The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG*

Rafael Jiménez Durán† Karsten Müller‡ Carlo Schwarz§

February 23, 2024

Abstract

We study the online and offline effects of content moderation on social media using the introduction of Germany's "Network Enforcement Act" (NetzDG), which fines social media platforms failing to remove hateful posts. We show that the law transformed social media discourse: posts became less hateful, refugee-related content less inflammatory, and the use of moderated platforms increased. The NetzDG also had offline effects by reducing anti-refugee hate crimes by 1% for every standard deviation in exposure to far-right social media use. The law reduced hate crimes partly by making it harder for perpetrators to coordinate, without changing attitudes toward refugees.

Keywords: Social Media, NetzDG, Content Moderation, Hate Crime, Refugees, Germany
JEL Codes: L82, J15, O38.

*We are grateful to Leonardo Bursztyn, Ruben Durante, Fabrizio Germano, Ruben Enikolopov, Sophie Hatte, Ro'ee Levy, Sulin Sardoschau, Joshua Tucker, David Yang, Noam Yuchtman, Ekaterina Zhuravskaya and seminar participants at NBER Political Economy Fall Meeting, Bocconi University, Cornell University, 2nd CEPR Workshop on Media, Technology, Politics, and Society, the Winter Meeting Econometric Society, EEA-ESEM conference, University of Cologne, University of Antwerp, University of Bristol, and Toulouse School of Economics for their helpful suggestions. Müller acknowledges financial support from a Singapore MoE start-up grant and Presidential Young Professorship (numbers A-0003319-01-00 and A-0003319-02-00).

†Università Bocconi, Department of Economics, IGIER, Stigler Center, rafael.jimenez@unibocconi.it.
‡National University of Singapore, Department of Finance, kmueller@nus.edu.sg.
§Università Bocconi, Department of Economics, IGIER, PERICLES, CEPR, CAGE, carlo.schwarz@unibocconi.it.

1 Introduction

One of the most frequently voiced charges against social media platforms, such as Facebook and Twitter, is that they have amplified existing societal tensions. Forty percent of Americans have experienced some form of online harassment (Anti-Defamation League, 2022), and many are concerned that hateful conversations on social media might contribute to the spread of hateful attitudes offline. Recent empirical evidence suggests that hateful posts on social media can indeed spill over into violent offline actions against ethnic and religious minorities (Bursztyn et al., 2019; Müller and Schwarz, 2021, 2023).

Social media companies have not sat idle in addressing these problems. Hate speech has been officially prohibited on YouTube since at least 2006, on Facebook since at least 2012, and on Twitter since 2015 (Twitter, 2015; Gillespie, 2018). However, efforts at content moderation remain highly controversial: Some people object that platforms are not moderating enough, while others raise concerns about online censorship. To evaluate whether content moderation policies are socially desirable, it is therefore crucial to understand whether content moderation can reduce online hate and offline violence.

This paper sheds light on this question by focusing on the first legal change explicitly aimed at incentivizing social media platforms to increase their moderation efforts: the German "Netzwerkdurchsetzungsgesetz" (Network Enforcement Act, henceforth NetzDG). The NetzDG was enacted on September 1, 2017, in response to a spike in online hate speech during the influx of more than one million refugees into Germany as a result of the 2015-2016 refugee crisis. The law marks a unique and unprecedented legal change that introduced penalties for large social media platforms of up to €50 million for failing to remove hateful content promptly.[1] As such, the law drastically changed social media providers' incentives to moderate such content and has been called a "key test for combatting hate speech on the internet" (Echikson and Knodt, 2022).

This paper investigates whether the increased content moderation efforts induced by the NetzDG decreased online and offline hate. In our analysis, we focus on toxic online content and real-life hate crimes targeting refugees, given the widespread nature of anti-refugee sentiment in the German context during that period (see Müller and Schwarz, 2021). Building on this existing work, we analyze the Twitter and Facebook accounts of followers of the far-right party Alternative for Germany ("Alternative für Deutschland," henceforth AfD). At the time the NetzDG was enacted, the AfD was the third-largest party in the German parliament, having risen on a platform of anti-refugee

[1]The NetzDG targeted social media companies with more than two million users. Besides Facebook and Twitter, the law applies or has applied to Change.org, Instagram, Google Plus, YouTube, Pinterest, Reddit, SoundCloud, TikTok, Twitch, and Jodel.

rhetoric. Importantly, the AfD also had, and still has, the largest Facebook following of any German party.

The empirical analysis proceeds in two parts. In the first part, we focus on the online effects of the NetzDG. To measure the impact of the law on the hatefulness of refugee-related Twitter content, we collect the universe of tweets that contain the word "refugee." We measure the hatefulness of these tweets using Google's Perspective API, a machine learning algorithm commonly used in industry applications and as a benchmark in academic studies. This algorithm assigns a "toxicity" score to each tweet, which can roughly be interpreted as the fraction of individuals considering it offensive or disrespectful. In a difference-in-differences analysis with Twitter data, we compare the content produced by "toxic users" and "non-toxic users" before and after the NetzDG was implemented. Intuitively, users who posted more toxic tweets before the law's passing should be more exposed to more stringent online content moderation. We consider two definitions of "toxic" users: those whose tweets, on average, fall into the top quartile of toxicity pre-NetzDG, or those who follow the AfD account on Twitter.

Consistent with an increase in content moderation efforts, we find an immediate and significant decrease in the toxicity of refugee-related tweets after the NetzDG became binding. Compared to the pre-period mean, there is a 19% (0.33SD) drop in the toxicity of tweets posted by users in the top quartile of pre-NetzDG toxicity and a drop of around 5% (0.08SD) for tweets posted by AfD followers. The results are robust to alternative definitions of "toxic users" and alternative measures of toxicity, including a measure of threats against an individual or group. We also document a similar reduction in the toxicity of overall tweets beyond refugee-related content.[2]

We provide two additional pieces of evidence on how the NetzDG changed the content of online discussions beyond its effect on toxicity. First, we analyze the frequency of words used by toxic relative to non-toxic users before and after the law. We find a clear shift away from inflammatory issues such as rape and other forms of sexual violence, terrorism, and Nazi comparisons in refugee-related Twitter content. Second, we train a machine learning topic model and show that left-leaning topics such as antisemitism, feminism, and concerns about neo-Nazis became more prevalent after the NetzDG, while refugee and terrorism-related content received less attention. We find no evidence for changes in discussions of censorship or disengagement with controversial political issues, which suggests that most users did not express concerns that the policy stifled freedom of expression online.

[2]Given that historical data from the Twitter API gives access to *surviving* tweets (i.e., the ones not removed by Twitter), these effects are likely driven by both a mechanical removal of hateful tweets and a deterrence effect on the production of hateful content.

As a last piece of evidence on the online effects, we study the impact of the NetzDG on the usage of different social media platforms. For this analysis, we collect a panel of web traffic data at the platform-country level, covering Germany and other Western countries and platforms targeted by the NetzDG and comparable untargeted platforms. Using a triple-difference design, we compare changes in the usage of platforms affected by the NetzDG relative to other platforms in Germany as opposed to other countries. These estimates suggest that the NetzDG increased the unique users of affected platforms by 8.1%. This finding is consistent with the idea that, for many users, platforms that enforce content moderation more stringently might be more attractive, perhaps because hate speech excludes some people from online conversations (Waldron, 2012).[3]

The second part of the paper studies the offline effects of the NetzDG. We investigate whether the policy-induced content moderation efforts also translated into fewer real-life hate crimes against refugees. For this analysis, we exploit municipality-level differences in the exposure to far-right social media content. To the extent that the NetzDG limited online hate speech, one would expect a decrease in the number of anti-refugee incidents in areas where more people were exposed to hateful content in the first place. Using two-way fixed effects regressions, we find that the introduction of the NetzDG led to a reduction of anti-refugee incidents in municipalities with many AfD Facebook followers. Specifically, municipalities with one standard deviation higher AfD followers per capita saw a 1% reduction in the number of anti-refugee incidents.

The underlying identification assumption of this approach is that in the absence of the NetzDG, municipalities with different prior exposures to hate speech on social media would have seen similar trends in anti-refugee incidents. In support of this assumption, we show that municipalities with different levels of AfD followers had similar trends in hate crimes in the period leading up to the enactment of the NetzDG. Our findings are also robust to controlling for many municipality characteristics and a large battery of sensitivity checks. For example, the estimates are not driven by differences in support for the AfD in the 2017 federal election, by general local social media or internet penetration, nor by the number of pre-existing refugees. The results are also unlikely to be driven by the tailing off of the refugee crisis itself. These main results remain unchanged if we consider Twitter-based (as opposed to Facebook-based) exposure measures, alternative variable transformations, different standard errors, or more restrictive fixed effects.

[3]Note that this finding does not necessarily imply that platforms find it profitable to increase their content moderation efforts. Content moderation—particularly when responding to user reports (as is the case of the NetzDG)—can be quite resource-intensive since it typically requires humans to review content manually (Gillespie, 2018).

We provide evidence that both the number of consumers and producers of anti-refugee content in a municipality matter for reducing anti-refugee incidents. Specifically, we find a stronger reduction of anti-refugee hate crimes, over and above what is predicted by the number of AfD followers (consumers), depending on the frequency with which users produce content on the AfD's Facebook page (as measured by posts, likes, comments, or shares). For example, municipalities with one standard deviation higher number of posts per AfD follower experience a further 0.5 percentage point reduction in the number of anti-refugee hate crimes after the NetzDG.

We further corroborate our main evidence using a synthetic control group approach, comparing overall hate crimes in Germany (including those unrelated to refugees) to other countries using a harmonized cross-country dataset. Specifically, we build a synthetic control for Germany using data for the period 2009-2020 from 21 donor countries, following the methodology of Abadie and Gardeazabal (2003) and Abadie et al. (2010). Using the full path of pre-intervention hate crimes as predictors, we find that the policy resulted in an annual decrease in 0.03 hate crimes per 10,000 inhabitants, or roughly 250 fewer hate crimes per year. This finding is robust to a battery of robustness checks, including placebo exercises that assign treatment to other countries or that focus on the overall rate of homicides (which should not be affected by the NetzDG).

Finally, we examine two plausible mechanisms for why content moderation prompted by the NetzDG has prevented anti-refugee hate crimes. The NetzDG could (1) make it harder to coordinate attacks online or (2) change individual attitudes towards refugees. We provide some evidence suggesting that the NetzDG made the coordination of anti-refugee incidents more difficult by differentiating incidents by the number of perpetrators. For this analysis, we hand-coded how many persons were involved in an attack on refugees for the 10,080 incidents in our data based on a description of each case. In line with the hypothesis that the NetzDG disrupted the ability to coordinate online, we find that the estimates are twice as large for anti-refugee incidents committed by multiple relative to single perpetrators. These findings are consistent with existing evidence in the literature that coordination may be an important mechanism in explaining the link between online and offline violence (Bursztyn et al., 2019; Müller and Schwarz, 2021).

We also analyze whether the NetzDG led to changes in the attitudes of social media users towards refugees using data from the German Socio-Economic Panel (GSOEP, Goebel et al., 2019). To this end, we investigate within-person changes in attitudes and actions toward refugees, comparing active social media users relative to non-users. We find no evidence for improved attitudes toward refugees considering all respondents or

conditioning on AfD supporters. These findings make it unlikely that more positive attitudes toward refugees predominantly explain the reduction in anti-refugee incidents.

Taken together, our findings suggest that the NetzDG had significant effects on online discourse and offline violence. However, our view is that the policy implications of these findings are not entirely obvious. Our evidence is consistent with the NetzDG helping reduce the use of social media for coordinating anti-refugee incidents without reducing its use for discussing controversial political issues. The increase in the unique number of social media users is also consistent with content moderation helping include more users in online conversations. That said, any complete welfare analysis would require investigating other offline effects beyond hate crimes, such as a potential undermining of freedom of speech, which is harder to measure with observational data.

Contribution to the literature. We mainly contribute to three strands of the literature. First, there is a growing literature on the real-life effects of social media. Existing work has investigated the impact of social media, among other outcomes, on mental health and well-being (Allcott et al., 2020; Braghieri et al., 2022), polarization (Sunstein, 2017; Allcott and Gentzkow, 2017; Boxell et al., 2017; Levy, 2021; Mosquera et al., 2020), protests (Enikolopov et al., 2020; Acemoglu et al., 2017; Fergusson and Molina, 2021; Howard et al., 2011), corruption and confidence in government (Enikolopov et al., 2018; Guriev et al., 2020), and voting (Bond et al., 2012; Jones et al., 2017; Fujiwara et al., 2023). See Zhuravskaya et al. (2020) for a review of the recent literature on the political effects of social media. Most closely related is the work that provides evidence of the impact of social media on hate crimes (Müller and Schwarz, 2021, 2023; Bursztyn et al., 2019; Du, 2023; Cao et al., 2023) for different social media platforms, countries, and minority groups. Despite this growing body of evidence, we know very little about how to effectively curb the adverse real-world effects of hateful social media content. To the best of our knowledge, our paper is the first to provide evidence about the offline impact of online content moderation policies.

Second, we contribute to a nascent literature that studies platform decisions and content moderation strategies in the context of hate speech and toxic content (Liu et al., 2021; Madio and Quinn, 2021; Kominers and Shapiro, 2024).[4] Jiménez Durán (2022) finds that changing beliefs about content moderation has an insignificant effect

[4]There is a parallel literature studying the moderation of misinformation. See Barrera et al. (2020); Henry et al. (2022); Guriev et al. (2023) for recent experimental work comparing different interventions targeting misinformation and Aridor et al. (2024) for a review. Another slightly related literature is the work on censorship of the internet and social media in autocratic regimes (Qin et al., 2017; Chen and Yang, 2019).

on consumer surplus. This finding suggests that the most sizeable welfare effects of content moderation could be due to its impact on out-of-platform outcomes, such as hate crimes. Beknazar-Yuzbashev et al. (2022) find that lowering users' exposure to toxic content on social media can decrease several measures of engagement and decrease the likelihood of posting subsequent toxic content.[5] Müller and Schwarz (2022) study the aftermath of Donald Trump's Twitter account deletion during the January 6 Capitol attack in the United States and document decreases in online toxicity but also platform engagement of Trump's followers vs. nonfollowers. Our findings on online toxicity are consistent with prior work by Andres and Slivko (2021), who estimate the effect of the NetzDG on the toxicity of right-wing Twitter users in Germany relative to Austria with a difference-in-differences design. In line with our results, they find that German AfD followers posted relatively less toxic content after the NetzDG. Different from their analysis, we focus on *within-country* variation for the results on online toxicity. In addition, our paper is the first to jointly study changes in online content, platform usage, hate crimes, and attitudes in the aftermath of the NetzDG.

Lastly, we speak to a broader literature on the effects of media on violence. Research by Yanagizawa-Drott (2014), DellaVigna et al. (2014), and Adena et al. (2015), for example, suggests that nationalist propaganda on the radio can increase the prevalence violence against minorities. Djourelova (2023) shows the effect of slanted language on attitudes toward immigrants. In other work, Dahl and DellaVigna (2009), Card and Dahl (2011), and Bhuller et al. (2013) investigate the effect of movies, TV, and the internet on different types of violence. Unlike social media, traditional media undergoes editorial processes and is easier to subject to regulatory oversight. Instead, social media companies indirectly shape content through platform design and content moderation. Nevertheless, our findings suggest that, even in this setting, a policy that imposes penalties (much like a Pigouvian tax) can affect online content and potentially reduce offline externalities.

2 Background

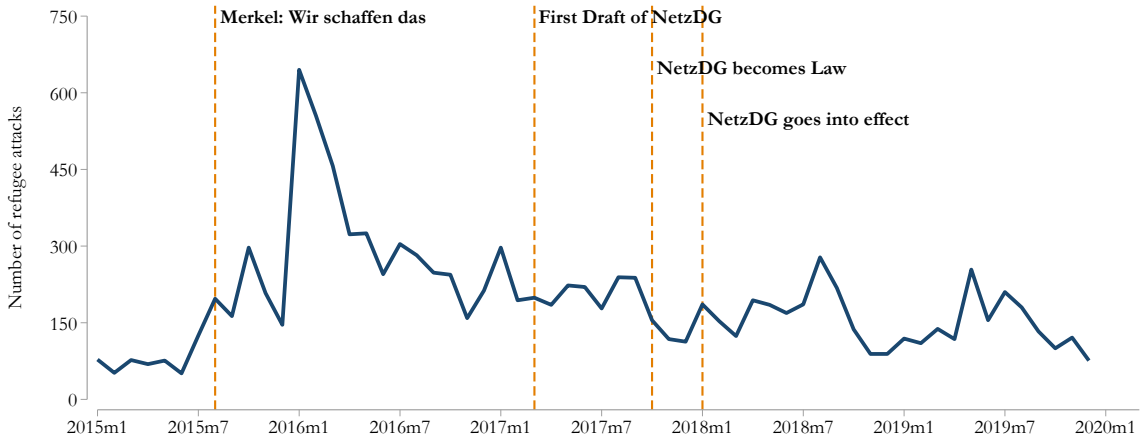
In August 2015, Chancellor Angela Merkel declared that Germany would welcome a large number of refugees from the Syrian Civil War and other conflicts who had arrived in Europe in the previous months. Following this "Wir schaffen das!"(we can do this)

[5]Their engagement decrease might seem at odds with the increase we document in the usage of platforms in response to higher moderation. However, toxicity could increase engagement while decreasing the utility from joining a platform (Beknazar-Yuzbashev et al., 2024). Our finding of an extensive margin effect is consistent with users, on net, deriving a higher benefit for more moderation.

speech, over 1.3 million refugees entered Germany over the 2015-2016 period. The inflow of refugees only slowed considerably after the European Union struck a deal with Turkey in March 2016, in which Turkey agreed to prevent Syrian refugees from crossing over to the EU in exchange for financial compensation (Parliament, 2016).

