.00	0.03	0.02	0.13	1.20	-0.41
	(0.00)	(0.02)	(0.02)	(0.01)	(0.24)	(1.46)	(2.70)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.25	0.25	0.47	0.14	0.50	0.44	0.71
Observations	160161	160161	160161	160161	160161	160161	12002
Mean of Outcome	0.00	0.07	0.68	0.06	1.02	18.19	13.29
SD of Outcome	0.05	0.26	0.46	0.24	2.05	9.60	34.27

Notes. The table shows the main coefficients of the difference-in-differences estimations comparing tweet characteristics by users affected and unaffected by NetzDG. The columns contain different outcome types: Col (1)-(4) are indicators for a Video, Photos, URL, or Media Link in the tweet. Col. (5) and (6) are counts for the number of hashtags and words. Col. (7) analyzes 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 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 year-month fixed effects, user 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 4: The Number of User Reports and Tweet Deletions due to NetzDG Reported by Twitter



Note: The semiannual numbers of tweets reported by users as hateful according to NetzDG (in the blue colour) and actually removed (in the orange colour) provided in Twitter NetzDG Reports.

7 The Mechanisms

There are three potential mechanisms that can drive the reduction in the average hate intensity scores:

- Following NetzDG, platforms increase the removal of UGC containing hate;
- Platform users decrease their expression of hate to avoid platform interventions;
- Users with preferences for hate expression exit the large platforms subject to NetzDG or multihome, expressing hatred on platforms with weaker moderation.

7.1 Content Removal

To measure the extent of an increased content removal on Twitter, we would need to access tweets taken down in Germany, which is not allowed by the platform. However, we retrieved the figures on user complaints and content removal that are officially provided by Twitter following the clause of NetzDG. Figure 4 suggests that user complaints on hateful tweets due to NetzDG as well as subsequent removal did not increase until 2019, and the numbers of removed tweets were strongly increasing in quarters 3 and 4 of 2019.

At the same time, our baseline results (see Figure 2) suggest a decrease in hate across tweets already in quarters 2 and 4 of 2018, i.e. in the period when, according to the graph, the numbers of complaints and deleted tweets were not growing.Moreover, experimental evidence from the field suggests that in quarter 3 of 2020, when the removal numbers were very high, the platform deleted about 2.1% of hateful tweets expressing Holocaust denial and hate towards disabilities that were not reported to Twitter and 3.5% of hateful tweets that were reported (Jiménez Durán 2022). Due to such low scale, we suggest that content removal is not driving the results in our setting.

In what follows, we examine the extent of the two remaining mechanisms, user self-censorship on Twitter and migration from the platform to other platforms that declare looser content moderation approaches.

7.2 Self-Censorship

To better assess the likelihood of users self-censoring themselves in the expression of hate, we look at changes in the distribution of hate at the user level. Specifically, we run a similar analysis to the volume of hate at the user-month level, now focusing on the parameters of the distribution of hate intensity. Our series of regressions use as dependent variables the minimum, the median, and the maximum values of hate intensity for each hate measure at the user-month level. The results in Table 10 suggest that while there is no change in the monthly minimum values, there is a strong significant shift to the left in the median values of hate intensity of each user measured by severe toxicity, toxicity, identity attacks, and insults.

The maximum values also decrease for toxicity, identity attacks, and insult, but these effects are marginally significant. Additionally, when we consider the entire dataset with all tweets of our users, the hate intensity shifts to the right, with increases in the median and mean for the probability of identity attacks (these results are available upon request). This shows that while the entire German segment follows the general trend on social networks towards an increase in mutual hate, the sensitive topics experience very significant and robust decrease in hate, i.e. "adjustment" in the language about these topics. *Importantly, if the shifts in the median of the hate intensity would have been due to tweets deletion by the platform, we would expect to measure stronger and more significant decreases in the maximum values, because the platform would logically focus on moderating tweets with the highest values of hate intensity.*