The large inflow of asylum seekers into Germany was accompanied by a flare-up in the number of anti-refugee incidents. The non-profit organization "Amadeu Antonio Stiftung" recorded more than 10,080 hate crimes targeting refugees in Germany between 2016 and 2020, visualized in Figure 1. Hate crimes spiked after Merkel's "Wir schaffen das" speech and peaked following the widely-reported 2016 New Year's Eve sexual assaults by refugees in Cologne. The frequency of these hate crimes also drew the attention of the international news media (see, for example, New York Times, 2017). Importantly, hate crimes against refugees continued even after the flow of refugees to Germany stopped following the EU-Turkey deal in March 2016.

Figure 1: Attacks on Refugees in Germany, 2015-19



Notes: This plot shows the monthly number of refugee attacks in Germany between 2015 and 2019 based on data from the Amadeu Antonio Stiftung, a non-profit organization. The dashed vertical lines mark the date of Merkel's "Wir schaffen das!" speech and important dates in the creation and approval of the NetzDG.

In previous research, Müller and Schwarz (2021) showed that social media played a role in this wave of anti-refugee crime. In particular, the Facebook page of the Alternative for Germany (AfD) became an important platform for the spread of anti-refugee content. The evidence suggests that these far-right Facebook pages helped propagate anti-refugee sentiment, and the exposure to such online content motivated real-world anti-refugee incidents.

In August 2015, Germany's Minister of Justice Heiko Maas demanded that social media companies should enforce existing laws prohibiting hate speech (Economist, 2018). In an open letter, Maas wrote: "The internet is not a lawless space where racist abuse and illegal posts can be allowed to flourish [...]." Due to what he deemed insufficient action by the social media companies, Maas introduced a first draft of the "Netzwerkdurchsetzungsgesetz" (NetzDG) in March 2017 to stem the wave of hateful content that was circulating on German social media.[6] The first draft of the NetzDG stated explicitly that "hate speech and other criminal content that cannot be effectively combated and prosecuted pose a great threat to peaceful coexistence in a free, open and democratic society" (authors' translation; Deutscher Bundestag, 2017). The NetzDG eventually passed the German parliament in September 2017. The NetzDG became law in October 2017 and penalties went into force on January 1st, 2018.

The NetzDG was "the first law that formalizes the process for platform takedown obligations" (Kohl, 2022). While it was not the first attempt at regulating online content moderation, the law marked a clear shift in the incentives of social media platforms. For the first time, a law established financial penalties of up to €50 million if social media companies with more than 2 million registered users in Germany failed to remove hateful content within 24 hours of being reported by German users.[7] The first companies to be covered by the law were Google (YouTube and Google+), Facebook (Facebook and Instagram), Twitter, and Change.org (Echikson and Knodt, 2022).[8] To incentivize users to report hateful content, the NetzDG required platforms to implement dedicated buttons to report violations against the law. Appendix Figure A.1 shows an example of such a reporting tool. The law also imposed an unprecedented transparency requirement

[6]Before the NetzDG, Maas had attempted to work with the major social media companies to reduce the prevalence of hate speech. In December 2015, the Task Force Against Illegal Online Hate Speech—formed by Facebook, Twitter, Google, and some anti-hate advocacy groups in Germany—signed a Code of Conduct. The companies agreed to remove hate speech promptly and to facilitate user reports. However, after several months, Maas noted that "the networks aren't taking the complaints of their own users seriously enough," which led him to introduce legislation with monetary penalties (Kaye, 2019). At the European level, Facebook, Microsoft, Twitter, and YouTube signed a voluntary Code of Conduct with the European Commission in May 2016 to review reported illegal content within 24 hours (Gillespie, 2018). See Gorwa (2019) for a compilation of formal and informal platform governance efforts around that time.

[7]After receiving user reports, companies typically evaluate whether the content violates their guidelines. If so, they take action on the content (e.g., delete a post globally). If the content does not violate their guidelines, it is assessed vis-à-vis the German Criminal Code listed in the NetzDG. If it is considered unlawful under the NetzDG, companies disable access to that content in Germany. See for example, a recent transparency report by Instagram: `https://transparency.fb.com/sr/netzdg-report-english-ig-jul-21`.

[8]Subsequently, other platforms such as Jodel, TikTok, Reddit, SoundCloud, Pinterest, and Twitch started providing the transparency reports required by the law. See `https://www.bundesanzeiger.de/pub/de/suchen2?7`.

for platforms to publish a biannual report on their content moderation activities (Heldt, 2019).

Similar provisions for social media platforms later became part of the "Online Safety Bill" in the United Kingdom and the "Digital Services Act" of the European Union, even though these laws do not implement direct financial penalties for platforms. Thus, the NetzDG provides a crucial testing ground for the effectiveness of such legislation. In the next section, we describe our main data sources that will allow us to investigate the impact of the NetzDG on online hate speech and offline hate crimes.

3 Data

Our main analysis builds on five separate datasets. First, we construct a database of refugee-related tweets that allows us to study the impact of the NetzDG on the toxicity of online content. Second, we construct a web traffic panel at the country-platform-quarter level that allows us to measure the effect of the law on the treated platforms' user base. Third, for our analysis of the offline effects of the NetzDG, we construct a municipality-quarter panel of anti-refugee incidents. Fourth, we use survey responses from the German Socio-Economic Panel (GSOEP) (Goebel et al., 2019) to study attitudes towards refugees. Fifth, for our synthetic control analysis, we build a cross-country panel of total hate crime. We describe the main data sources for each dataset in the following.

Refugee-related Twitter Content

To provide evidence for the effects of the NetzDG on the toxicity of social media content, we create a tweet-level dataset measuring the online toxicity of refugee-related tweets. We focus on Twitter data because Facebook, unfortunately, does not allow the collection of posts directly from user profiles. In contrast, Twitter provides rich post and user data, and, importantly, it is also one of the twelve platforms that have been subject to the NetzDG.

We use the full-archive search endpoint of Twitter's Academic API and obtain all tweets containing the word "Flüchtling" (German for *refugee*) between January 2016 and December 2019. As discussed in Section 2, the focus on refugee-related Twitter content is motivated by the increases in online hate speech that occurred during the refugee crisis and the existing evidence that links this online content to offline violence. We thus investigate the effect of the NetzDG on the hatefulness of refugee-related German tweets. In total, this dataset contains 811,332 tweets. Appendix Figure A.2 plots the

monthly number of tweets mentioning the word "Flüchtling" (refugee), which shows no downward shift in the number of refugee-related tweets after the implementation of the NetzDG. We also investigate changes in the overall discourse on Twitter by collecting all other tweets posted by the users in our sample. To identify the political leaning of users, we additionally scraped the Twitter follower lists of all major German parties. These lists allow us to identify which Twitter users follow the AfD's Twitter account.

We measure the hatefulness of online content using Google's Perspective API (Wulczyn et al., 2017; Dixon et al., 2018). This API returns a machine-learning-based "toxicity" score between 0 and 1 (where 1 is the most toxic). The score is interpreted as the fraction of people who consider the content to be "toxic," which is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." Besides the main toxicity measure, the API also provides other scores, which include severe toxicity, identity attack, insult, profanity, and threat.[9]

Appendix Table A.2 contains summary statistics for our sample of refugee tweets. On average, refugee-related tweets have a toxicity score equal to 0.4. To get a sense of what kind of language these numbers imply: "Ich mag keine Flüchtlinge" (I don't like refugees) has a toxicity score equal to 0.41, and "Flüchtlinge sind Müll" (Refugees are trash) has a toxicity of 0.8. Around 29% of tweets in the sample were posted by AfD followers, and 50% of them were posted by users following at least one political party. In Appendix Table Table A.1, we provide several examples of toxic tweets in our data.

Platform Usage Panel

To measure the effect of the NetzDG on the usage of websites targeted by the law, we construct a panel dataset with web traffic data at the website-country-quarter level. We obtain the number of unique users and total visits between January 2017 (the earliest available data) and 2019, for seven major online platforms in 36 OECD countries from Semrush.com.[10] Four of the selected platforms (Instagram, Twitter, YouTube, and Facebook) were the first ones to be subject to the NetzDG in Germany, while three were never subject to it (Amazon, Netflix, and Wikipedia). We focus on the largest platforms because web traffic data is estimated from clickstream data from a panel of

[9] See `https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US`.

[10] These countries are those that joined the OECD before 2020, see `https://www.oecd.org/about/document/ratification-oecd-convention.htm`.

users' browsing behaviour (collected from browser extensions and mobile applications), which is imprecise for smaller platforms.[11]

Municipal Anti-Refugee Hate Crime Panel

The analysis of the offline effects of the NetzDG is based on a panel dataset for the number of anti-refugee hate crimes for each German municipality between January 2016 and December 2019, aggregated at the quarterly level. The underlying data on anti-refugee hate crimes were collected by the Amadeu Antonio Foundation and Pro Asyl (a pro-asylum NGO).[12] Information on around three-quarters of these incidents comes from administrative police data reported based on parliamentary requests. All of the 10,080 anti-refugee crimes are classified into four groups. The most common cases are property damage to refugee homes (7,814 incidents), followed by assault (1,693), incidents during anti-refugee protests (72), and arson (153). 348 events are classified as "suspected cases" that are still under investigation. We provide one example for each class of anti-refugee incidents in Appendix Table A.3. We are able to link incidents to their corresponding municipality because they are geo-coded with exact longitude and latitude. We assign these incidents to municipalities using shape files provided by the ©GeoBasis-DE/BKG 2016 website.[13]

Most of the additional municipality-level variables are based on the replication data from Müller and Schwarz (2021).[14] The main measures of far-right Facebook usage from Müller and Schwarz (2021) which we use in our analysis are based on the number of AfD Facebook followers in each municipality, which was obtained by hand-collecting and geo-coding a place of residence for 34,389 users who interacted with AfD's Facebook's page as of October 2017. The motivation to use the AfD's Facebook page is that the AfD is a right-wing populist party whose Facebook page is arguably the key platform for anti-refugee content online and has a broader reach than the Facebook page of any other German party. Moreover, we focus on Facebook because it is the most widely used platform in the German setting. We augment the data with information about the

[11] In particular, it is likely that there are not enough observations at the country-quarter-website level for smaller platforms. See `https://www.semrush.com/blog/what-is-clickstream-data/`. For example, Change.org, which was the one other platform among those initially subject to the NetzDG, has only 1.6% of Facebook's traffic according to Semrush estimates.

[12] These data are available at `https://www.mut-gegen-rechte-gewalt.de/service/chronik-vorfaelle`.

[13] The analysis is conducted on the level of 4,679 German municipalities ("Gemeindeverwaltungsverband"). After removing uninhabited areas, we are left with 4,466 municipalities in our sample. We use the level of the "Gemeindeverwaltungsverband" instead of "Gemeinden" since the area and population of these administrative areas are more similar.

[14] The underlying reproduction file is available here.

activity of each user. This allows us to construct the number of posts, likes, comments, and shares for each AfD user.[15]

We visualize the relationship between far-right Facebook usage and hate crimes in Figure 2. The map shows quintiles of AfD Facebook usage per capita overlayed with the location of anti-refugee incidents (orange dots). There is considerable geographical variation in both incidents and AfD users. Appendix Table A.4 presents summary statistics for anti-refugee incidents, our measure of exposure to online hate speech (AfD users per capita), and our control variables. The unit of analysis is a municipality-quarter. There are 10,080 anti-refugee incidents in our sample. There was at least one incident in every quarter of our study period, and 48% of municipalities experienced at least one incident. On average, municipalities have 3 AfD users per 10,000 inhabitants and 80% have at least one AfD user.

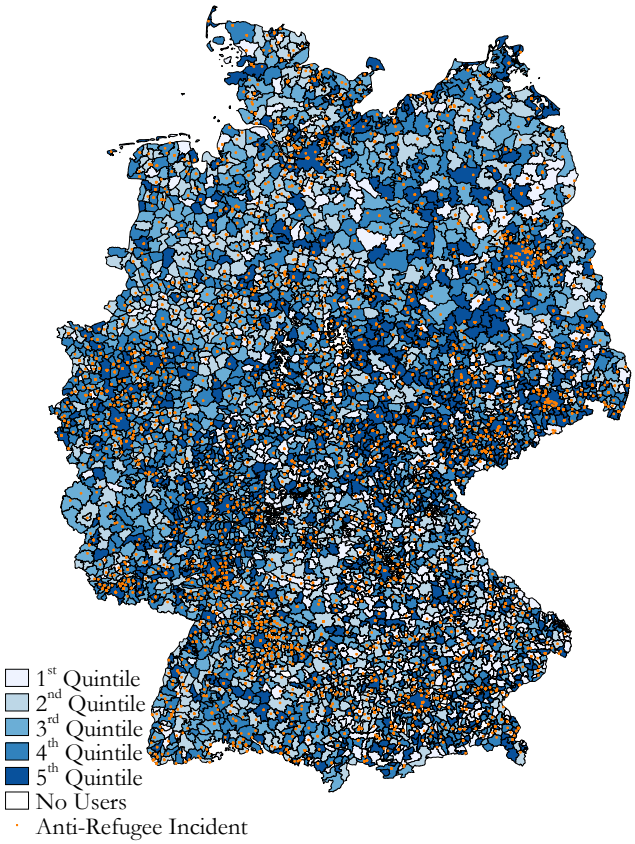
To control for the number of Facebook users in a municipality, we create a measure using Google search. In particular, we use a list of the names of over 2,000 German cities as well as all German municipalities and use the Google Search API to obtain the number of people who indicate living in each municipality on their Facebook profile. To do so, we search for "Lives in: *City Name*" restricted to Facebook.com, where *City Name* corresponds to either a city's or municipality's name. These Google searches return the number of Facebook user profiles where people indicate living in a particular municipality, which should be a sound proxy for the number of local Facebook users.

We further construct an alternative exposure measure based on the number of AfD Twitter followers in a municipality. For this measure, we use the location information from the user profiles that we have collected for our analysis of Twitter content. This data allows us to verify our findings based on the exposure to hateful content on an alternative social media platform.

Finally, we add municipality-level socio-economic controls and measures of voting and media consumption behavior. The main source of socioeconomic data is the German Statistical Office, which disseminates regional data via www.regionalstatistik.de. For each municipality, we can measure population by age group, GDP per worker, population density, and the vote results for the German Federal Election in September 2017. We also have data on the immigrant and asylum-seeker population share. Data on broadband internet availability comes from the Federal Ministry of Transport and Digital Infrastructure (BMVI). To measure the popularity of traditional media, we use data for 2016-2017 newspaper sales from the "Zeitungsmarktforschung Gesellschaft der deutschen Zeitungen (ZMG)" (Society for Market Research of German Newspapers),

[15]The shares were not included in the replication file but stem from the same Facebook scraping.

Figure 2: Map AfD Facebook Users and Anti-Refugee Incidents



Notes: The shading of the maps indicates the quintiles of the distribution of AfD users per capita for the municipalities in Germany. Orange dot indicate anti-refugee incidents.

normalized by municipality population. Data on other types of crimes by county and year come from the Bundeskriminalamt (BKA)'s Police Crime Statistics.

Survey Data on Attitudes Towards Refugees

To study whether the NetzDG was associated with changes in attitudes towards refugees, we extract a set of relevant questions from the German Socio-Economic Panel (GSOEP) (Goebel et al., 2019). The GSOEP is the largest yearly household panel survey in Germany. In the 2016 and 2018 waves of the GSOEP, respondents were asked a battery of questions about their attitudes toward refugees. For example, respondents were asked if refugees are good for the 1) economy, 2) culture, or 3) their place of residence.

Furthermore, the GSOEP asks whether respondents have taken actions to help refugees (e.g., by donating or volunteering).

We harmonize the coding of all questions into indicator variables such that 1 represents a positive attitude or action towards refugees. We additionally created indices that capture the average response of a respondent across the attitude and action questions. Further, we use questions on social media habits to create an indicator for respondents who use social media at least once a week. We provide summary statistics for the full set of questions and derived variables in Appendix Table A.6.

Cross-Country Hate Crime Panel

We construct a cross-country panel of hate crime incidents for the years 2009-2020, which enables us to construct a synthetic Germany. The most comprehensive hate crime database covering several countries is compiled by the Organization for Security and Co-operation in Europe (OSCE). We obtained the reported hate crimes for each of the 57 member States of the OSCE, as well as meta-data describing measurement changes over time.[16]

The data that Germany reports to the OSCE, however, include online hate speech offenses. To avoid picking up a spurious effect from changes in hate speech reporting due to the NetzDG, we obtain data on violent hate crimes (which do not include hate speech) from Germany's Federal Ministry of the Interior and Homeland (BMI) from the table *Übersicht "Hasskriminalität:" Entwicklung der Fallzahlen 2001 – 2021*. Violent hate crimes include bomb attacks, arson attacks, homicides (including attempts), robberies, physical injuries, and violent property damages. Lastly, we gathered population counts from the World Bank's World Development Indicators.

Table A.7 summarizes the data availability for the OSCE members and the filters that we impose in order to build a balanced panel of countries, which we describe in more detail in Appendix A. We excluded micro-states, countries that changed their measurement of hate crimes after the NetzDG, and countries with more than 50% (six) missing observations in 2009-2020. To retain as many countries as possible, we linearly interpolate the gaps for the remaining countries but discard those with missing values at the beginning or end of the series. The resulting dataset contains 21 countries in addition to Germany. Appendix Figure D.3 shows the evolution of hate crimes in Germany and the raw mean of the donor countries. Unsurprisingly, we find that the large differences in

[16]The underlying data can be downloaded from `https://hatecrime.osce.org/country`. The information on reporting changes is available `https://hatecrime.osce.org/national-frameworks-country#dataCollection`.

pre-existing trends across countries make a traditional differences-in-differences analysis impossible.

4 Online Effects of the NetzDG

In the first part of the paper, we investigate the online effects of the NetzDG. The analysis proceeds in three steps. We start by studying the impact on the toxicity of social media content, particularly when it is related to refugees. Then, we investigate whether the NetzDG affected content changes among other dimensions, such as word frequencies and topics. Finally, we analyze the effects of the law on platform usage.

4.1 Impact of the NetzDG on Online Toxicity

As outlined in Section 2, the NetzDG marked a clear shift in the incentives of social media companies to moderate online content. In Online Appendix B, we provide a theoretical framework to derive predictions of such a change in incentives on the prevalence of hateful content. Within the framework, we interpret the NetzDG as a tax that increases the marginal cost of the prevalence of unmoderated hate speech on social media platforms. In the context of a dominant platform—such as Facebook in Germany, which had a 95% market share of daily active users in 2018 (Bundeskartellamt, 2019)—the framework predicts that this policy should result in a decrease in the equilibrium amount of unmoderated hate speech on the platform.

Empirical Strategy

Our strategy compares changes in the toxicity of refugee-related tweets posted by users producing particularly toxic content to other Twitter users, before and after the implementation of the NetzDG. In particular, we expect to see a decrease in the average toxicity of refugee-related tweets posted by more "exposed" users relative to others. We compare toxicity before and after the NetzDG even if, technically, the law could also affect content that was posted before its implementation. However, since the NetzDG relied heavily on users flagging content, newly posted content was more likely to be reported, as it featured more prominently in users' timelines. As a result, the NetzDG disproportionately affected platforms' incentive to delete or hide content posted *after* it went into effect. Note that any content moderation of social media content that was posted before the NetzDG would bias our results towards 0. Our estimates are, therefore, likely a lower bound.

With this in mind, we estimate a difference-in-differences regression of the following form:

$$Toxicity_{iut} = \theta \cdot Toxic\,User_u \times Post\;NetzDG_t + \phi_u + \mu_t + \psi_{iut}, \tag{1}$$

where $Toxicity_{iut}$ denotes the toxicity score of tweet i posted by user u on day t, based on the coding from the Google Perspective API. The main independent variable is the interaction between our exposure measure—an indicator variable for highly toxic users ($Toxic\;User_u$)—-and the post-NetzDG dummy ($Post\;NetzDG_t$). $Post\;NetzDG_t$ is equal to 1 starting in the fourth quarter of 2017 (October 1, 2017), when the NetzDG took effect.

We show results for two definitions of $Toxic\,User_u$. One version defines exposed users as those that sent particularly toxic content before the NetzDG.[17] As a second definition of $Toxic\,User_u$, we use Twitter followers of the AfD, motivated by the fact that the AfD positioned itself as a clear anti-refugee voice in Germany (Müller and Schwarz, 2021). As we show in Appendix Figure C.1, AfD Twitter followers are far more likely to post toxic refugee-related tweets. To avoid the estimates from picking up shifts in the composition of users, we restrict this analysis to users who were active in the pre-period and joined Twitter before January 2016.

Results

Table 1 presents the results from estimating equation (1). Columns (1) and (2) show the results for users who posted highly toxic content before the NetzDG, while columns (3) and (4) show the results for AfD users. All specifications indicate a significant reduction in the toxicity of tweets after the NetzDG. The results hardly change when we include user-specific linear time trends (see columns (2) and (4)). The estimates for highly toxic users in column (2) suggest that the NetzDG was associated with a reduction in the toxicity of tweets of around 19% (0.33SD) relative to the mean. To provide a more intuitive understanding for the coefficient of -0.073, this is approximately the difference in the toxicity of the statements "Flüchtlingsabschaum muss raus aus Deutschland" (refugee scum must be removed from Germany), with a toxicity score of 0.92, and the statement "Diese Flüchtlinge sollen raus aus Deutschland" (these refugees should get out of Germany) with a toxicity score of 0.84. The magnitude for tweets posted by AfD

[17] In our baseline results, highly toxic users are defined as those above the 75th percentile of the pre-period toxicity distribution. In Appendix Table C.1, we show that our results hold irrespective of the precise cutoff.

users (column (4)) is 5% (0.08SD).[18] These results suggest that the reduction in toxicity is smaller for tweets posted by users with a stronger ideological attachment to the AfD.

Table 1: NetzDG and Refugee-related Online Toxicity

	Dep. var.: Toxicity Measures			
	(1)	(2)	(3)	(4)
Highly Toxic User × Post	-0.084***	-0.073***		
	(0.004)	(0.006)		
AfD follower × Post			-0.016***	-0.018***
			(0.003)	(0.004)
User FE	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes
Linear Time Trend		Yes		Yes
Observations	277,135	277,135	277,135	277,135
Pre-Period Mean of DV	0.39	0.39	0.39	0.39
R^2	0.28	0.34	0.28	0.34

Notes: This table presents the results of estimating Equation (1) where the dependent variable is the toxicity of tweets containing the word "Flüchtling" (refugee) (bounded between 0 and 1). In columns (1) and (2) $Toxic\,User_u$ is an indicator variable equal to 1 if a users' tweets before the NetzDG were on average above the 75th percentile of the toxicity distribution. In columns (3) and (4), $Toxic\,User_u$ is an indicator variable that is equal to 1 if a Twitter user follows the AfD's account. All regressions control for user and day fixed effects. Columns (2) and (4) additionally control for user-specific linear time trends. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure 3 shows a dynamic event study version of these specifications, which replaces the *Post* indicator variable with dummies for the quarters around when the NetzDG became binding. Panel (a) shows the event study for highly toxic Twitter users, and Panel (b) that for AfD Twitter users. These figures suggest that the refugee-related tweets posted by highly toxic users and AfD followers followed similar trends of toxicity compared to other Twitter users up to 2017q3. Afterward, toxicity quickly and persistently decreased after the NetzDG became law in 2017q4. For highly toxic users, the estimates consistently remain below their pre-period level. For AfD followers, we observed a decrease until the end of 2018, followed by a spike in toxicity in 2019q1. This spike is likely explained by the court case against a refugee for the rape and murder of a 14-year-old. This court case attracted considerable public attention (e.g., Spiegel, 2019) and was instrumentalized by the AfD to stoke renewed fears of refugees (e.g.,

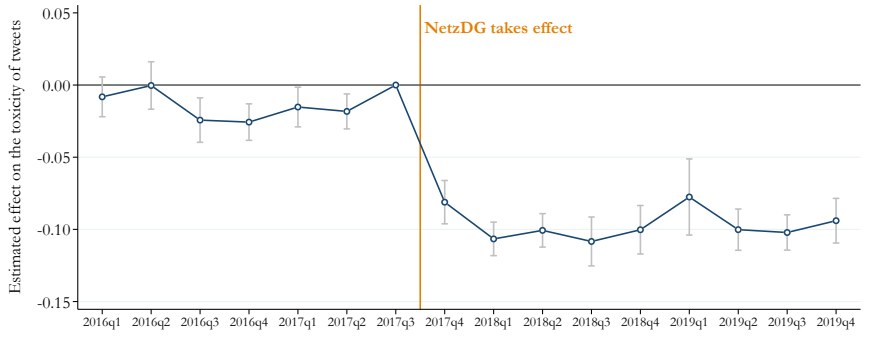
[18] Andres and Slivko (2021) find a reduction of around 2.5% in the monthly volume of hateful tweets on migration and religion sent in Germany relative to Austria.

Frankfurter Rundschau, 2019). We observe a synchronous but much smaller uptick in toxicity of highly toxic users in 2019q1.

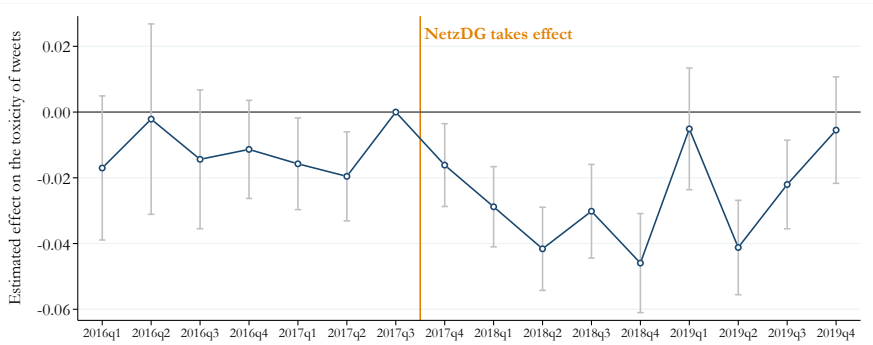
We conduct several robustness checks to validate our findings. In Appendix Table C.1, we consider different cutoffs of pre-period toxicity. Across all specifications, we find a reduction in online toxicity after the passing of the NetzDG. Appendix Table C.2 presents robustness exercises using the different measures of toxicity produced by Google's Perspective API. The effect is consistently significant and negative across almost all toxicity measures. Finally, Appendix Table C.3 presents estimates of the impact of the NetzDG on the overall amount of refugee-related content users produce. For both highly toxic users and AfD followers, the number of refugee-related tweets increased after the passage of the NetzDG.

Figure 3: NetzDG and Online Toxicity of Refugee-related Content

(a) Highly Toxic Users



(b) AfD Followers



Notes: Panels A and B plot the coefficients from event study versions of Equation (1). In Panel (a), we define $Toxic\,User_u$ equal to 1 if a user was in the top quartile of toxicity pre-NetzDG, and 0 otherwise. In Panel (b), we define $Toxic\,User_u$ equal to 1 if a user followed the AfD. The dependent variable is the toxicity of tweets containing the word refugee ("Flüchtling"). The omitted category is the 3rd quarter of 2017, the quarter before the passing of the NetzDG (indicated with the vertical line). The whiskers indicate 95% confidence intervals based on standard errors clustered by user.

Appendix Figure C.2 provides additional evidence on the effectiveness of the NetzDG by investigating the toxicity of all Twitter content without restricting the sample to refugee-related tweets. The figure plots an event study for the average toxicity of tweets sent by highly toxic users in a balanced user-level panel.[19] Similar to the results for refugee-related content, we observe a significant reduction in overall online toxicity after the NetzDG. The regression estimates from this analysis are shown in Appendix Table C.4. In this table, we additionally show that the NetzDG did not decrease the number of tweets sent by highly toxic users. Similar to the results for the toxicity of refugee-related Twitter content, these findings are robust to different definitions of toxic users (see Appendix Table C.5) and hold independently of the toxicity measure we use (see Appendix Table C.6). Notably, column (6) of Table C.6 presents negative effects on a measure of the threatening language of tweets, which may suggest a role of content moderation in obstructing the online coordination of violent acts.[20] We will revisit this hypothesis in Section 5.2.

Taken together, these different pieces of evidence suggest that the NetzDG induced a reduction in the hatefulness of online content. This reduction is likely driven by a combination of three factors. First, online platforms significantly increased their moderation efforts after the NetzDG. From NetzDG compliance reports, we know that Twitter received close to 500,000 reports in 2018 and removed at least 50,000 tweets (Twitter, 2018a,b). Besides the direct removal of content, platforms could have adjusted their algorithms to reduce the exposure of German users to hateful content. Second, as toxic tweets provoke further toxic tweets (e.g., Müller and Schwarz, 2023, 2022), content moderation may have a multiplier effect by which the removal of toxic content prevents additional toxic tweets downstream. Third, the NetzDG could have deterred users from posting toxic content in the first place by affecting their first or second-order beliefs. For example, users may have become concerned about the potential legal repercussions of posting toxic messages (even though actual legal cases are extremely rare). Alternatively, the NetzDG could have changed users' second-order beliefs about how acceptable other users find toxic content.

Given that these channels interact with each other in equilibrium, it is impossible to disentangle their contribution to the aggregate effect we document. Importantly, all three of these mechanisms are in line with the interpretation that the NetzDG was

[19] We do not consider AfD users as "exposed" in this exercise because their toxicity scores for non-refugee topics are not particularly high.

[20] Perspective API defines threats as "Describes an intention to inflict pain, injury, or violence against an individual or group." See `https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages`.

effective in reducing the toxicity of social media content, which is what matters for our analysis. This drop in toxicity motivates the analyses in the second part of the paper, where we examine whether the NetzDG-induced reduction in hateful online rhetoric also affected real-life anti-refugee incidents.

4.2 Impact of the NetzDG on Online Content

As the second step of our analysis, we investigate other changes in the online discourse besides measures of toxicity. In particular, we study changes in *what* gets discussed instead of only focusing on *how* issues are discussed. We use two approaches. First, we study changes in the frequency of words used by toxic relative to non-toxic Twitter users in refugee-related tweets. Second, we use a machine learning topic model to analyze overall shifts in the issues that get discussed online.

Changes in Word Frequencies

As the first test, we analyze changes in word frequencies of refugee-related Twitter content for toxic relative to non-toxic users before and after the NetzDG. Let p_{wgt} be the probability of word w being used by group $g \in \{0, 1\}$ (non-toxic (0) or toxic (1) user) in period $t \in \{0, 1\}$ (before (0) or after (1) the NetzDG). We calculate:

$$\Delta_w = (p_{w11} - p_{w10}) - (p_{w01} - p_{w00})$$

Put differently, we calculate the change in word frequencies for toxic and non-toxic users after the NetzDG. Then, we take the difference in these changes between toxic and non-toxic users. This allows us to understand which words saw the greatest changes in usage among toxic users relative to other Twitter users.[21]

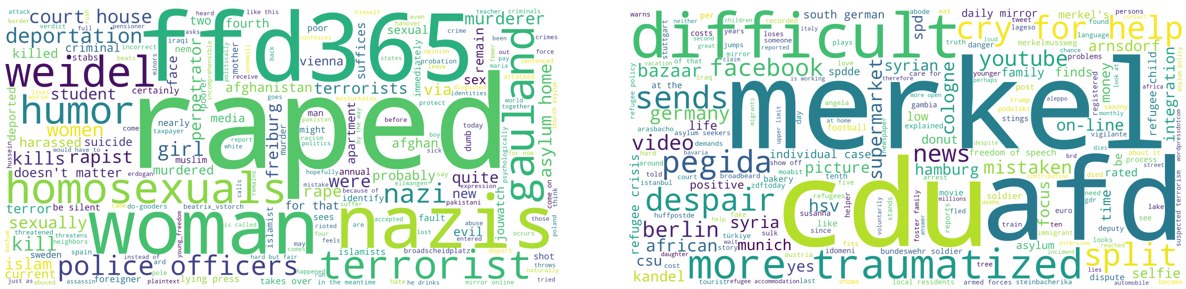
Figure 4 visualizes the results from this analysis. Panel (a) shows words with decreasing frequency, and panel (b) shows words with increasing frequency. For convenience, we have translated everything into English. From this analysis, three findings stand out. First, there is a clear shift away from inflammatory issues such as rape and other forms of sexual violence, terrorism, and Nazi comparisons. The prominently decreasing term "ffd365" was a now-defunct right-wing news website. Second, there is an increase in mentions of mainstream political actors (Merkel, CDU, AfD) and words that suggest a more nuanced debate on the topics, mentioning difficulties, integration, and traumatization. Third, we observe increased mentions of "freedom of speech." This

[21]For the calculation of word frequencies, we cast all words in lowercase, exclude stopwords (very frequent words), and restrict our analysis to the 1,000 most frequent words.

could hint at increased concerns about restrictions to online content, even though we do not frequently observe discussions about the NetzDG and censorship.