Table 10: Panel: The Changes in Hate Intentsity Distribution by User and Month

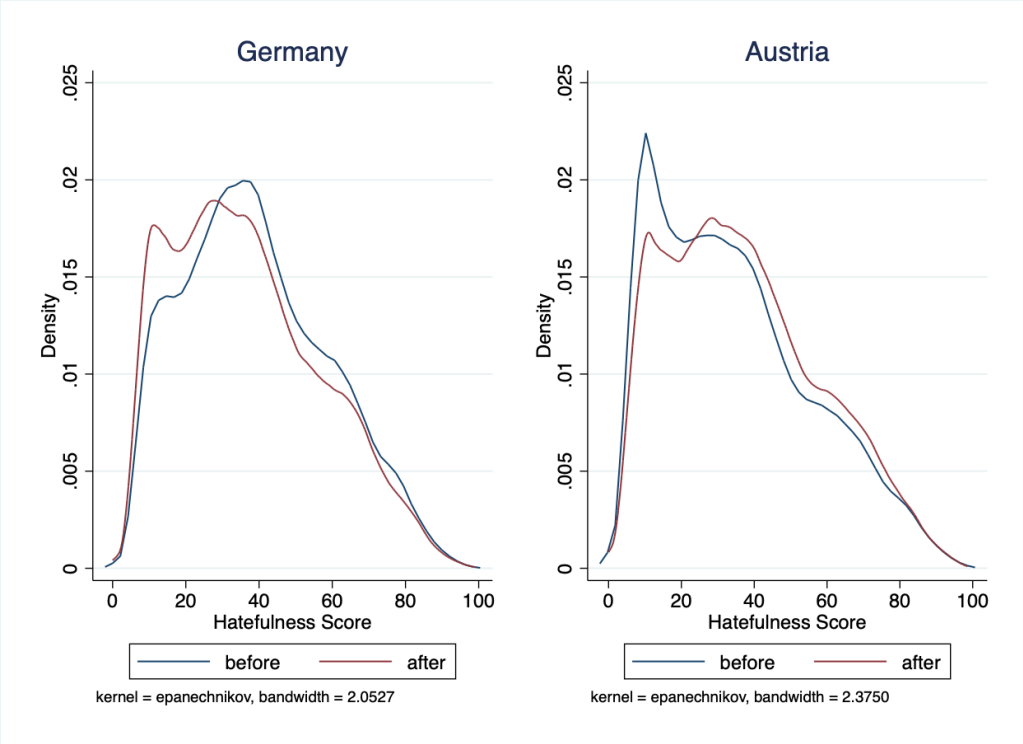
The Treatment Effect						
	(1)	(2)	(3)	(4)	(5)	(6)
	sev. Toxicity	Toxicity	Threat	ID Attack	Profanity	Insult
Min. Hate Score	0.46	0.13	0.60	0.38	-0.00	0.15
	(0.79)	(0.99)	(0.65)	(1.28)	(0.00)	(0.89)
Median Hate Score	-1.25*	-2.22***	-0.41	-2.48**	-0.00	-2.04***
	(0.74)	(0.74)	(0.75)	(1.01)	(0.00)	(0.70)
Max. Hate Score	-2.06	-1.96*	-3.13**	-2.05*	-0.02	-2.15*
	(1.32)	(1.03)	(1.52)	(1.15)	(0.02)	(1.10)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9546	9546	9546	9546	9546	9546
Groups	492	492	492	492	492	492

Notes. The table shows the coefficients of interest of the panel difference-in-difference estimations at the user-month level. The dependent variables (in rows) are the minimum, median, and maximum values of the corresponding hate intensity measures (in columns) computed at the user-month level. 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 terrorist attacks. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We additionally compare 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 5 display the average of the mean hate score of each tweet. Figure 5 shows that the distribution of the hate score shifts towards higher scores in the middle part of the distribution in Austria after January 2018 and towards lower scores in Germany. Again, if the decreases in hate we measure were driven by the platform removing content, we would expect to observe a stronger change in the right tail of the distribution in Germany than in the middle part of the distribution. Hence, the treatment effect seems to be driven by users moderating their language in tweets about sensitive topics.

Figure 5: Distribution of Average Hate Intensity by Time Period and Treatment Status



Notes. These plots display the distribution of the hate intensity scores of each tweet by untreated (Austria) and treated (Germany) groups of users before and after NetzDG. Observations range from 1.5 years before to 1.5 years after NetzDG. The hate intensity scores are calculated as the averages of the scores of the six hate dimensions for each tweet.

7.3 User Migration

Anecdotal evidence suggests that users with strong preferences for uploading and viewing hateful content migrate to platforms with weaker or no content moderation in response to the efforts of large social media platforms such as Twitter and Facebook to moderate 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. 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

[16] Spiegel Article

and media campaigns by Facebook and Twitter promising to increase their moderation efforts.[17]

These migration patterns might suggest relocation rather than mitigation of hate speech, although Rauchfleisch and Kaiser 2021 and Ali et al. 2021 suggest that deplatforming is still an effective tool to combat online hate speech due to the lower reach of hatred on smaller platforms. Moreover, as soon as smaller platforms grow in the number of users due to migration from the dominant platforms and reach the cutoff of 2 million users, they become subject to NetzDG and are forced to either moderate hate speech or leave the market.

We address the user migration by analyzing the user composition in our sample. Table 11 shows the share and the number of users in our sample i) who tweeted only before NetzDG was introduced, ii) only after NetzDG was introduced, and iii) who tweeted both before and after the introduction of NetzDG. About half of the users in our estimation sample (i.e., who tweeted about sensitive topics) was present on Twitter both before and after the introduction of NetzDG.

Table 11: User Composition

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

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

Consistently with the general growth path of social media platforms (Hölig and Hasebrink 2020), more users joined than left our sample. This pattern is stronger in Germany than in Austria, and, surprisingly, the share of users leaving the sample is lower in Germany than in Austria.

Additionally, we replicate our baseline analysis using only tweets from users who tweeted in our sample at least twice before and after NetzDG. The results in Table 28 (in Appendix) are confirming the baseline results and the effects are measured more precisely. Hence, our baseline results are driven by the users who continue tweeting on the platform after the implementation of NetzDG.

The analysis in this section suggests that the main driver of the decrease in hate due to NetzDG in our setting is *self-censoring in the expression of hatred*. The effect of the regulation on the behaviour of the social network users is consistent with the findings of Huang, Hong, and Burtch 2016 that the integration of a social network into the review platform changes the volume and quality of UGC via shifts in user behaviour rather than in user composition.