Figure 4: Changes in Word Frequencies – Refugee Tweets

(a) Decreasing Frequency **(b) Increasing Frequency**



Notes: This figure shows word clouds that visualize the relative word frequency changes for toxic compared to non-toxic users after the NetzDG among refugee-related tweets. Panel (a) shows words with decreasing relative frequency, while panel (b) shows words with increasing relative frequency. The size of the words is proportional to the frequency change.

Appendix Figure C.3 displays the word frequency changes for all tweets without restricting to refugee-related content. Panel (a) again shows the words with decreasing frequency, and panel (b) shows the words with increasing frequency. There is an overall shift away from political topics and refugee-related issues in particular. There are also fewer mentions of political leaders (e.g., Angela Merkel, Erdogan), political organizations (e.g., AfD, CDU, Pegida), and refugee-related terms (e.g., refugee, Islam, Muslims, terror). Lastly, we also observe a shift from mainstream news outlets (e.g., Welt, Spiegelonline, Zeitonline, NTV) to video platforms like YouTube.

Changes in Topics

As a second analysis, we investigate overall topic changes around the NetzDG using machine learning topic models. Topic models describe a range of techniques that make it possible to automatically group similar text documents into topics. Each topic, in turn, is described by a set of frequent topic words.[22] We use the topic-modeling technique *top2vec* (Angelov, 2020), which combines pre-trained semantic embedding—a technique to represent text as vectors with low dimensionality—with clustering algorithms to

[22]In recent years, topic-modeling techniques have found countless applications in the social sciences and have, among many others, been used to analyze journal articles (Griffiths and Steyvers, 2004), transcripts of the Federal Reserve's Open Market Committee (Hansen et al., 2018), the history of economic thought (Ambrosino et al., 2018), ideologies (Draca and Schwarz, 2024), and parenting styles (Rauh and Renée, 2023). See Gentzkow et al. (2019) and Ash and Hansen (2023) for more details.

identify topics. This model has at least three crucial advantages for our setting. First, the use of pre-trained semantic embeddings allows the model to use information from vast outside corpora to infer the relationships between words, which is particularly helpful for short texts such as tweets. Second, top2vec automatically finds the number of topics instead of us having to choose a topic in an ad-hoc manner. Third, top2vec is able to infer far more finely-grained topics than other commonly used methods such as Latent Dirichlet Allocation.[23]

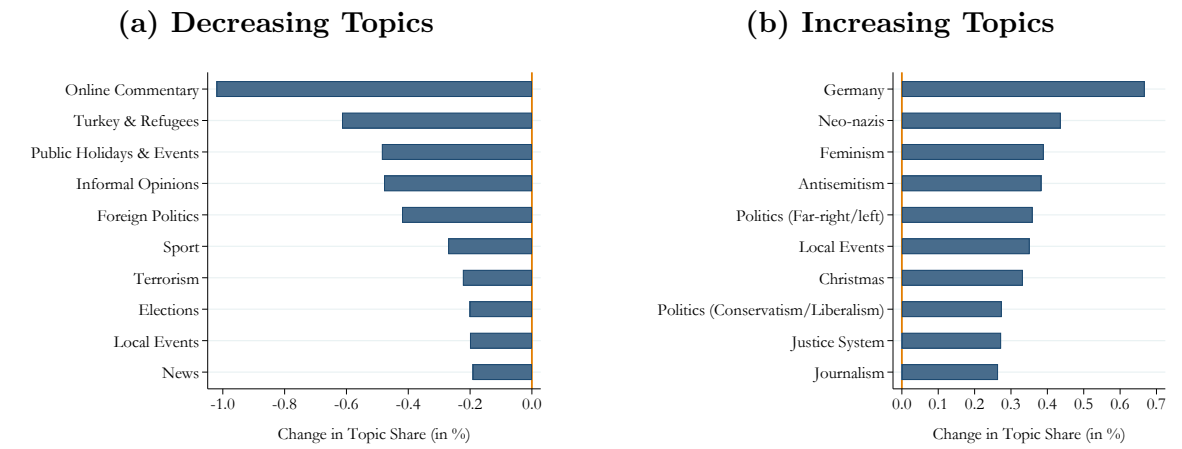
As the training of topic models is computationally intensive, we restrict this analysis to a random subset of one million tweets, which nevertheless should suffice to accurately capture overall topic dynamics. As a preliminary step, we remove links and mentions of accounts from the tweets, as well as the hashtag sign ("#"). This procedure ensures, for example, that "#refugee" is treated equally to the word "refugee." We then fit top2vec to this corpus using the "distiluse-base-multilingual-cased" embeddings model (Reimers and Gurevych, 2019) and "hdbscan" (Campello et al., 2013) as a clustering algorithm. To ensure that the topics are meaningful, we additionally specify that the model creates clusters with at least 250 tweets. The resulting model creates 278 topics. For the analysis, we then calculate the share of each topic among all tweets for the period from 2016q1 to 2017q3 (pre-period) and for the period from 2017q4-2019q4 (post-period).

The results from this analysis are shown in Figure 5. Panel (a) shows the topics with the largest decrease in their topic share, and panel (b) shows the topics with the largest increase. The y-axis contains the topic labels we manually assigned based on the topic words and a reading of the tweets. Appendix Table C.7 shows the full list of words for each of the 20 topics.

The topic model suggests an increased discussion of left-leaning topics like anti-semitism, feminism, and concerns about neo-Nazis. We also see more debate around Germany in general and German Turks in particular. Topics of decreasing importance are, among others, Turkey and refugees, terrorism, and foreign policy. The first two topics, in particular, are important electoral issues for the Alternative for Germany. Overall, the topic model results are consistent with a shift of Twitter discussion towards somewhat more left-leaning topics in the German context. In line with the results based on word frequencies, we do not observe a strong rise in discussions of censorship. The results also do not suggest that people disengage from controversial political issues.

[23]Latent Dirichlet Allocation is the most widely used topic model (Blei et al., 2003). See Schwarz (2023) for a Stata implementation.

Figure 5: Changes in Topics – All Tweets



Notes: This figure shows the ten topics with largest decrease (panel a) or increase (panel b) in their topic share after the NetzDG. The topics were created using the top2vec (Angelov, 2020) topic model. The y-axis lists the five most important topic words for each topic.

4.3 Impact of the NetzDG on Platform Usage

As the last step of our analysis on the online effects of the NetzDG, we investigate its impact on platform usage. One major concern with the NetzDG was that it could stifle the usage of the moderated platforms. While we found no effect of the NetzDG on the number of tweets of toxic users relative to non-toxic users (see Appendix Table C.4 and Table C.4), there could nonetheless be significant changes in the overall usage of moderated platforms. We investigate this possibility based on web traffic data from seven major online platforms in 36 OECD countries provided by Semrush. By "moderated," we mean the four platforms initially subject to the NetzDG in Germany (Instagram, Twitter, YouTube, and Facebook), while by "unmoderated" we mean three that were not (Amazon, Netflix, and Wikipedia).

Equipped with these data, we estimate the following triple-difference regression:

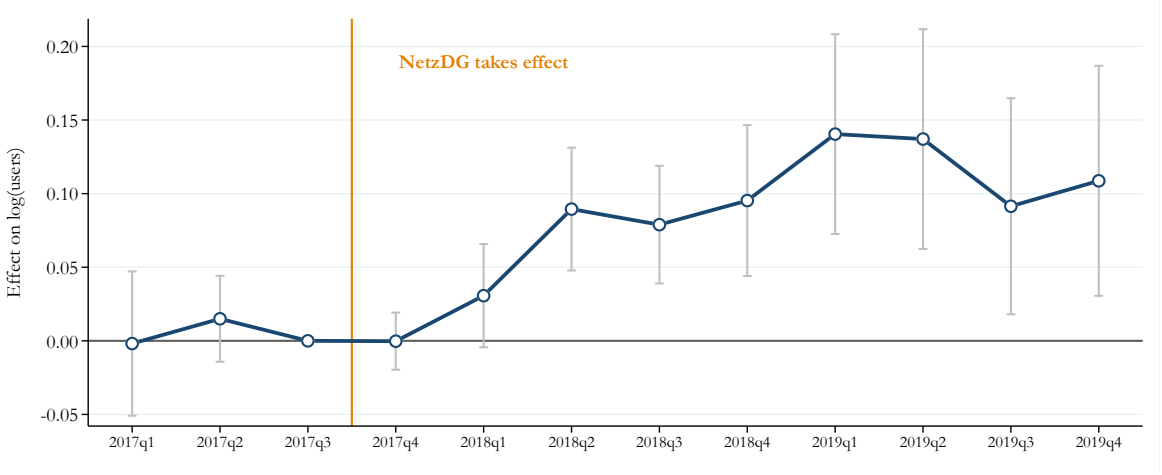
$$\begin{aligned} Usage_{ict} = {} & \beta_1 \cdot Moderated\ Platform_i \times Germany_c \times Post\ NetzDG_t \\ & + \beta_2 \cdot Moderated\ Platform_i \times Post\ NetzDG_t \\ & + \beta_3 \cdot Germany_c \times Post\ NetzDG_t \\ & + \gamma_i + \omega_c + \delta_t + \epsilon_{ict}, \end{aligned} \tag{2}$$

where $Usage_{ict}$ is either the log number of users (unique visitors) or the log number of total visits of platform i in country c in quarter t. $Moderated\ Platform_i$ is an indicator

of whether the platform is subject to the NetzDG in Germany, $Germany_c$ is an indicator for Germany, and $Post\ NetzDG_t$ is an indicator for the quarters after the NetzDG. All regressions include platform (γ_i), country (ω_c), and quarter (δ_t) fixed effects. The main coefficient of interest β_1 measures relative changes in the usage of moderated platforms vis-à-vis unmoderated platforms in Germany relative to changes of usage of the same platforms in other countries.

The identifying assumption underlying these regressions is that without the NetzDG, the relative use of moderated and unmoderated platforms in Germany would have followed similar trends as in other countries (Olden and Møen, 2022). We provide support for this assumption by testing for pre-trends in Figure 6. We find that relative platform usage in Germany followed similar trends, before the NetzDG, when compared to the other countries in our sample. Note that this figure begins in 2017q1 as these are the earliest quarters for which Semrush web traffic data exist. After the passage of the NetzDG from 2017q4 onwards, we find overall significantly positive estimates for the usage of the moderated platforms in Germany.

Figure 6: The Effect of the NetzDG on Platform Usage



Notes: This figure plots coefficients from event study versions of Equation (2). The dependent variable is the log number of users. The omitted category is the 3rd quarter of 2017, the quarter before the passing of the NetzDG (indicated with the vertical line). The whiskers indicate 95% confidence intervals based on standard errors clustered by country.

These findings are confirmed by the regression estimates in Table 2. The estimates suggest that the quarterly usage of moderated platforms increased by 8% when measured by the number of users (columns (1) and (2)) and by 9% based on the total number of visits (columns (3) and (4)). The estimates remain unchanged when we include additional interacted fixed effects (columns (2) and (4)).

Table 2: Regression Estimates: NetzDG and Platform Usage

	Dep. var.: Log(Users)		*Dep. var.: Log(Visits)*	
	(1)	(2)	(3)	(4)
Germany × Platform × Post	0.081***	0.081***	0.089***	0.089***
	(0.022)	(0.022)	(0.029)	(0.029)
Country FE	Yes		Yes	
Year-Quarter FE	Yes		Yes	
Platform FE	Yes		Yes	
Country × Year-Quarter FE		Yes		Yes
Platform × Year-Quarter FE		Yes		Yes
Country × Platform FE		Yes		Yes
Observations	3,024	3,024	3,024	3,024
Pre-Period Mean of DV	15.86	15.86	17.62	17.62
R^2	0.93	0.99	0.92	0.99

Notes: This table presents the results of estimating Equation (2) where the dependent variable is the log number of unique users or the log of the total visits to a website. *Germany* is an indicator variable equal to 1 for Germany, and 0 otherwise. $Platform$ is an indicator equal to 1 for the platforms targeted by the NetzDG (Instagram, Twitter, YouTube, and Facebook), and 0 for those that were not (Amazon, Netflix, Wikipedia). *Post* is a dummy variable equal to 1 for observations after 2017q3. Standard errors in parentheses are clustered at the country level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Why did the NetzDG result in an increase in traffic to these websites? The persistent effects in Figure 6 suggest that the pattern is unlikely to be explained by a temporary increase in the salience or popularity of the treated platforms due to the passage of the law. Instead, these findings are consistent with an increase in the demand for these websites, suggesting that most users prefer a marginal increase in moderation. These results are consistent with the evidence in Jiménez Durán (2022), who documents a positive effect of reporting hate speech on the engagement of users who are attacked by hateful posts. Appendix Figure C.4 provides further estimates at the platform level. Platforms with a higher reliance in user reports for content moderation (Instagram and Twitter) are also the ones where the effect is stronger.[24]

[24]As argued by Jiménez Durán (2022), this pattern is to be expected, given that platforms moderate up to a point where the marginal benefit (an increase in user engagement) equals marginal cost. We expect platforms with a higher reliance on user reports to have a higher marginal cost because they likely rely more on human reviewers than platforms that proactively remove content (which typically rely heavily on automated systems). See, for example, Meta's proactive detection strategy: `https://transparency.fb.com/policies/improving/proactive-rate-metric/`. In 2019q3 (the earliest for which there is data), 55.9% of content violating Instagram's rules was found through user reports. This compares to 19.3% for Facebook (`https://transparency.fb.com/reports/community-standards-enforcement/hate-speech`). On YouTube, non-automated video removals amounted to 20.4% of removals (but this figure also includes offenses other than hate speech, see `https://transparencyreport.google.com/youtube-policy/removals`). In contrast, in 2020, close to half

5 Offline Effects of the NetzDG

The second part of the paper investigates the offline effects of the NetzDG. The analysis proceeds in three steps. First, we study if the NetzDG-induced decrease in online toxicity also led to a reduction in the prevalence of anti-refugee hate crimes. Second, we investigate two potential mechanisms that could explain changes in the anti-refugee incidents. Lastly, we analyze the impact of the NetzDG on overall hate crimes in Germany using a cross-county synthetic control design.

5.1 Did the NetzDG Reduce the Number of Anti-Refugee Hate Crimes?

To estimate the effect of the NetzDG on anti-refugee hate crimes, we exploit variation in the exposure of different German municipalities to anti-refugee content. Intuitively, we expect places with a higher exposure to this type of content to be disproportionately affected by the NetzDG relative to places with a lower exposure.

Empirical Strategy

This intuition gives rise to the following empirical strategy:

$$y_{it} = \theta \cdot AfD\ Users\ p.c._i \times Post\ NetzDG_t + \mathbf{X_{it}'}\beta + \gamma_i + \delta_t + \epsilon_{it}, \tag{3}$$

where our main outcome of interest, y_{it}, is the inverse hyperbolic sine of the number of anti-refugee incidents in municipality i in quarter t.[25] The main independent variable is the interaction between the number of AfD Facebook users per capita ($AfD\ Users\ p.c._i$) and a time dummy ($Post\ NetzDG_t$) which is equal to one for the period starting in 2017q4 when the NetzDG became law. The regression includes a full set of municipality and time fixed effects. The municipality fixed effects control for any baseline difference in the number of anti-refugee incidents (e.g., due to the higher presence of refugees), while the time fixed effects account for any Germany-wide change in the number of anti-refugee incidents (e.g., due to national news events).

Table A.5 plots the mean and standard deviation of a large number of municipality characteristics by quartiles of our exposure variable, $AfD\ Users\ p.c._i$. More exposed municipalities tend to be somewhat larger and more likely to vote for the AfD, Linke, or

of violating content on Twitter was flagged by humans (`https://www.fastcompany.com/90528941/twitter-automatically-flags-more-than-half-of-all-tweets-that-violate-its-rules`).

[25] In Appendix Table D.2, we show that the results are robust to other variable transformations.