[17] Politico Article

8 Conclusion

In the past few years, social media platforms reportedly made numerous efforts to design complex infrastructure and mechanisms for combating harmful content. However, the public concern remained that large tech companies were not doing enough to remove hate speech as only a very small share of user complaints was addressed. Hence, German policy makers additionally imposed large platforms "to take on their responsibility in [the] question of deleting criminal content" (Heiko Maas, federal minister for justice and consumer protection)[18] and passed the law obliging large social media platforms to timely remove user-reported hateful content. Among critiques of the law, politicians and civil society mention the threats of policing digital communication and restricting the freedom of speech. At the same time, the opponents of the law claimed that platforms will not comply with the law due to the lack of clarity and this law only would increase legal uncertainty. Our paper contributes to this discussion showing that after the implementation of NetzDG the intensity of hate in German tweets decreased. We measure the causal effect of NetzDG using a quasi-experimental setting in which we compare the content generated by right-wing sympathizers in the German Twittersphere compared to the Austrian Twittersphere.

Although Twitter claims that the NetzDG did not affect content moderation as most of hate speech is removed due to its internal governance policy (Liesching et al. 2021), we find that legal regulation can contribute to restraining harmful content even when platforms already have governance rules for the same purpose. While the platform's governance rules apply to tweets by users located in both Austria and Germany, our results suggest an additional reduction in hate speech by users located in Germany. We find robust effects of the regulation on decreasing the intensity of (severe) toxicity, profanity, insults, and identity attacks in Germany as opposed to Austria by about 6-11% of the mean. Moreover, the volume of potentially unlawful tweets decreased by 11% in the number of original hateful tweets. Additionally, we find that hateful tweets generally receive higher user engagement. Hence, the reduction in the number of original hateful tweets decreases the exposure to hate even more due to prevented impressions and user engagement with hateful content.

Our analysis uncovers the underlying mechanisms of the law's effect. We show that the treatment effect is only present in tweets on sensitive topics such as religion and migration with higher average scores of hate intensity. Tweets on topics with lower average hate scores and other tweeting style characteristics such as the number of words or uploaded images are not affected by the law. This suggests that NetzDG is successful in targeting relevant topics without significantly affecting not targeted content. Moreover, we address the three potential mechanisms driving the decrease in hate intensity due to NetzDG. Although data limitations do not allow us to rule out that platforms delete more hateful UGC, we show that our results are mostly driven by the self-censorship of platform users who limit themselves in their expression of hatred subsequent to NetzDG.

The implementation of NetzDG inspired many countries to design similar national laws and,

[18] https://www.politico.eu/article/germany-unveils-law-with-big-fines-for-hate-speech-on-social-media/

later, NetzDG became a model for the EU-wide regulation Digital Services Act which comes into force in 2024. Hence, our findings are of high relevance for the policy makers, as they inform about potential outcomes and mechanisms through which the laws of similar design could affect the prevalence of illegal content on social media platforms in other countries.

Future research should assess the long term effects of counter-hate legislation on social media and the content censored by the platform to better evaluate the effect of regulation on the strategic incentives of social media platforms. Additionally, researchers could examine a number of issues regarding the optimal design of the regulation as emerging studies hint that platforms can respond strategically to obligations provided by external rules.

References

Ali, Shiza, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini (2021). "Understanding the Effect of Deplatforming on Social Networks". In: *13th ACM Web Science Conference 2021*, pp. 187–195.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow (2020). "The welfare effects of social media". In: *American Economic Review* 110.3, pp. 629–76.

Allcott, Hunt and Matthew Gentzkow (2017). "Social media and fake news in the 2016 election". In: *Journal of economic perspectives* 31.2, pp. 211–236.

Bail, Christopher A, Brian Guay, Emily Maloney, Aidan Combs, D Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky (2020). "Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017". In: *Proceedings of the national academy of sciences* 117.1, pp. 243–250.

Beknazar-Yuzbashev, George, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski (2022). "Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment". In: *Available at SSRN*.

BLINDED (2021). "BLINDED". In: *ICIS Proceedings* 11.

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.

Borwankar, Sameer, Jinyang Zheng, and Karthik Natarajan Kannan (2022). "Democratization of Misinformation Monitoring: The Impact of Twitter's Birdwatch Program". In: *Available at SSRN 4236756*.

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin (2022). "Social media and mental health". In: *American Economic Review* 112.11, pp. 3660–3693.

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.

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.

Echikson, William and Olivia Knodt (2018). "Germany's NetzDG: A key test for combatting online hate". In: *CEPS Policy Insight*.

ElSherief, Mai, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding (2018). "Hate lingo: A target-based linguistic analysis of hate speech in social media". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 1.

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.

Ershov, Daniel, Juan S Morales, et al. (2021). *Sharing news left and right: The effects of policies targeting misinformation on social media*. Tech. rep. Collegio Carlo Alberto.

Feher, Adam (2023). "How to enforce platforms' liability?" In: *Working paper*.

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*.

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera (2021). "Social media, sentiment and public opinions: Evidence from# Brexit and# USElection". In: *European Economic Review* 136, p. 103772.

Griffin, Rachel (2021). "New School Speech Regulation and Online Hate Speech: A Case Study of Germany's NetzDG". In: *Available at SSRN: 3920386*.

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.

Han, Xiaochuang and Yulia Tsvetkov (2020). "Fortifying toxic speech detectors against veiled toxicity". In: *arXiv preprint arXiv:2010.03154*.

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

Huang, Ni, Yili Hong, and Gordon Burtch (2016). "Social network integration and user content generation: Evidence from natural experiments". In: *MIS Quarterly, Fox School of Business Research Paper* 17-001.

Jiménez Durán, Rafael (2022). "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter". In: *Available at SSRN: 4044098*.

Jiménez Durán, Rafael, Karsten Müller, and Carlo Schwarz (2022). "The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG". In: *Available at SSRN 4230296*.

Kasakowskij, Thomas, Julia Fürst, Jan Fischer, and Kaja J Fietkiewicz (2020). "Network enforcement as denunciation endorsement? A critical study on legal enforcement in social media". In: *Telematics and Informatics* 46, p. 101317.

King, Gary and Richard Nielsen (2019). "Why propensity scores should not be used for matching". In: *Political analysis* 27.4, pp. 435–454.

Liesching, Marc, Chantal Funke, Alexander Hermann, Christin Kneschke, Carolin Michnic, Linh Nguyen, Johanna Prüßner, Sarah Rudolph, and Vivien Zschammer (2021). *Das NetzDG in der praktischen Anwendung: Eine Teilevaluation des Netzwerkdurchsetzungsgesetzes*. Carl Grossmann Verlag.

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

Mallipeddi, Rakesh, Ramkumar Janakiraman, Subodha Kumar, and Seema Gupta (2021). "The effects of social media content created by human brands on engagement: Evidence from indian general election 2014". In: *Information Systems Research* 32.1, pp. 212–237.

Mondal, Mainack, Leandro Araújo Silva, and Fabriécio Benevenuto (2017). "A measurement study of hate speech in social media". In: pp. 85–94.

Morales, Juan S (2020). "Perceived popularity and online political dissent: Evidence from Twitter in Venezuela". In: *The International Journal of Press/Politics* 25.1, pp. 5–27.

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

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

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.

Rauchfleisch, Adrian and Jonas Kaiser (2021). "Deplatforming the far-right: An analysis of YouTube and BitChute". In: *Available at SSRN: 3867818*.

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.

Tworek, Heidy and Paddy Leerssen (2019). "An analysis of Germany's NetzDG law". In: *Translantic Working Group*.

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.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018). "The spread of true and false news online". In: *science* 359.6380, pp. 1146–1151.

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov (2020). "Political effects of the internet and social media". In: *Annual review of economics* 12, pp. 415–438.

Table 12: 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	58.09524	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	81.43812	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.
Threat	65.58015	Describes an intention to inflict pain, injury, or violence against an individual or group.
Identity Attack	91.92697	Negative or hateful comments targeting someone because of their identity.
Profanity	32.70008	Swear words, curse words, or other obscene or profane language.
Insult	65.19685	Insulting, inflammatory, or negative comment towards a person or a group of people.

Notes. 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`

Appendices

A Further Information on Data

Table 13: 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 14: 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

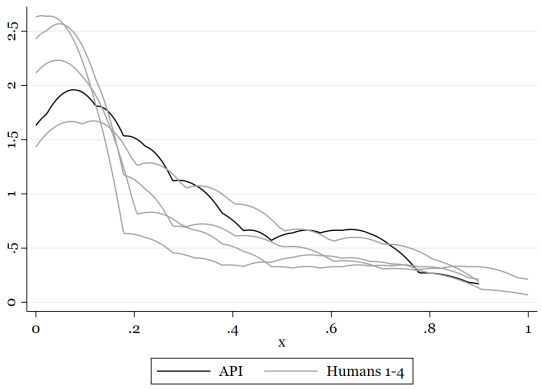
Table 15: Summary Table of Tweet Characteristics

	Total					Germany					Austria				
	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
Severe Toxicity	29.98	29	24.13	0.00	100.00	31.05	29	23.69	0.00	100.00	28.32	18	24.70	0.00	100.00
Toxicity	42.81	45	22.47	0.00	100.00	43.68	46	21.74	0.00	100.00	41.47	42	23.47	0.00	100.00
Threat	34.73	21	24.90	0.00	99.93	36.21	22	25.45	0.00	99.93	32.47	20	23.86	0.00	99.46
Identity Attack	57.42	61	27.59	0.00	100.00	58.61	63	26.58	0.00	100.00	55.58	60	28.98	0.00	100.00
Profanity	20.48	11	20.47	0.00	100.00	20.51	11	20.12	0.00	100.00	20.42	11	20.99	0.00	99.95
Insult	37.24	36	22.06	0.00	99.72	37.54	36	21.42	0.00	99.72	36.78	36	23.01	0.00	99.72
No. of Retweets	4.00	0	19.35	0.00	911.00	4.89	0	22.17	0.00	911.00	2.62	0	13.84	0.00	779.00
No. of Likes	7.03	0	36.12	0.00	1711.00	8.28	0	41.28	0.00	1398.00	5.12	0	26.19	0.00	1711.00
No. of Replies	1.12	0	5.62	0.00	292.00	1.28	0	6.71	0.00	292.00	0.86	0	3.30	0.00	275.00
Video in tweet	0.00	0	0.05	0.00	1.00	0.00	0	0.04	0.00	1.00	0.00	0	0.06	0.00	1.00
Photo	0.07	0	0.26	0.00	1.00	0.08	0	0.28	0.00	1.00	0.05	0	0.23	0.00	1.00
URL	0.68	1	0.46	0.00	1.00	0.74	1	0.44	0.00	1.00	0.59	1	0.49	0.00	1.00
Link to media outlet	0.06	0	0.24	0.00	1.00	0.08	0	0.27	0.00	1.00	0.03	0	0.18	0.00	1.00
No. of Words	18.19	15	9.60	1.00	57.00	18.33	15	9.63	1.00	57.00	17.96	15	9.56	1.00	52.00
Tweeted at night	0.08	0	0.26	0.00	1.00	0.08	0	0.28	0.00	1.00	0.06	0	0.24	0.00	1.00
Terrorist attack in country	0.02	0	0.12	0.00	1.00	0.02	0	0.15	0.00	1.00	0.00	0	0.06	0.00	1.00
Election in country	0.01	0	0.10	0.00	1.00	0.01	0	0.11	0.00	1.00	0.01	0	0.08	0.00	1.00