Green party, but these differences are quantitatively small. To control for potential other drivers of trends in hate crimes over time, the vector ($\mathbf{X_{it}}$) includes control variables, which we also interact with the $Post\ NetzDG_t$ dummy. We cluster standard errors at the county level.[26]

As is standard for difference-in-differences designs, our identifying assumption is that in the absence of the NetzDG, municipalities with different prior exposures to hate speech on social media would have experienced a similar trend in hate crimes. The coefficient θ, therefore, measures the extent to which the NetzDG was associated with a differential change in the number of anti-refugee incidents in municipalities with a higher exposure to anti-refugee content on Facebook.

Results

Table 3 shows our main results. Column (1) contains estimates of our baseline specification using Equation (3), controlling only for log population interacted with the *Post* indicator to control for any changes in hate crimes due to population differences. In the following columns, we add controls for potential confounders. Throughout the different specifications, the point estimates remain stable and indicate that a one standard deviation increase in AfD Facebook users per capita results in a 1% (relative) reduction in quarterly hate crimes. As a benchmark, Müller and Schwarz (2021) find that a one standard deviation increase in AfD Facebook users per capita is associated with a 10% higher probability of a weekly anti-refugee incident relative to the mean. Our estimate on the effect of the NetzDG seems plausible given the 5% reduction in hateful online content we identified for AfD users in the previous section.

In column (2), we control for the vote share of the AfD and all other major parties at the municipality level. These controls account for any change in anti-refugee incidents around the time of the NetzDG that can be explained by the political leaning of a municipality. We find that the coefficient for the AfD vote share is positive and significant. This result highlights the clear distinction between offline support for the AfD and online exposure to hateful content, the former of which is unaffected by the NetzDG. Controlling for the AfD vote share further allows us to mitigate concerns about many other contemporaneous shocks. The reason is that shocks other than the NetzDG that may disproportionately reduce anti-refugee attacks in right-leaning areas should affect AfD voters similarly to AfD Facebook users. Our results instead point toward the importance of an online channel.

[26] In Appendix Table D.3, we show robustness for alternative levels of clustering.

Column (3) adds Facebook users per capita in a municipality as a control to account for any changes in anti-refugee incidents that could be explained by unobservable confounders that correlate with a municipality's affinity to social media. In a similar spirit, we add a control for the access to broadband internet in column (4), which is our preferred specification, since it controls for population, voting, and media consumption behavior. The coefficients on Facebook users per capita and broadband internet access are small and statistically indistinguishable from 0. In other words, after accounting for the exposure to far-right Facebook usage, a town's social media or internet penetration does not matter for its elasticity with respect to the NetzDG. This finding suggests that, in line with our hypothesis and the evidence in the first part of the paper, the NetzDG mattered for people who were exposed to anti-refugee content instead of the effects being driven by access to social media or the internet. Finally, in column (5), we include a wealth of additional control variables (see Appendix A for details), all of which we interact with the *Post* indicator. The inclusion of these 19 additional control variables again has little impact on the magnitude, sign, and statistical significance of our main estimate.

Event study

Figure 7 visualizes the coefficients from an event study version of regression Equation (3), with 2017q3 (the quarter before the NetzDG became law) as the excluded period. We find no evidence for pre-existing trends in this specification. The pre-period coefficients are statistically insignificant and close to 0. We only observe a statistically significant reduction in the number of anti-refugee incidents after the increase of content moderation efforts in 2017q4. Moreover, this negative effect appears to be persistent and stable over the two years following the NetzDG.

Alternative Explanations

As with any difference in difference estimate, identification requires the absence of other contemporaneous shock that differentially affects areas with many AfD Users. Two possible candidates for such shocks could be 1) the end of the refugee crisis, and 2) the 2017 federal election. We discuss these events in turn.

First, our findings cannot be easily explained by some form of mean reversion in the number of anti-refugee incidents due to the end of the refugee crisis in Germany. As discussed in Section 2, the inflow of refugees to Germany had already stopped in March 2016 when the EU struck a deal with Turkey to prevent the further entry of refugees

Table 3: Effect of NetzDG on Anti-Refugee Hate Crime

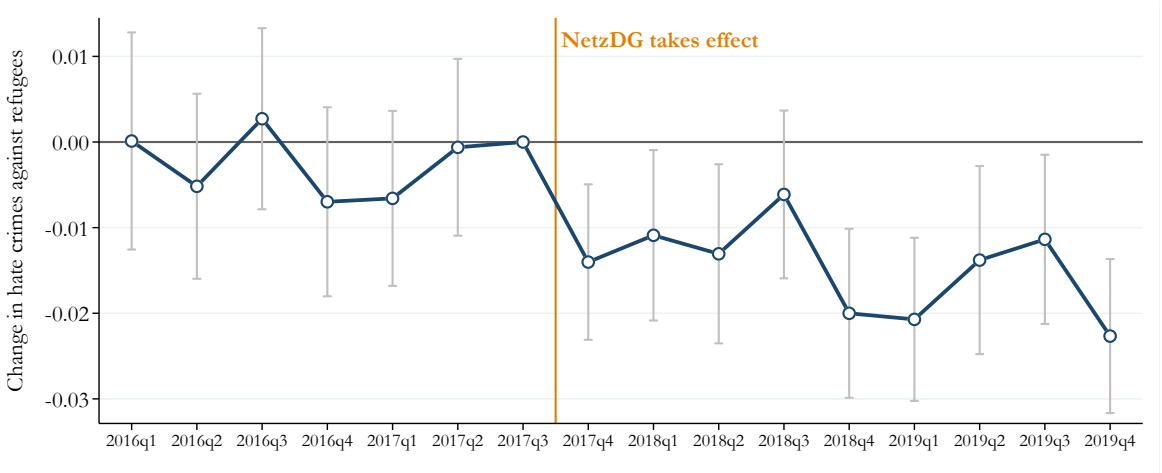
	Dep. var.: Asinh(Anti-Refugee Hate Crimes)				
	(1)	(2)	(3)	(4)	(5)
AfD Facebook users p.c. (std) × Post	-0.012***	-0.009***	-0.009***	-0.009***	-0.008***
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
AfD vote share (std) × Post		0.034***	0.034***	0.036***	0.031***
		(0.012)	(0.012)	(0.012)	(0.012)
Facebook users p.c (std) × Post			0.002	0.002	0.003
			(0.002)	(0.002)	(0.002)
Broadband internet (std) × Post				0.005	0.001
				(0.003)	(0.004)
Municipality FE	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes
Ln(Pop.) × Post	Yes	Yes	Yes	Yes	Yes
Election Controls × Post		Yes	Yes	Yes	Yes
All Controls (19) × Post					Yes
Observations	71,456	71,008	71,008	71,008	68,736
Pre-Period Mean of DV	0.12	0.12	0.12	0.12	0.12
R^2	0.44	0.44	0.44	0.44	0.45

Notes: This table presents the results of estimating Equation (3), where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes in a municipality in a given quarter. *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. All regressions include municipality and quarter fixed effects as well as a control for the natural logarithm of population, interacted with *Post*. See text for a detailed description of the additional control variables. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

from Syria to Europe. Therefore, the total number of refugees was nearly constant around the introduction of the NetzDG. Further, the effect we find occurs over a year after this important demarcation point of the refugee crisis. It is further worth noting that the exposure measure is largely uncorrelated with the number of refugees (see Appendix Table A.5). As a result, including the municipality number of refugees as a control does not change the estimates (see column 5 Table 3). Moreover, any such mean reversion should also affect municipalities with many AfD voters in a similar way, which is rejected by the estimated positive coefficient on the AfD vote share.

Second, the 2017 federal election is unlikely to drive our findings because we include controls for the electoral results of all major German parties in our regressions. The inclusion of these variables makes hardly any difference to the magnitude and significance of our coefficient of interest. Moreover, the positive coefficient for the AfD vote share contradicts the idea that the end of the election period was associated with a drop in the number of anti-refugee incidents. Instead, these results suggest that the unexpectedly strong showing of the AfD in the 2017 federal elections, where it became

Figure 7: Event Study Hate Crime



Notes: This figure plots the coefficients from running an event study version of regression Equation (3). The dependent variable is the inverse hyperbolic sine of the number of anti-refugee incidents. The omitted category is the 3rd quarter of 2017, the quarter before the passing of the NetzDG (indicated with the vertical line). The whiskers indicate 95% confidence intervals based on standard errors clustered by county.

the third strongest party and the first far-right party in the German parliament since 1945, may have emboldened its supporters. This hypothesis meshes well with evidence in Bursztyn et al. (2020), who document that the election of Donald Trump in the United States increased people's willingness to publicly express xenophobic views.

Finally, our results effectively exploit cross-sectional exposure in the residual variation in AfD Facebook usage that is not explained by either AfD or Facebook affinity. This strategy makes it unlikely that any other event that may have occurred contemporaneously with the passage of the NetzDG biases our estimates. In order for such an event to be a potential confounder, it would have to simultaneously reduce anti-refugee incidents in municipalities with many AfD Facebook users but at the same increase anti-refugee incidents in municipalities with AfD voters *and* leave municipalities with many Facebook users unaffected.

Robustness

To further probe the robustness of our findings, we perform additional robustness checks. First, Online Appendix Table D.1 shows that, with the exception of cases of arson (which are rare), the NetzDG affected all categories of anti-refugee incidents (i.e., assault, demonstration, suspected attacks, and other miscellaneous property attacks). The strongest response is for assaults and other property attacks. The effect on severe

incidents such as assaults, which seem difficult to "fake", also makes it less likely that the estimates capture differential changes in reporting incidents. Besides, any overall change in the reporting of incidents would be absorbed by the time fixed effects. It is also worth noting that the passage of the NetzDG and the surrounding debate on hate crime should, if anything, make it more likely for victims to report incidents.

Table 4 presents a battery of additional robustness exercises. Column (2) shows robustness to the inclusion of federal state × quarter fixed effects (see column (2)). This specification exploits variation within the same federal state at the same point in time, and hence accounts for any potential changes in law enforcement that might have been introduced by the state governments. These fixed effects will also absorb any differential shock that might affect a specific federal state (e.g., local elections). Column (3) excludes January and February 2016 from the data, which constitute the largest spike in anti-refugee incidents. This exclusion leaves the estimates unchanged and highlights that the findings are not driven by these outliers in the number of incidents. Similarly, the findings are robust to excluding municipalities without anti-refugee incidents, without AfD users, or with few refugees per capita (columns (4), (5), and (6), respectively). Throughout these exercises, the estimates remain statistically significant, making it unlikely that they are driven by zeroes in the dependent or independent variables.

Table 4: Robustness Tests

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)					
	Baseline	Federal State × Quarter FE	Exclude Q1 2016	Exclude Attack= 0	Exclude AfD User= 0	Exclude Few Refugees
	(1)	(2)	(3)	(4)	(5)	(6)
AfD Facebook users p.c. (std) × Post	-0.009***	-0.008***	-0.009***	-0.016***	-0.009***	-0.016***
	(0.002)	(0.002)	(0.002)	(0.005)	(0.003)	(0.004)
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Fed. State × Quarter FE		Yes				
Ln(Pop/) × Post	Yes	Yes	Yes	Yes	Yes	Yes
AfD vote share × Post	Yes	Yes	Yes	Yes	Yes	Yes
Election Controls × Post	Yes	Yes	Yes	Yes	Yes	Yes
Facebook users p.c × Post	Yes	Yes	Yes	Yes	Yes	Yes
Broadband internet × Post	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71,008	71,008	66,570	36,384	64,736	56,656
Pre-Period Mean of DV	0.12	0.12	0.10	0.23	0.12	0.14
R^2	0.44	0.45	0.45	0.42	0.44	0.46

Notes: This table presents the results of estimating municipality-quarter-level regressions as in Equation (3) where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes. *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. All regressions include municipality and quarter fixed effects, as well as controls for the logarithm of population, the AfD vote share, Facebook users per capita, and broadband internet access, all interacted with *Post*. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Third, Appendix Table D.2 shows that the estimates are robust independently of the functional form of the dependent and independent variables. In particular, we explore transformations of the dependent variable (refugee attacks) in inverse hyperbolic sine (baseline), counts, or the log number of refugee incidents per capita. Neither of these changes alter our findings (see columns (1-3)). Columns (4-6) then replace the main independent variable with an indicator of whether a municipality has an above-median number of AfD users per capita. This exercise serves three purposes. First, it allows us to rule out concerns about outliers in the number of AfD users per capita. Second, this dummy specification does not rely on functional form assumptions, because it simply picks up changes in the mean number of anti-refugee incidents after the NetzDG in a canonical difference-in-differences setting. Third, this transformation also alleviates concerns that the findings are driven by heterogeneous treatment effects in the two-way fixed effects estimation (De Chaisemartin and d'Haultfoeuille, 2023), as our results also hold in this dummy specification.

Fourth, we repeat the analysis based on the number of AfD Twitter followers (as opposed to Facebook followers) in a municipality. Appendix Table D.4 shows that the results are virtually identical with this alternative measure of exposure to the NetzDG. Appendix Figure D.2 also presents the corresponding event study estimates.

Fifth, we perform a leave-one-out analysis excluding one municipality at a time. The results are shown in Appendix Figure D.1. The estimates are highly stable throughout. As such, our findings do not appear to be driven by outliers or any particular municipality.

Finally, Appendix Table D.3 shows that the estimates remain statistically significant irrespective of the level of clustering of the standard errors. Specifically, the results are similar when standard errors are clustered at 1) the county level (baseline), 2) the county and quarter level, 3) the municipality level, or 4) the municipality and quarter level.

Consumers vs Producers of Online Hate

The NetzDG could affect the prevalence of anti-refugee incidents by changing the willingness of either the consumers or producers of anti-refugee online content to commit acts of violence against refugees. We examine these hypotheses by investigating heterogeneity in our estimates depending on the amount of content posted on the AfD's Facebook page.

If the effect we document is driven by the presence of producers of anti-refugee posts in a municipality, the impact of the NetzDG should be stronger in such areas. Table 5 explores this possibility by including different measures of content production

in the regressions. In particular, we measure "production" using the average number of posts, comments, likes, and shares sent by each AfD user in a given municipality before the passing of the NetzDG. Note that these regressions are only estimated for municipalities for which we can identify at least one AfD user. The results suggest that the effect of the NetzDG is stronger in municipalities in which users were more actively producing content on the AfD's Facebook page. This holds independent of the measure of usage intensity. The coefficient in column (1) suggests that a one standard deviation increase in the number of posts per AfD user is associated with an additional 0.5 percentage point reduction in the number of anti-refugee hate crimes.

Table 5: Heterogeneity by User Activity

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)			
	(1)	(2)	(3)	(4)
AfD Facebook users p.c. (std) × Post	-0.022***	-0.022***	-0.022***	-0.022***
	(0.004)	(0.004)	(0.004)	(0.004)
Post per AfD User (std) × Post	-0.005***			
	(0.001)			
Likes per AfD User (std) × Post		-0.005***		
		(0.001)		
Comments per AfD User (std) × Post			-0.004***	
			(0.001)	
Shares per AfD User (std) × Post				-0.004***
				(0.002)
Municipality FE	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes
Ln(Pop.) × Post	Yes	Yes	Yes	Yes
Observations	57,008	57,008	57,008	57,008
Pre-Period Mean of DV	0.14	0.14	0.14	0.14
R^2	0.45	0.45	0.45	0.45

Notes: This table presents the results from estimating Equation (3) for municipalities with at least one AfD Facebook user. The dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes in a municipality and quarter. *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. We additionally include different measures of Facebook activity per AfD user before the NetzDG in regressions, also standardized to have a mean of 0 and a standard deviation of 1. All regressions include municipality and quarter fixed effects, as well as a control for the logarithm of population interacted with *Post*. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

In line with the idea that the consumers of hateful online posts matter independently, we find that the effect of the number of AfD Facebook users per capita in a municipality (first row) has predictive power over and above the presence of local producers, highlighting the importance of passive exposure. This pattern holds for all measures of engagement, such as likes, comments, or shares. It is also worth noting

that, in the subset of municipalities with at least one AfD user, the effect of the NetzDG is stronger than in our baseline specification. Together, these results suggest that the NetzDG affected the likelihood of committing anti-refugee acts for both consumers and producers of anti-refugee content.

5.2 Potential Mechanisms

We now shed light on two different mechanisms that may partially explain why an increase in content moderation could reduce hate crimes. First, we analyze whether the NetzDG made it more difficult for potential perpetrators of anti-refugee incidents to coordinate online. Second, we study whether the NetzDG influenced attitudes towards refugees.

Did the NetzDG Affect Collective Action?