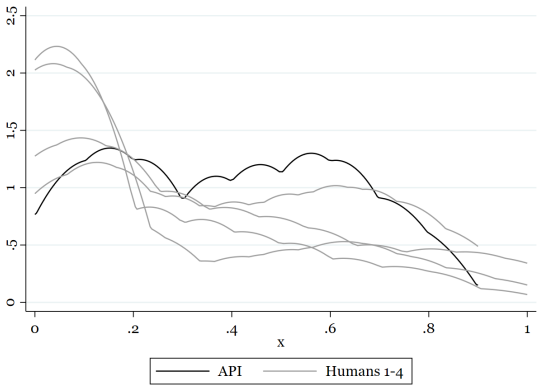
Table 16: Raw Correlation among Outcome Variables

	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					

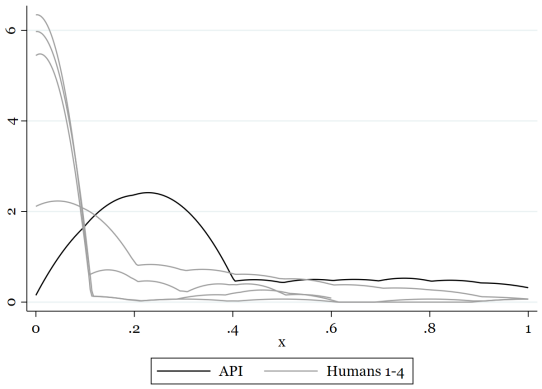
Figure 6: Distributions of the Hate Scores by Perspective API and four Human Classifiers



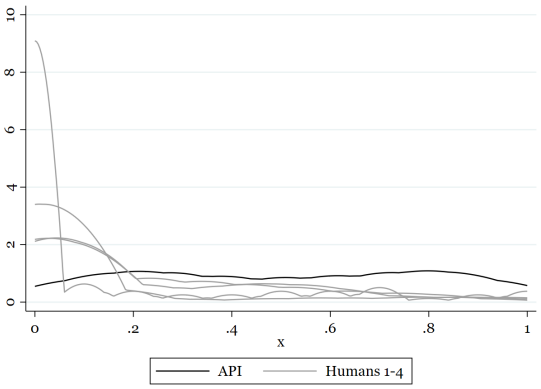
(a) Severe Toxicity



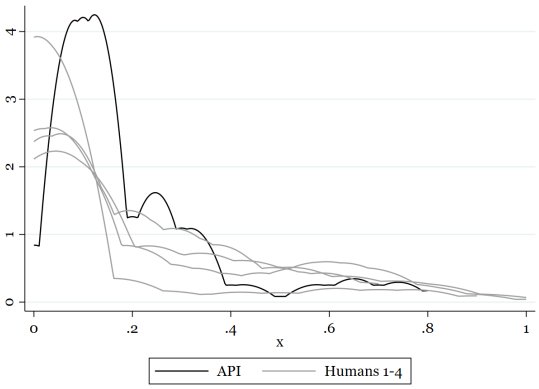
(b) Toxicity



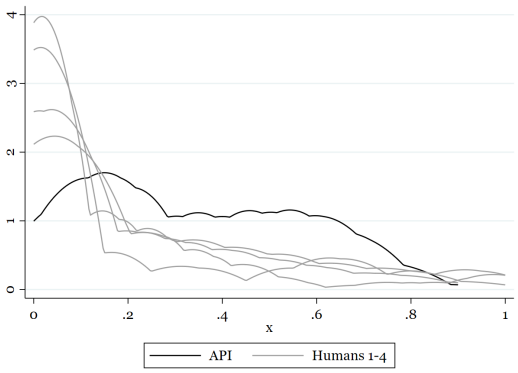
(c) Threat



(d) Identity Attack



(e) Profanity



(f) Insult

Notes. The plot shows the scores with regard to the six hate dimensions as estimated by the algorithm (Perspective API) and four human classifiers.

B Baseline and Robustness Checks

Table 17: OLS

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID Attack	(5) Profanity	(6) Insult
Germany	3.23	3.04*	3.82***	3.65*	0.51	1.49
	(1.99)	(1.56)	(1.39)	(1.99)	(1.32)	(1.57)
Treated after T.=1	-2.18*	-3.26***	-0.87	-4.69***	-1.70*	-3.14***
	(1.31)	(1.12)	(1.00)	(1.59)	(0.91)	(1.17)
Tweeted at night	2.75***	2.83***	0.31	3.28***	2.44***	3.11***
	(0.88)	(0.83)	(0.56)	(1.15)	(0.76)	(0.94)
verified=1	-5.26***	-4.45***	-3.19***	-3.54**	-4.60***	-4.56***
	(1.47)	(1.36)	(0.82)	(1.79)	(1.08)	(1.52)
No. of 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	-0.20	-0.02	-0.16	-0.06	-0.20	0.06
	(0.23)	(0.22)	(0.25)	(0.29)	(0.21)	(0.23)
Wednesday	-0.68***	-0.49**	-0.77***	-0.77***	-0.51**	-0.38
	(0.23)	(0.23)	(0.27)	(0.28)	(0.21)	(0.24)
Thursday	-0.53**	-0.53**	-0.85***	-0.46*	-0.41*	-0.26
	(0.26)	(0.23)	(0.27)	(0.28)	(0.22)	(0.24)
Friday	0.08	0.17	0.20	-0.42	-0.02	0.12
	(0.27)	(0.27)	(0.23)	(0.32)	(0.23)	(0.25)
Saturday	0.23	0.62**	-0.49	0.59	0.24	0.72**
	(0.37)	(0.31)	(0.34)	(0.36)	(0.31)	(0.32)
Sunday	0.13	0.31	-0.39	0.47	0.08	0.29
	(0.27)	(0.24)	(0.32)	(0.31)	(0.22)	(0.25)
Terrorist attack in country	0.23	0.28	1.37**	-0.25	0.01	-0.07
	(0.57)	(0.54)	(0.69)	(0.71)	(0.42)	(0.49)
Election in country	-0.45	-1.01	-0.40	-1.02	-0.72	-0.64
	(0.67)	(0.67)	(0.76)	(0.81)	(0.57)	(0.64)
Constant	38.70***	49.59***	45.28***	60.64***	26.12***	41.80***
	(3.46)	(3.02)	(2.51)	(4.37)	(2.59)	(3.02)
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	160407	160407	160407	160407	160407	160407
Mean of Outcome	29.97	42.81	34.73	57.41	20.48	37.24
SD of Outcome	24.13	22.47	24.90	27.59	20.47	22.06

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 18: 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.	-1.89***	-2.15***	-0.76	-2.63***	-1.33***	-2.18***
	(0.68)	(0.54)	(0.71)	(0.81)	(0.50)	(0.55)
Tweeted at night	0.72	0.71	0.13	0.08	0.87	0.74
	(0.66)	(0.61)	(0.40)	(0.75)	(0.59)	(0.63)
Tuesday	-0.31	-0.13	-0.24	-0.22	-0.32*	-0.06
	(0.20)	(0.19)	(0.24)	(0.24)	(0.18)	(0.19)
Wednesday	-0.55**	-0.39*	-0.69***	-0.72***	-0.42**	-0.32
	(0.22)	(0.22)	(0.26)	(0.25)	(0.20)	(0.21)
Thursday	-0.53**	-0.54**	-0.81***	-0.56**	-0.41**	-0.29
	(0.22)	(0.21)	(0.25)	(0.24)	(0.20)	(0.22)
Friday	-0.08	0.07	0.09	-0.48*	-0.11	0.03
	(0.25)	(0.25)	(0.23)	(0.29)	(0.22)	(0.22)
Saturday	0.11	0.41	-0.38	0.36	0.12	0.50*
	(0.31)	(0.28)	(0.31)	(0.33)	(0.28)	(0.29)
Sunday	0.03	0.10	-0.21	0.21	-0.07	0.01
	(0.23)	(0.21)	(0.29)	(0.26)	(0.19)	(0.21)
Terrorist attack in country	0.25	0.28	1.36*	-0.14	-0.10	-0.18
	(0.50)	(0.51)	(0.70)	(0.65)	(0.40)	(0.47)
Election in country	-0.32	-0.68	-0.55	-0.33	-0.59	-0.27
	(0.56)	(0.55)	(0.71)	(0.63)	(0.50)	(0.53)
Constant	30.70***	43.51***	35.27***	58.45***	21.01***	37.90***
	(0.23)	(0.21)	(0.28)	(0.29)	(0.21)	(0.20)
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	160165	160165	160165	160165	160165	160165
Mean of Outcome	29.973	42.810	34.735	57.419	20.476	37.242
SD of Outcome	24.126	22.463	24.900	27.590	20.468	22.059

Clustered standard errors in parentheses

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

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.	-2.14***	-2.45***	-0.54	-3.27***	-1.25**	-2.36***
	(0.75)	(0.58)	(0.72)	(0.87)	(0.53)	(0.58)
Constant	29.80***	42.81***	34.77***	57.55***	20.14***	37.15***
	(0.29)	(0.22)	(0.25)	(0.32)	(0.18)	(0.20)
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	29.10	42.00	34.60	56.39	19.75	36.37
SD of Outcome	23.52	22.24	24.86	27.56	19.84	21.71