Following the literature on the effects of social media on collective action (e.g., Enikolopov et al., 2020; Manacorda and Tesei, 2020; Fergusson and Molina, 2021), we investigate whether the NetzDG was able to interrupt the ability of potential perpetrators to coordinate anti-refugee incidents. As an example, the NetzDG could make it harder to learn about the willingness of others to carry out acts of violence against refugees. Recall from the result in column (6) of Table C.6 that the NetzDG made the tone of tweets less threatening. To examine a potential coordination mechanism, we rerun our main analysis but split anti-refugee incidents based on the number of perpetrators. Note that we could only hand-code the number of perpetrators for 9% of the anti-refugee incidents in our data, which leads to mechanically smaller coefficients in these regressions.

Table 6 presents the results from this analysis. Panel (a) shows the results for anti-refugee incidents with a single perpetrator, whereas Panel (b) shows the estimates for multiple perpetrators. While the estimates in both panels are statistically significant, the effect of the NetzDG on incidents with multiple perpetrators is, in all cases, twice as large as for incidents with a single perpetrator. These findings highlight the "social" component of anti-refugee incidents and suggest social media may help facilitate collecting action (in this case, violent attacks on refugees). This evidence is also in line with previous work by Müller and Schwarz (2021), who document a stronger effect of social media on hate crimes with multiple perpetrators, as well as the evidence in Bursztyn et al. (2019) of a coordination mechanism of social networks in the Russian context.

Table 6: Regression Estimates: Effect of NetzDG on Hate Crime

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)				
	(1)	(2)	(3)	(4)	(5)
Panel (a): Single Perpetrators					
AfD Facebook users p.c. (std) × Post	-0.002***	-0.001**	-0.001**	-0.001**	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	71,456	71,008	71,008	71,008	68,736
Pre-Period Mean of DV	0.00	0.00	0.00	0.00	0.00
R^2	0.14	0.14	0.14	0.14	0.14
Panel (b): Multiple Perpetrators					
AfD Facebook users p.c. (std) × Post	-0.004***	-0.002**	-0.002**	-0.002**	-0.002**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Municipality FE	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes
Ln(Pop/) × Post	Yes	Yes	Yes	Yes	Yes
AfD vote share × Post		Yes	Yes	Yes	Yes
Election Controls × Post		Yes	Yes	Yes	Yes
Facebook users p.c × Post			Yes	Yes	Yes
Broadband internet × Post				Yes	Yes
All Controls (19) × Post					Yes
Observations	71,456	71,008	71,008	71,008	68,736
Pre-Period Mean of DV	0.01	0.01	0.01	0.01	0.01
R^2	0.20	0.20	0.20	0.20	0.21

Notes: This table presents the results of estimating Equation (3), where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes in a municipality in a given quarter. *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. All regressions include municipality and quarter fixed effects as well as a control for the natural logarithm of population, interacted with *Post*. See text for a detailed description of the additional control variables. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Did the NetzDG affect Attitudes toward Refugees?

The NetzDG may also have decreased hate crimes because it changed attitudes toward refugees, for example by reducing animus of social media users towards refugees. To examine this idea, we use data from the German Socio-Economic Panel (GSOEP) (Goebel et al., 2019). Specifically, we exploit the panel nature of GSOEP to study within-person changes in attitudes towards refugees using a regression of the following form:

$$y_{it} = \theta \cdot Social\ Media\ User \times Post\ NetzDG_t + \mathbf{X_{it}'}\beta + \gamma_i + \delta_t + \epsilon_{it}, \tag{4}$$

where y_{it} is the response to one of five questions on the impacts of refugees on the economy and society. As described above, we recode these questions into indicator variables such that 1 represents a positive attitude towards refugees. We additionally create an index capturing the average response of a respondent across these five questions. $Social\ Media\ User$ is an indicator for respondents who use social media at least once a week. $Post\ NetzDG_t$ is a dummy for the period after the NetzDG. As the data only contains two survey waves, $Post\ NetzDG_t$ is 1 for the year 2018 and 0 for 2016. $\mathbf{X_{it}}$ contains controls for the gender and age of the respondents, which we interact with the $Post$ indicator. Finally, γ_i and δ_t are full sets of respondent and survey wave fixed effects. As a result, β measures whether respondents who use social media developed more positive attitudes towards refugees between 2016 and 2018 relative to respondents who did not use social media.

Table 7: Changes in Attitudes Towards Refugees

	Dep. var.: Refugees are ...					
		Positive for the			A Chance in the	
	Index	Economy	Culture	Place of Living	Short-term	Long-term
	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a): All Respondents						
Social Media User × Post	-0.008	0.001	-0.011	0.001	-0.019**	-0.010
	(0.006)	(0.009)	(0.009)	(0.008)	(0.009)	(0.009)
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Survey Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	36,698	36,296	36,280	36,254	36,268	36,144
Pre-Period Mean of DV	0.50	0.60	0.57	0.53	0.24	0.54
R^2	0.82	0.73	0.75	0.75	0.66	0.74
Panel (b): AfD Voters						
Social Media User × Post	-0.002	0.029	-0.034	-0.009	0.004	0.004
	(0.017)	(0.034)	(0.030)	(0.024)	(0.019)	(0.026)
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes
Survey Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,566	2,550	2,558	2,548	2,554	2,550
Pre-Period Mean of DV	0.16	0.25	0.18	0.15	0.06	0.16
R^2	0.71	0.64	0.64	0.66	0.55	0.66

Notes: This table presents the results of estimating Equation (4), where the dependent variables are different measures for positive attitudes towards refugees. *Social Media Users* is an indicator for respondents who use social media at least once a week. All regressions include individual and survey year fixed effects as well as a control for the respondent's gender and age, interacted with *Post*. See the text for a detailed description of the variables. Robust standard errors in parentheses are clustered by individual. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 7 plots the results. Overall, there is no evidence for positive changes in attitudes for the period after the NetzDG. All estimates are small and, except one, statistically insignificant. The only significant estimate is *negative*, which would reject the hypothesis that the NetzDG improved attitudes towards refugees. These null results hold both for all GSOEP respondents (Panel a) as well as respondents who express support for the AfD (Panel b). As an additional test, we also investigate changes in pro-refugee *actions* (as opposed to opinions) in Appendix Table D.5. The estimates are again mostly small and, if anything, negative. Taken together, these findings provide evidence against the idea that the NetzDG has caused a reduction in anti-refugee incidents primarily by changing attitudes towards refugees.

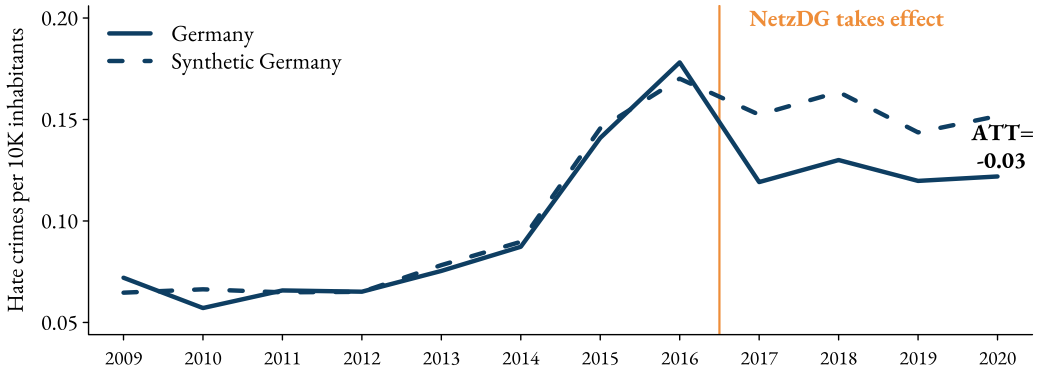
5.3 Synthetic Control Estimates

As the last piece of our analysis, we provide additional evidence for the offline effects of the NetzDG based on synthetic control estimates. This serves two purposes. First, it lends further credence to our main findings using a completely different data source and empirical strategy. Second, the synthetic control estimates allow us to investigate the effect on the total (non-refugee related) number of hate crimes in Germany, which are only available at the country-year level. More specifically, we build a synthetic control group for Germany using data from 21 donor countries from the OSCE, following the methodology of Abadie and Gardeazabal (2003) and Abadie et al. (2010). The dependent variable is the yearly number of hate crimes per 10,000 inhabitants, and we use as predictors the full path of lagged outcomes, as recommended by Ferman et al. (2020). Because some of the donor countries changed their data collection in the pre-period, we add as a predictor an indicator of whether there was a change in measurement. Because the NetzDG became law in the fourth quarter of 2017, we define 2017 as the treatment year. This approach is more conservative than using 2018 as the treatment year since backdating the intervention does not mechanically bias the estimator (Abadie, 2021).

Figure 8 reports the main estimates from this exercise. This figure shows that the number of hate crimes per 10,000 inhabitants in the synthetic Germany based on the 21 donor countries by construction closely tracks the observed hate crimes until the year the NetzDG was enacted. After the NetzDG goes into force, we find a drop in the number of hate crimes relative to the synthetic control. The average treatment effect (ATE) in the 2018-2020 post-period is -0.0301 hate crimes per 10,000 inhabitants, or 250 fewer hate crimes per year. Appendix D.3. presents additional information, such

as the weights used to construct the synthetic control (Table D.6) and the pre-period balance of the predictors (Table D.7).[27]

Figure 8: Evolution of Hate Crimes in Germany vs. Synthetic Germany



Notes: This figure presents the evolution of hate crimes per 10,000 inhabitants in Germany and a synthetic Germany. The synthetic control uses all lagged outcomes as predictors, as well as a dummy variable indicating whether there were measurement changes in the pre-period.

We can reject the null hypothesis that the ATT is non-negative with a p-value of 0.045, constructed based on in-space placebo tests as in Abadie et al. (2010), where "placebo effects" are computed assuming that each of the donor countries is treated. Intuitively, this exercise shows that the magnitude of the treatment effect in Germany is an outlier relative to the placebo effects estimated among countries in the donor pool. Figure D.5 provides visual evidence of this intuition by plotting the histograms of the (one-sided) ratio of the mean square predicted errors (MSPE) after vs. before the NetzDG in Germany and in the donor countries.

Table D.8 shows that this result is robust to a battery of additional checks. First, we investigate alternative ways of dealing with missing data (no interpolation or including a dummy for interpolated values). Second, we explore alternative transformations of the outcome variable (logarithm and levels). Third, we consider alternative end dates, which result in a different donor pool based on differences in data availability. Fourth, we consider alternative sets of donor countries (leaving out donor countries or restricting our estimates to OECD members). Overall, the results remain similar throughout these robustness checks, suggesting that the NetzDG contributed to reducing the aggregate number of hate crimes in Germany relative to other countries.

[27]The weights overall seem intuitive, with countries like Poland, Italy, and Austria receiving a large weight (close to 10% each). The one outlier is the large weight of Lithuania (55%). In robustness checks in Table D.8, we confirm that our results do not change when we remove Lithuania from the data.

As a placebo exercise, we replicate the estimation on the number of homicides per 10,000 inhabitants, based on the assumption that it is unlikely that the NetzDG impacted the overall homicide rate. If our results were driven by other policies implemented by Germany that coincided with the NetzDG and also impacted hate crimes, such as a change in law enforcement, we would also expect to see an effect on this outcome variable. As Figure D.6 shows, there is no evidence of an effect on homicides; the estimate is positive in some years and negative in others. The effect size is only one-fourth of the estimate for hate crimes (relative to the level of pre-treatment outcomes). Moreover, as opposed to our estimates on hate crimes, the magnitude of the effect is small compared to the placebo effect on the donor countries, which is reflected in a p-value of 0.44.

6 Discussion

Much attention has been devoted to the spread of hateful content on social media. The controversial German NetzDG was in large part a reaction to the prevalence of hateful messages on social media platforms and the perceived limited effort of these platforms to moderate this content. By leveraging this unique quasi-experiment, this paper is the first to show that content moderation, induced by regulation, can indeed achieve its primary aim of reducing hateful sentiments online and decreasing the incidence of hate crimes against minorities offline.

While reducing hate is undoubtedly an important aim, we want to caution against taking this finding as blanket support for content moderation. This study does not and cannot evaluate the full schedule of costs and benefits of online censorship and its potential impact on legitimate online debate. For example, one of the main reasons why the NetzDG has been controversial is its potential misuse to undermine freedom of expression and stifle political dissent (Kaye, 2018). We do not find evidence that the law increased online discussions of censorship or that users disengaged from controversial political issues. However, Figure A.3 shows that an expert-opinion-based index that measures freedom of expression in Germany decreased after the passage of the law, moving it from third place in 2016 (between Switzerland and Belgium) to seventeenth place in 2020 (between New Zealand and Uruguay). While it is unclear whether this decrease is driven by the NetzDG, it highlights the need for more research to understand the effects of this law on freedom of expression and offline political discussion. As such, we believe our findings should best be interpreted as a valuable starting point for understanding the online and offline effects of content moderation on social media.

References

Abadie, A. (2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature 59*(2), 391–425.

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association 105*(490), 493–505.

Abadie, A. and J. Gardeazabal (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review 93*(1), 113–132.

Acemoglu, D., T. A. Hassan, and A. Tahoun (2017, 08). The Power of the Street: Evidence from Egypt's Arab Spring. *The Review of Financial Studies 31*(1), 1–42.

Adena, M., R. Enikolopov, M. Petrova, V. Santarosa, and E. Zhuravskaya (2015). Radio and the Rise of The Nazis in Prewar Germany. *The Quarterly Journal of Economics 130*(4), 1885–1939.

Allcott, H., L. Braghieri, S. Eichmeyer, and M. Gentzkow (2020, March). The Welfare Effects of Social Media. *American Economic Review 110*(3), 629–76.

Allcott, H. and M. Gentzkow (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives 31*(2), 211–36.

Ambrosino, A., M. Cedrini, J. B. Davis, S. Fiori, M. Guerzoni, and M. Nuccio (2018). What topic modeling could reveal about the evolution of economics. *Journal of Economic Methodology 25*(4), 329–348.

Andres, R. and O. Slivko (2021). Combating Online Hate Speech: The Impact of Legislation on Twitter. *ZEW-Centre for European Economic Research Discussion Paper* (21-103).

Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Anti-Defamation League (2022). Online Hate and Harassment. The American Experience 2022. *Center for Technology and Society*. Accessed: 2022-09-11.

Aridor, G., R. Jiménez Durán, R. Levy, and L. Song (2024). The economics of social media. *Available at SSRN*.

Ash, E. and S. Hansen (2023). Text algorithms in economics. *Annual Review of Economics 15*, 659–688.

Barrera, O., S. Guriev, E. Henry, and E. Zhuravskaya (2020). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of Public Economics 182*, 104123.

Beknazar-Yuzbashev, G., R. Jiménez Durán, J. McCrosky, and M. Stalinski (2022). Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN*.

Beknazar-Yuzbashev, G., R. Jiménez Durán, and M. Stalinski (2024). A model of harmful yet engaging content on social media. *Available at SSRN*.

Bhuller, M., T. Havnes, E. Leuven, and M. Mogstad (2013). Broadband Internet: An Information Superhighway to Sex Crime? *Review of Economic Studies 80*(4), 1237–1266.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*(Jan), 993–1022.

Bond, R. M., C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler (2012). A 61-Million-Person Experiment in Social Influence and Political Mobilization. *Nature 489*(7415), 295.

Boxell, L., M. Gentzkow, and J. M. Shapiro (2017). Greater Internet Use Is Not Associated with Faster Growth in Political Polarization Among US Demographic Groups. *Proceedings of the National Academy of Sciences of the United States of America*, 201706588.

Braghieri, L., R. Levy, and A. Makarin (2022). Social Media and Mental Health.

Bundeskartellamt (2019). Bundeskartellamt Prohibits Facebook From Combining User Data From Different Sources. `https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Pressemitteilungen/2019/07_02_2019_Facebook.html`. Accessed: 2022-07-14.

Bursztyn, L., G. Egorov, R. Enikolopov, and M. Petrova (2019). Social Media and Xenophobia: Evidence from Russia. Working Paper 26567, National Bureau of Economic Research.

Bursztyn, L., G. Egorov, and S. Fiorin (2020). From Extreme to Mainstream: The Erosion of Social Norms. *American Economic Review 110*(11), 3522–48.

Campello, R. J., D. Moulavi, and J. Sander (2013). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172. Springer.

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

Card, D. and G. B. Dahl (2011). Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior. *The Quarterly Journal of Economics 126*(1),

103–143.

Chen, Y. and D. Y. Yang (2019). The Impact of Media Censorship: 1984 or Brave New World? *American Economic Review 109*(6), 2294–2332.

Dahl, G. and S. DellaVigna (2009). Does Movie Violence Increase Violent Crime? *The Quarterly Journal of Economics*, 677–734.

De Chaisemartin, C. and X. d'Haultfoeuille (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal 26*(3), C1–C30.