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.	-1.72**	-2.11***	-0.51	-2.76***	-1.27**	-2.20***
	(0.76)	(0.57)	(0.81)	(0.82)	(0.55)	(0.58)
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	135883	135883	135883	135883	135883	135883
Mean of Outcome	29.69	42.64	34.63	57.28	20.41	37.18
SD of Outcome	23.99	22.41	24.84	27.56	20.45	22.02

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.	-0.80	-0.75	0.75	-1.53	-0.60	-1.09
	(0.77)	(0.73)	(0.95)	(0.97)	(0.56)	(0.71)
Constant	30.31***	43.13***	34.34***	58.17***	20.71***	37.73***
	(0.37)	(0.35)	(0.46)	(0.45)	(0.26)	(0.33)
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	29.97	42.81	34.73	57.42	20.48	37.24
SD of Outcome	24.13	22.46	24.90	27.59	20.47	22.06

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

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

C User Engagement

Table 22: Cross Section with FE: User Engagement with Potentially Unlawful Tweets - Log Retweets

	(1) sev. Toxicity	(2) Toxicity	(3) Threat
Germany			
× AfterT	0.09	0.09	0.09
	(0.06)	(0.06)	(0.05)
Severely toxic	0.07***		
	(0.02)		
Germany			
× Severely toxic	-0.03		
	(0.02)		
AfterT			
× Severely toxic	-0.03		
	(0.02)		
Germany			
× AfterT			
× Severely toxic	0.04		
	(0.03)		
Toxic		0.04***	
		(0.02)	
Germany			
× Toxic		-0.02	
		(0.02)	
AfterT			
× Toxic		-0.05*	
		(0.03)	
Germany			
× AfterT			
× Toxic		0.05	
		(0.04)	
Threat			0.06***
			(0.02)
Germany			
× Threat			-0.01
			(0.02)
AfterT			
× Threat			0.03
			(0.02)
Germany			
× AfterT			
× Threat			0.04
			(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	160161	160161	160161
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 23: Cross Section with FE: User Engagement with Potentially Unlawful Tweets - Log Retweets

	(1) ID Attack	(2) Profanity	(3) Insult
Germany			
× AfterT	0.08	0.09	0.09
	(0.06)	(0.06)	(0.06)
ID Attack	0.05***		
	(0.01)		
Germany			
× ID Attack	-0.02		
	(0.02)		
AfterT			
× ID Attack	-0.01		
	(0.02)		
Germany			
× AfterT			
× ID Attack	0.05*		
	(0.03)		
Profanity		0.04	
		(0.02)	
Germany			
× Profanity		-0.02	
		(0.03)	
AfterT			
× Profanity		-0.03	
		(0.04)	
Germany			
× AfterT			
× Profanity		0.05	
		(0.05)	
Insult			0.04*
			(0.02)
Germany			
× Insult			-0.03
			(0.03)
AfterT			
× Insult			-0.05*
			(0.03)
Germany			
× AfterT			
× Insult			0.08*
			(0.05)
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	160161	160161	160161
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 24: Cross Section with FE: User Engagement with Potentially Unlawful Tweets - Log Likes

	(1) sev. Toxicity	(2) Toxicity	(3) Threat
Germany			
× AfterT	0.00	0.00	0.01
	(0.07)	(0.07)	(0.07)
Severely toxic	0.10***		
	(0.02)		
Germany			
× Severely toxic	-0.03		
	(0.03)		
AfterT			
× Severely toxic	-0.07**		
	(0.04)		
Germany			
× AfterT			
× Severely toxic	0.07		
	(0.05)		
Toxic		0.09***	
		(0.02)	
Germany			
× Toxic		-0.01	
		(0.03)	
AfterT			
× Toxic		-0.07	
		(0.04)	
Germany			
× AfterT			
× Toxic		0.05	
		(0.06)	
Threat			0.00
			(0.03)
Germany			
× Threat			0.01
			(0.03)
AfterT			
× Threat			0.04
			(0.03)
Germany			
× AfterT			
× Threat			-0.01
			(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	160161	160161	160161
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 25: Cross Section with FE: User Engagement with Potentially Unlawful Tweets - Log Likes

	(1) ID Attack	(2) Profanity	(3) Insult
Germany			
× AfterT	-0.01	0.01	0.00
	(0.07)	(0.07)	(0.07)
ID Attack	0.08***		
	(0.02)		
Germany			
× ID Attack	-0.02		
	(0.02)		
AfterT			
× ID Attack	-0.01		
	(0.03)		
Germany			
× AfterT			
× ID Attack	0.08**		
	(0.03)		
Profanity		0.07**	
		(0.03)	
Germany			
× Profanity		0.00	
		(0.04)	
AfterT			
× Profanity		-0.02	
		(0.05)	
Germany			
× AfterT			
× Profanity		0.05	
		(0.07)	
Insult			0.10***
			(0.02)
Germany			
× Insult			-0.01
			(0.03)
AfterT			
× Insult			-0.08*
			(0.04)
Germany			
× AfterT			
× Insult			0.10
			(0.06)
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	160161	160161	160161
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 26: Cross Section with FE: User Engagement with Potentially Unlawful Tweets - Log Replies