DellaVigna, S., R. Enikolopov, V. Mironova, M. Petrova, and E. Zhuravskaya (2014, July). Cross-Border Media and Nationalism: Evidence from Serbian Radio in Croatia. *American Economic Journal: Applied Economics 6*(3), 103–32.

Deutscher Bundestag (2017). Drucksache 18/12356. `https://dserver.bundestag.de/btd/18/123/1812356.pdf`. Accessed: 2022-08-04.

Dixon, L., J. Li, J. Sorensen, N. Thain, and L. Vasserman (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73.

Djourelova, M. (2023). Persuasion through slanted language: Evidence from the media coverage of immigration. *American Economic Review 113*(3), 800–835.

Draca, M. and C. Schwarz (2024). How polarized are citizens? measuring ideology from the ground-up. *Forthcoming Economic Journal*.

Du, X. (2023). Symptom or Culprit? Social Media, Air Pollution, and Violence.

Echikson, W. and O. Knodt (2022). Germany's NetzDG: A Key Test for Combatting Online Hate. *Available at SSRN: https://ssrn.com/abstract=3300636*.

Economist (2018). In Germany, Online Hate Speech Has Real-World Consequences.

ElSherief, M., V. Kulkarni, D. Nguyen, W. Y. Wang, and E. 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*, Volume 12.

Enikolopov, R., A. Makarin, and M. Petrova (2020). Social Media and Protest Participation: Evidence from Russia. *Econometrica 88*(4), 1479–1514.

Enikolopov, R., M. Petrova, and K. Sonin (2018). Social media and corruption. *American Economic Journal: Applied Economics 10*(1), 150–174.

Fergusson, L. and C. Molina (2021, April). Facebook Causes Protests. Documentos CEDE 018002, Universidad de los Andes - CEDE.

Ferman, B., C. Pinto, and V. Possebom (2020). Cherry picking with synthetic controls. *Journal of Policy Analysis and Management 39*(2), 510–532.

Frankfurter Rundschau (2019). Fall Susanna F. in Wiesbaden: Wie rechte Stimmungsmache mit Flüchtlingskriminalität funktioniert. *https://www.fr.de/rhein-main/soziale-medien-helfen-rechten-kampagnen-11847580.html*.

Fujiwara, T., K. Müller, and C. Schwarz (2023). The Effect of Social Media on Elections: Evidence from the United States. *Journal of the European Economic Association*.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as data. *Journal of Economic Literature 57*(3), 535–574.

Gillespie, T. (2018). *Custodians of the Internet*. Yale University Press.

Goebel, J., M. M. Grabka, S. Liebig, M. Kroh, D. Richter, C. Schröder, and J. Schupp (2019). The German Socio-Economic Panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik 239*(2), 345–360.

Gorwa, R. (2019). The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content. *Internet Policy Review 8*(2), 1–22.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National academy of Sciences 101*(suppl_1), 5228–5235.

Guriev, S., E. Henry, T. Marquis, and E. Zhuravskaya (2023). Curtailing False News, Amplifying Truth. *Amplifying Truth (October 29, 2023)*.

Guriev, S., N. Melnikov, and E. Zhuravskaya (2020). 3G Internet and Confidence in Government*. *The Quarterly Journal of Economics*.

Han, X. and Y. Tsvetkov (2020). Fortifying Toxic Speech Detectors Against Veiled Toxicity. *arXiv preprint arXiv:2010.03154*.

Hansen, S., M. McMahon, and A. Prat (2018). Transparency and deliberation within the fomc: A computational linguistics approach. *The Quarterly Journal of Economics 133*(2), 801–870.

Heldt, A. P. (2019). Reading Between the Lines and the Numbers: An Analysis of the First Netzdg Reports. *Internet Policy Review 8*(2).

Henry, E., E. Zhuravskaya, and S. Guriev (2022). Checking and Sharing Alt-facts. *American Economic Journal: Economic Policy 14*(3), 55–86.

Howard, P. N., A. Duffy, D. Freelon, M. Hussain, W. Mari, and M. Maziad (2011). Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring? *Working Paper*.

Jiménez Durán, R. (2022). The Economics of Content Moderation: Theory and Experimental Evidence From Hate Speech on Twitter. *Available at SSRN*.

Jones, J. J., R. M. Bond, E. Bakshy, D. Eckles, and J. H. Fowler (2017, 04). Social Influence and Political Mobilization: Further Evidence From a Randomized Experiment

in the 2012 U.S. Presidential Election. *PLOS ONE 12*(4), 1–9.

Kaye, D. (2018). Mandate of the special rapporteur on the promotion and protection of the right to freedom of opinion and expression. *OL ITA 1*(2018), 20.

Kaye, D. A. (2019). *Speech Police: The Global Struggle to Govern the Internet*. Columbia Global Reports.

Kohl, U. (2022). Platform Regulation of Hate Speech–A Transatlantic Speech Compromise? *Journal of Media Law*, 1–25.

Kominers, S. D. and J. M. Shapiro (2024). Content Moderation with Opaque Policies. Working Paper 32156, National Bureau of Economic Research.

Levy, R. (2021, March). Social Media, News Consumption, and Polarization: Evidence from a Field Experiment. *American Economic Review 111*(3), 831–70.

Liu, Y., P. Yildirim, and Z. J. Zhang (2021). Social Media, Content Moderation, and Technology. *arXiv preprint arXiv:2101.04618*.

Madio, L. and M. Quinn (2021). Content Moderation and Advertising in Social Media Platforms. *Available at SSRN 3551103*.

Manacorda, M. and A. Tesei (2020). Liberation Technology: Mobile phones and Political Mobilization in Africa. *Econometrica 88*(2), 533–567.

Mosquera, R., M. Odunowo, T. McNamara, X. Guo, and R. Petrie (2020). The Economic Effects of Facebook. *Experimental Economics 23*(2), 575–602.

Müller, K. and C. Schwarz (2021). Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association 19*(4), 2131–2167.

Müller, K. and C. Schwarz (2022). The Effects of Online Content Moderation: Evidence from President Trump's Account Deletion. *Available at SSRN 4296306*.

Müller, K. and C. Schwarz (2023). From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics 15*(3), 270–312.

New York Times (2017). Seeking Asylum in Germany, and Finding Hatred, By Ainara Tiefenthäler, Shane O'neill and Andrew Michael Ellis .

Olden, A. and J. Møen (2022). The Triple Difference Estimator. *The Econometrics Journal 25*(3), 531–553.

Parliament, E. (2016). European parliament resolution of 14 april 2016 on the 2015 report on turkey.

Pemstein, D., K. L. Marquardt, E. Tzelgov, Y.-t. Wang, J. Krusell, and F. Miri (2018). The v-dem measurement model: latent variable analysis for cross-national and cross-temporal expert-coded data. *V-Dem working paper 21*.

Qin, B., D. Strömberg, and Y. Wu (2017). Why does china allow freer social media? protests versus surveillance and propaganda. *Journal of Economic Perspectives 31*(1), 117–140.

Rauh, C. and L. Renée (2023). How to measure parenting styles? *Review of Economics of the Household 21*(3), 1063–1081.

Reimers, N. and I. Gurevych (2019, 11). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Schwarz, C. (2023). Estimating text regressions using txtreg_train. *The Stata Journal 23*(3), 799–812.

Spiegel (2019). Was bei den Ermittlungen gegen Ali B. schief lief. *https://www.spiegel.de/panorama/justiz/fall-susanna-f-in-wiesbaden-ali-b-was-bei-den-ermittlungen-schief-lief-a-1220080.html*. Accessed: 2024-01-27.

Sunstein, C. R. (2017). *# Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Twitter (2015). Fighting Abuse to Protect Freedom of Expression. `https://blog.twitter.com/en_us/a/2015/fighting-abuse-to-protect-freedom-of-expression`. Accessed: 2022-09-11.

Twitter (2018a). Twitter Netzwerkdurchsetzungsgesetzbericht: Januar - Juni 2018.

Twitter (2018b). Twitter Netzwerkdurchsetzungsgesetzbericht: Juli - Dezember 2018.

Vidgen, B., S. Hale, S. Staton, T. Melham, H. Margetts, O. Kammar, and M. Szymczak (2020). Recalibrating Classifiers for Interpretable Abusive Content Detection. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pp. 132–138.

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Wulczyn, E., N. Thain, and L. Dixon (2017). Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th international conference on world wide web*, pp. 1391–1399.

Yanagizawa-Drott, D. (2014). Propaganda and Conflict: Evidence from the Rwandan Genocide. *The Quarterly Journal of Economics 129*(4), 1947–1994.

Zhuravskaya, E., M. Petrova, and R. Enikolopov (2020). Political Effects of the Internet and Social Media. *Annual Review of Economics 12*.

Online Appendix: The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG

This Online Appendix consists of four parts.

1. Appendix A provides additional details on the data sources
2. Appendix B provides a theoretical framework for the empirical analysis.
3. Appendix C presents additional results on the online effects of the NetzDG.
4. Appendix D presents additional results on the offline effects of the NetzDG.

A Additional Details on the Data

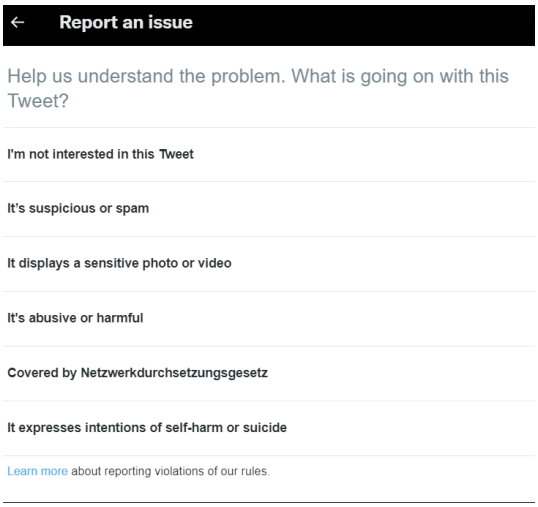
Table A.1: Translated Example of Toxic Refugee Tweets

Date	Post	Toxcity
2016-03-08	@{user} Oh {expletive} you (you refugee, go back to your country)	0.99
2020-09-12	@{user} @{user} I burn your pets like Hitler burned the fat corps of your ugly, {stream of expletives} grandparents in the concentration camps after they showered. You {expletive} refugee.	0.99
2017-11-19	@{user} It is normal. Every piece of trash that hurts, rapes and murders is given more attention than the victim. even more so when a fucking refugee. The police are instructed not to intervene so harshly	0.89
2018-03-07	@{user} What the fuck. I didn't vouch for any refugee. And now I have to pay for the stupidity of the do-gooders with my taxes. I find it an impudence.	0.86
2016-02-04	@{user} You're not a refugee otherwise, it would be free ;) the stupid German pays.	0.84

Notes: This table reports five example of toxic refugee tweets. The tweets were translated by the authors. Usernames, expletives, and links were masked.

Figure A.1: How Users Twitter Can Report Content Covered by the NetzDG

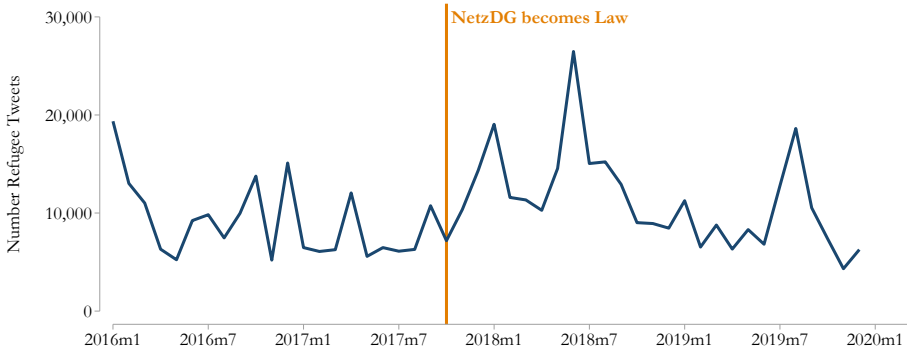
(a) Main Reporting Field



(b) Reason for Reporting

Notes: These screenshots show how Twitter users located in Germany can report content violating the NetzDG. Panel (a) shows the main reporting field a user sees when clicking on "report an issue" for a given tweet. Note that "Covered by the Netzwerkdurchsuchungsgesetz" is its own category. Panel (b) shows that the next prompt requires the user to specify a category, where "Hass schürende/verfassungswidrige Inhalte", "Gewalt/Bedrohung/Aufforderung zu Straftaten", "Beleidigung/Üble Nachrede", and "Terrorismus" refer directly to online hate speech or incitement of violence.

Figure A.2: Time Series Refugee Tweets



Notes: The time-series plot shows the monthly number of tweets mentioning the word "Flüchtling" (refugee) between 2016 and 2019.

Table A.2: Summary Statistics Toxicity Refugee Tweets

Variable	Mean	SD	p50	Min	Max	N
Toxicity Measures						
Toxicity	0.40	0.22	0.00	0.40	1.00	298,846
Sev. Toxicity	0.31	0.24	0.00	0.29	1.00	298,846
Identity Attack	0.51	0.25	0.00	0.52	1.00	298,846
Insult	0.34	0.20	0.00	0.32	1.00	298,846
Profanity	0.22	0.21	0.00	0.12	1.00	298,846
Threat	0.41	0.29	0.00	0.24	1.00	298,846
User Measures						
AfD Twitter Followers	0.29	0.45	0.00	0.00	1.00	298,846
Party Twitter Followers	0.50	0.50	0.00	0.00	1.00	298,846
Pre-Period Tox $\geq$ 50pct	0.53	0.50	0.00	1.00	1.00	298,846
Pre-Period Tox $\geq$ 75pct	0.25	0.43	0.00	0.00	1.00	298,846
Pre-Period Tox $\geq$ 90pct	0.10	0.30	0.00	0.00	1.00	298,846
Pre-Period Tox $\geq$ 95pct	0.05	0.22	0.00	0.00	1.00	298,846

Notes: This table displays the mean, standard deviation, median, minimum, maximum, and number of observations for the variables used in the tweet-level analysis.

Table A.3: Examples of Anti-Refugee Incidents

Date	Place	Description	Type
15.06.2018	Ismaning	Two 28-year-old Germans met a group of people from Eritrea on the train. At first, the two made racist comments about them. After getting off at Ismaning train station, one of the two 28-year-olds pulled a 21-year-old from the group to the ground and kicked him. The injured person lost consciousness and had to be treated in a hospital.	Assault
27.09.2018	Werdau	A 25-year-old is said to have thrown an incendiary device onto the grounds of an asylum accommodation. Half an hour before the crime, the man threatened residents and security staff of the shelter that he would set the facility and the people living there on fire. He then left the crime scene and returned with the incendiary device to throw it over the entrance gate.	Arson
20.03.2016	Steinhagen	Garbage containers under the carport of an asylum accommodation catch fire for an unknown reason. Two garbage cans burn out completely in the fire.	Suspected Case
02.07.2016	Zirndorf	25 neo-Nazis from the alliance "Franken wehrt sicht" demonstrated in the afternoon under the motto "Zirndorf says no to the home - citizen dialogue now!" in front of an asylum accommodation.	Demonstration
01.09.2018	Leipzig	Two masked men riot with a baseball bat and a pool cue in front of the house where a 31-year-old asylum seeker lives with his wife and five children.	Other cases

Notes: This table reports one example for each class of anti-refugee incidents in the data. The descriptions were translated by the authors.

Table A.4: Summary Statistics

Variable	Mean	SD	p50	Min	Max	N
Anti-Refugee Incidents						
Anti-refugee incidents	0.14	1.07	0.00	0.00	115.00	71,456
Anti-refugee incidents (arson)	0.00	0.06	0.00	0.00	9.00	71,456
Anti-refugee incidents (demonstration)	0.00	0.04	0.00	0.00	4.00	71,456
Anti-refugee incidents (assault)	0.02	0.23	0.00	0.00	15.00	71,456
Anti-refugee incidents (other)	0.11	0.86	0.00	0.00	88.00	71,456
Anti-refugee incidents (suspected cases)	0.00	0.11	0.00	0.00	13.00	71,456
Main Variables						
AfD users per capita (in %)	0.03	0.02	0.00	0.03	0.11	71,456
Log(Population)	9.15	0.93	5.81	9.10	15.07	71,456
Vote share AfD	14.86	7.01	3.13	12.85	44.86	71,008
Facebook users per capita	0.08	0.12	0.00	0.05	0.91	71,456
Share broadband internet (in %)	83.00	10.66	43.50	84.60	100.00	71,456
Additional Control Variables						
GDP per worker	63094.77	9846.31	46835.00	62207.00	136763.00	71,152
Population density	281.92	381.64	6.55	144.77	4653.18	71,456
Immigrants per capita	13.96	7.63	1.82	13.78	49.72	69,632
Refugees per capita	0.01	0.01	0.00	0.01	0.10	71,456
Registered domains per capita	0.14	0.06	0.06	0.13	1.39	71,456
Mobile broadband speed	11.90	2.33	6.24	11.60	24.41	71,456
Newspaper sales per capita	0.09	0.08	0.00	0.09	1.64	70,800
Vote share CDU	36.45	7.10	19.88	35.74	64.48	71,008
Vote share SPD	18.55	7.04	4.68	17.23	46.70	71,008
Vote share Linke	7.84	4.37	1.57	6.16	26.10	71,008
Vote share Green	7.03	3.50	0.87	6.66	25.47	71,008
Vote share FDP	9.70	2.87	3.38	9.29	27.52	71,008
Vote share NPD	0.49	0.41	0.00	0.31	2.01	71,456
Voter turnout	76.44	3.14	65.93	76.46	83.88	71,456
Average age	44.97	2.28	26.80	44.70	56.20	69,168
Share population 0-25	24.73	3.18	13.78	25.19	37.14	69,168
Share population 25-50	33.35	2.04	21.67	33.32	45.37	69,168
Share population 50-75	32.58	3.14	21.97	32.14	50.08	69,168
Share population 75+	9.34	1.81	3.58	9.22	17.65	69,168

Notes: This table displays the mean, standard deviation, median, minimum, maximum, and number of observations of the variables used in the municipality-quarter panel.