	(1) sev. Toxicity	(2) Toxicity	(3) Threat
Germany			
× AfterT	-0.02	-0.02	-0.02
	(0.05)	(0.05)	(0.05)
Severely toxic	0.03**		
	(0.02)		
Germany			
× Severely toxic	0.00		
	(0.02)		
AfterT			
× Severely toxic	-0.02		
	(0.02)		
Germany			
× AfterT			
× Severely toxic	-0.02		
	(0.03)		
Toxic		0.05***	
		(0.01)	
Germany			
× Toxic		-0.02	
		(0.02)	
AfterT			
× Toxic		-0.05**	
		(0.02)	
Germany			
× AfterT			
× Toxic		0.02	
		(0.04)	
Threat			0.01
			(0.01)
Germany			
× Threat			0.00
			(0.01)
AfterT			
× Threat			0.01
			(0.01)
Germany			
× AfterT			
× Threat			-0.01
			(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	160161	160161	160161
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 27: Cross Section with FE: User Engagement with Potentially Unlawful Tweets - Log Replies

	(1) ID attack	(2) Profanity	(3) Insult
Germany			
× AfterT	-0.03	-0.02	-0.02
	(0.05)	(0.05)	(0.04)
ID Attack	0.05***		
	(0.02)		
Germany			
× ID Attack	-0.03		
	(0.02)		
AfterT			
× ID Attack	-0.03		
	(0.02)		
Germany			
× AfterT			
× ID Attack	0.03		
	(0.02)		
Profanity		0.03	
		(0.02)	
Germany			
× Profanity		0.01	
		(0.02)	
AfterT			
× Profanity		-0.00	
		(0.03)	
Germany			
× AfterT			
× Profanity		-0.01	
		(0.04)	
Insult			0.04**
			(0.02)
Germany			
× Insult			-0.01
			(0.03)
AfterT			
× Insult			-0.03
			(0.03)
Germany			
× AfterT			
× Insult			0.02
			(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	0.43	0.43	0.43
Observations	160161	160161	160161
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 28: 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.	-1.86**	-2.24***	-0.57	-2.90***	-1.30**	-2.37***
	(0.77)	(0.60)	(0.77)	(0.91)	(0.53)	(0.60)
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	29.09	41.56	34.37	56.00	19.39	35.80
SD of Outcome	23.82	22.41	24.78	27.70	19.75	21.78

Notes. The table replicates the baseline analysis, but for the subset of users who are observed at least twice before and after NetzDG came into effect. It shows the main coefficients of the difference-in-differences estimations comparing the hate intensity in tweets by those staying users that are affected and unaffected by the law (NetzDG). The columns contain the outcome measures discussed in the data section: Continuous scores ranging from 0 to 100 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 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 year-month fixed effects, user 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

Table 29: Triple Differences: The Effect of NetzDG on the Intensity of Hate in Tweets for Users following a German and Austrian Populist Party

	(1) sev. Toxicity	(2) Toxicity	(3) Threat	(4) ID Attack	(5) Profanity	(6) Insult
Germany	5.31*	5.30*	4.75***	7.90*	1.48	0.00
	(3.08)	(2.76)	(1.72)	(4.25)	(1.88)	(.)
AfterT=1	0.00	0.00	0.00	0.00	0.00	0.00
	(.)	(.)	(.)	(.)	(.)	(.)
Germany × AfterT=1	-4.17**	-5.47***	-1.56	-8.73***	-2.84**	-2.21***
	(1.93)	(1.76)	(1.18)	(2.73)	(1.23)	(0.57)
follow both countries=1	2.28	3.24	0.50	6.77	1.06	0.00
	(3.82)	(3.31)	(2.05)	(4.43)	(2.53)	(.)
Germany × follow both countries=1	-1.98	-1.87	-0.95	-5.46	0.51	0.00
	(3.86)	(3.38)	(2.32)	(4.62)	(2.66)	(.)
AfterT=1 × follow both countries=1	-1.79	-2.54	0.71	-5.89**	-1.78	-1.24
	(2.40)	(2.11)	(1.44)	(2.73)	(1.76)	(1.01)
Germany × AfterT=1 × follow both countries=1	2.60	2.82	-0.15	7.31**	1.01	0.57
	(2.86)	(2.53)	(1.65)	(3.38)	(2.03)	(1.50)
Year-Month FE	Yes	Yes	Yes	Yes	Yes	Yes
Account age FE	Yes	Yes	Yes	Yes	Yes	Yes
User ID	No	No	No	No	No	Yes
R^2	0.04	0.05	0.02	0.05	0.03	0.18
Observations	160406	160406	160406	160406	160406	160165
Mean of Outcome	29.97	42.81	34.73	57.41	20.48	37.24
SD of Outcome	24.13	22.47	24.90	27.59	20.47	22.06

Notes. The table shows the main coefficients of the triple-difference estimations comparing the hate intensity in tweets by users affected and unaffected by the law (NetzDG) and following populist parties of either one or both countries. The columns contain the outcome measures discussed in the data section: Continuous scores ranging from 0 to 100 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 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 year-month fixed effects and fixed effects for the account age in months when the respective tweet was posted. User fixed effects are dropped as the information if a user follows one or two parties is invariant for a user. Standard errors are clustered at the user level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.010$.

Download ZEW Discussion Papers:

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