Table A.5: Summary Statistics by Quartile of AfD Facebook Users Per Capita

	1st Quartile		2nd Quartile		3rd Quartile		4th Quartile	
Variable	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Anti-refugee incidents	0.041	0.256	0.077	0.374	0.114	0.466	0.332	2.023
Anti-refugee incidents (arson)	0.000	0.022	0.002	0.050	0.002	0.050	0.004	0.094
Anti-refugee incidents (demonstration)	0.000	0.011	0.000	0.021	0.000	0.021	0.003	0.073
Anti-refugee incidents (assault)	0.004	0.084	0.009	0.111	0.017	0.157	0.065	0.415
Anti-refugee incidents (other)	0.034	0.222	0.063	0.325	0.090	0.386	0.250	1.617
Anti-refugee incidents (suspected cases)	0.001	0.044	0.003	0.069	0.004	0.111	0.011	0.171
AfD users per capita (in %)	0.002	0.004	0.019	0.004	0.034	0.005	0.063	0.018
Log(Population)	8.605	0.728	9.287	0.630	9.370	0.875	9.357	1.170
Vote share AfD	14.665	6.828	13.480	6.153	14.663	6.731	16.645	7.848
Facebook users per capita	0.064	0.121	0.084	0.131	0.086	0.115	0.086	0.098
Share broadband internet (in %)	82.737	9.859	83.633	10.184	83.196	11.256	82.433	11.215
GDP per worker	63297.784	9717.812	63976.647	10014.253	63726.393	9901.268	61373.485	9532.920
Population density	202.268	293.691	261.068	306.674	314.564	385.356	349.824	491.318
Immigrants per capita	12.913	6.617	15.095	7.253	15.016	7.726	12.837	8.495
Refugees per capita	0.010	0.005	0.011	0.005	0.011	0.007	0.011	0.007
Registered domains per capita	0.142	0.055	0.143	0.048	0.142	0.049	0.138	0.069
Mobile broadband speed	11.737	2.321	11.855	2.389	11.937	2.296	12.064	2.300
Newspaper sales per capita	0.117	0.085	0.086	0.071	0.083	0.071	0.084	0.073
Vote share CDU	38.718	7.284	37.010	6.760	35.746	6.635	34.311	6.968
Vote share SPD	17.033	6.751	19.426	7.012	19.450	6.848	18.288	7.251
Vote share Linke	6.809	3.916	7.303	3.810	7.865	4.162	9.381	5.060
Vote share Green	7.146	3.569	7.512	3.400	7.023	3.320	6.447	3.636
Vote share FDP	9.344	2.826	10.172	2.884	10.020	3.007	9.270	2.659
Vote share NPD	0.468	0.387	0.425	0.356	0.475	0.397	0.597	0.471
Voter turnout	76.904	3.006	76.836	2.980	76.368	3.057	75.669	3.333
Average age	44.687	2.301	44.621	2.069	44.980	2.119	45.608	2.465
Share population 0-25	25.294	3.170	25.326	2.970	24.672	2.957	23.624	3.307
Share population 25-50	33.519	2.017	33.496	1.885	33.343	1.923	33.050	2.267
Share population 50-75	32.236	3.149	32.116	2.915	32.588	2.919	33.378	3.393
Share population 75+	8.951	1.791	9.062	1.639	9.397	1.716	9.948	1.921

Notes: This table displays the mean, standard deviation, of the variables used in the municipality-year-quarter panel, split by quartiles of AfD Facebook users per capita (the "exposure" variable in the difference-in-differences analysis).

Table A.6: Summary Statistics GSOEP

Variable	Mean	SD	p50	Min	Max	N
Pro-refugee Attitudes						
Index refugee attitudes	0.49	0.38	0.00	0.60	1.00	36,912
Refugees are good for economy	0.59	0.49	0.00	1.00	1.00	36,682
Refugees are good for culture	0.56	0.50	0.00	1.00	1.00	36,677
Refugees are good for place of living	0.52	0.50	0.00	1.00	1.00	36,665
Refugee are a chance (Short-term)	0.25	0.44	0.00	0.00	1.00	36,667
Refugee are a chance (Long-term)	0.52	0.50	0.00	1.00	1.00	36,598
Pro-refugee Actions						
Index refugee actions	0.15	0.23	0.00	0.00	1.00	36,979
Action: Donated (Last Year)	0.27	0.44	0.00	0.00	1.00	36,861
Action: Donated (Future)	0.30	0.46	0.00	0.00	1.00	36,332
Action: Volunteered (Last Year)	0.07	0.25	0.00	0.00	1.00	36,761
Action: Volunteered (Future)	0.11	0.31	0.00	0.00	1.00	36,225
Action: Demonstrated (Last Year)	0.05	0.21	0.00	0.00	1.00	36,733
Action: Demonstrated (Future)	0.08	0.27	0.00	0.00	1.00	36,202
Respondent Characteristics						
Social media user	0.63	0.48	0.00	1.00	1.00	41,644
AfD voter	0.06	0.24	0.00	0.00	1.00	41,644
Female	0.53	0.50	0.00	1.00	1.00	41,644
Age	49.23	17.11	18.00	48.00	103.00	41,643

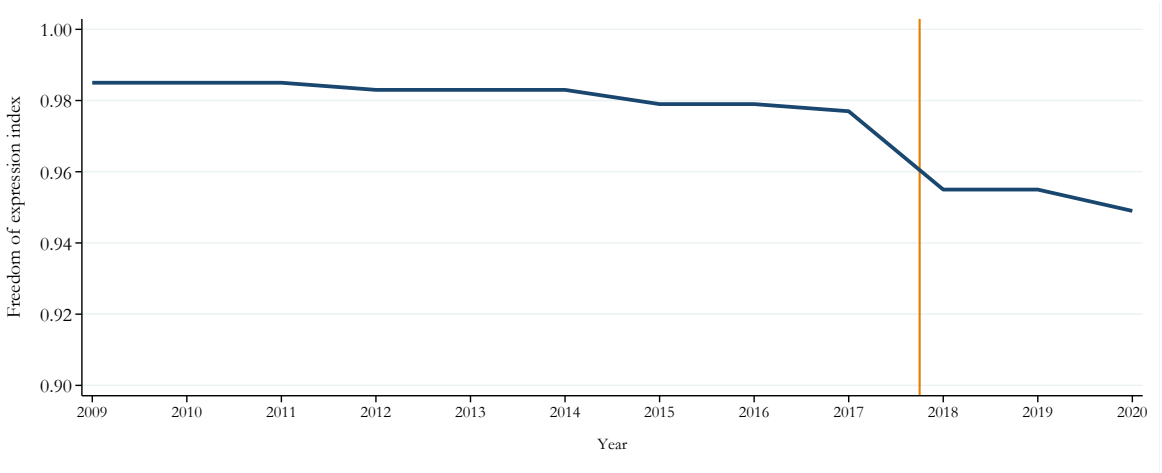
Notes: This table displays the mean, standard deviation, median, minimum, maximum, and number of observations for the variables from GSOEP used in the attitudes analysis.

Table A.7: OSCE Members and Data Filters

OSCE State	No data 2009-2020	Microstate	Data changes 2017-2020	7+ missings 2009-2020	End gaps
Albania				×	×
Andorra		×			
Armenia				×	×
Austria					
Azerbaijan				×	×
Belarus				×	×
Belgium					
Bosnia and Herzegovina					
Bulgaria					
Canada			×		
Croatia					
Cyprus					
Czech Republic					
Denmark					
Estonia				×	
Finland					
France					×
Georgia			×		
Germany					
Greece			×		
Holy See		×			×
Hungary			×		
Iceland					×
Ireland			×		
Italy					
Kazakhstan					×
Kyrgyzstan				×	×
Latvia					×
Liechtenstein		×			
Lithuania					
Luxembourg	×			×	×
Malta	×	×		×	×
Moldova					
Monaco	×			×	×
Mongolia					×
Montenegro				×	×
Netherlands			×		
North Macedonia				×	×
Norway			×		
Poland					
Portugal					
Romania				×	×
Russian Federation				×	×
San Marino	×	×		×	×
Serbia			×		
Slovakia					
Slovenia			×	×	
Spain					
Sweden			×		
Switzerland					
Tajikistan	×			×	×
Turkey					
Turkmenistan	×			×	×
UK					
US					
Ukraine					
Uzbekistan				×	×

Notes: This table presents the list of the 57 OSCE member States and the selection criteria used to filter them. Germany and the donors in the baseline specification are bolded. "No data 2009-2020" indicates that there was no data for that period. "Microstate" indicates microstates. "End gaps" indicates missing data at the beginning or end of the series, even after interpolation (i.e., countries that would require extrapolation to be balanced). "7+ missings 2009-2020" indicates that the raw data has more than 7 years of missing values. "Data changes 2017-2020" indicates changes in the measurement of hate crimes in that period.

Figure A.3: Freedom of Expression Index, 2009-2020



Notes: This graph shows the Freedom of Expression Index for Germany obtained from the V-Dem dataset (Pemstein et al., 2018). This index aggregates the ratings provided by multiple country experts who respond to questions regarding government censorship efforts, the harassment of journalists, media self-censorship, media bias, freedom of discussion for ordinary citizens, and freedom of academic and cultural expression. For reference, the mean value of the index pre-NetzDG across countries was 0.68 and the standard deviation was 0.28.

B Theoretical Framework

This model builds on the microfoundation laid out in Jiménez Durán (2022). The model assumes that there is a single platform on which two types of users—"Acceptable" (A) and "Hater" (H)—interact with each other. The platform chooses a moderation rate $c \in [0, 1]$ that determines the proportion of hateful content that survives on the platform. Moreover, by carefully choosing its advertising frequencies, the platform can effectively choose the engagement of each type of user; that is, the amount of time they spend consuming content. Let T^A denote the aggregate engagement of acceptable users and T^H denote the aggregate engagement of hateful users post-moderation.

The platform faces inverse demands $p^\theta(T^A, T^H, c)$, $\theta \in \{A, H\}$. These objects equal the amount of dollars that advertisers are willing to pay per minute of ad times the amount of time that users are willing to spend watching ads per minute of content consumed.[28] The platform also has costs $\phi(T^A, T^H, c)$ and is required by a regulator to pay an expected penalty $\tau > 0$ for each unit of hateful content that it fails to moderate. Hence, its problem becomes:

$$\max_{T^A, T^H, c} \; p^A(T^A, T^H, c)T^A + p^H(T^A, T^H, c)T^H - \phi(T^A, T^H, c) - \tau T^H. \tag{B.1}$$

We interpret the implementation of the NetzDG as a marginal increase in the expected regulatory penalty; $d\tau > 0$. In other words, the policy resulted in an increase in the marginal cost of unmoderated hate speech. In this case, it is easy to show that, if the second-order conditions of problem (B.1) hold, the amount of surviving hateful content on the platform decreases in response to an increase in fines; $dT^H/d\tau < 0$.[29]

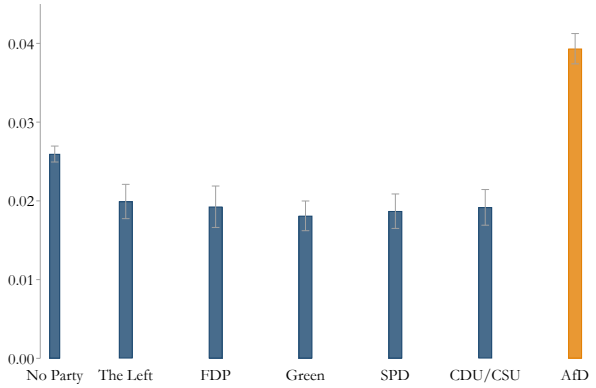
[28] In the notation of Jiménez Durán (2022), $p^\theta(T^A, T^H, c) = a^\theta(T^A, T^H, c)P^\theta(T^A, T^H, c)$, where a^θ denotes the advertisers' willingness to pay and P^θ denotes the advertising load for type θ. In this paper, we allow the platform to be a price-setter in the ads market.

[29] To see why, rewrite problem (B.1) as $\max_{T^H} \; \widetilde{\pi}(T^H) - \tau T^H$, where $\widetilde{\pi}(T^H)$ denotes the maximized profits (pre-penalties) for a given T^H. Applying the implicit function theorem to the first-order condition of this problem yields $dT^H/d\tau = 1/\widetilde{\pi}''$. The second-order condition of the problem requires that $\widetilde{\pi}'' < 0$.

C Additional Results on Online Effects

C.1. Additional Results for the Toxicity of Refugee Tweets

Figure C.1: Toxicity of Refugee Twitter Content by Party in Pre-Period



Notes: The figure shows bar graphs with the frequency of tweets with a toxicity larger than 0.8, which is a commonly-used cutoff for classifying hate speech in the literature (ElSherief et al., 2018; Han and Tsvetkov, 2020; Vidgen et al., 2020), depending on which German party users follow before the passing of the NetzDG.

Table C.1: Robustness: Threshold of Pre-Period Toxicity

	Dep. var.: Toxicity Measures			
	(1)	(2)	(3)	(4)
Pre-Period Tox $\geq$ 50pct $\times$ Post	-0.080***			
	(0.002)			
Pre-Period Tox $\geq$ 75pct $\times$ Post		-0.084***		
		(0.004)		
Pre-Period Tox $\geq$ 90pct $\times$ Post			-0.129***	
			(0.005)	
Pre-Period Tox $\geq$ 95pct $\times$ Post				-0.183***
				(0.005)
User FE	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes
Observations	277,135	277,135	277,135	277,135
Pre-Period Mean of DV	0.39	0.39	0.39	0.39
R^2	0.28	0.28	0.28	0.28

Notes: This table presents the results of estimating Equation (1) where the dependent variable is the toxicity of tweets containing the word "Flüchtling" (refugee) (bounded between 0 and 1). $Toxic\ User_u$ is an indicator variable equal to 1 if a users' tweets before the NetzDG were on average above the 50th, 75th, 90th, or 95th percentile. All regressions control for user and day fixed effects. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table C.2: Robustness: Toxicity Measures – Refugee Tweets

	Dep. var.: Toxicity measured by:					
	Toxicity	Severe Toxicity	Identity Attack	Insult	Profanity	Threat
	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a): Highly Toxic Users						
Highly Toxic User × Post	-0.084***	-0.084***	-0.082***	-0.073***	-0.071***	-0.058***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.005)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	277,135	277,135	277,135	277,135	277,135	277,135
Pre-Period Mean of DV	0.39	0.30	0.50	0.33	0.21	0.41
R^2	0.28	0.27	0.26	0.27	0.25	0.29
Panel (b): AfD Followers						
AfD follower × Post	-0.016***	-0.017***	-0.023***	-0.017***	-0.019***	0.004
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.004)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	277,135	277,135	277,135	277,135	277,135	277,135
Pre-Period Mean of DV	0.39	0.30	0.50	0.33	0.21	0.41
R^2	0.28	0.27	0.26	0.27	0.24	0.29

Notes: This table presents the results of estimating Equation (1), where the dependent variable is the measure of toxicity listed in the top row, bounded between 0 and 1, calculated based on tweets containing the word refugee ("Flüchtling"). In panel (a), we use and indicator variable equal to 1 if a users' tweets before the NetzDG were on average above the 75th percentile. In panel (b) *AfD follower* is an indicator variable that is equal to 1 if a Twitter user follows the AfD's account. All regressions control for AfD follower and day fixed effects. Robust standard errors in parentheses are clustered by users. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

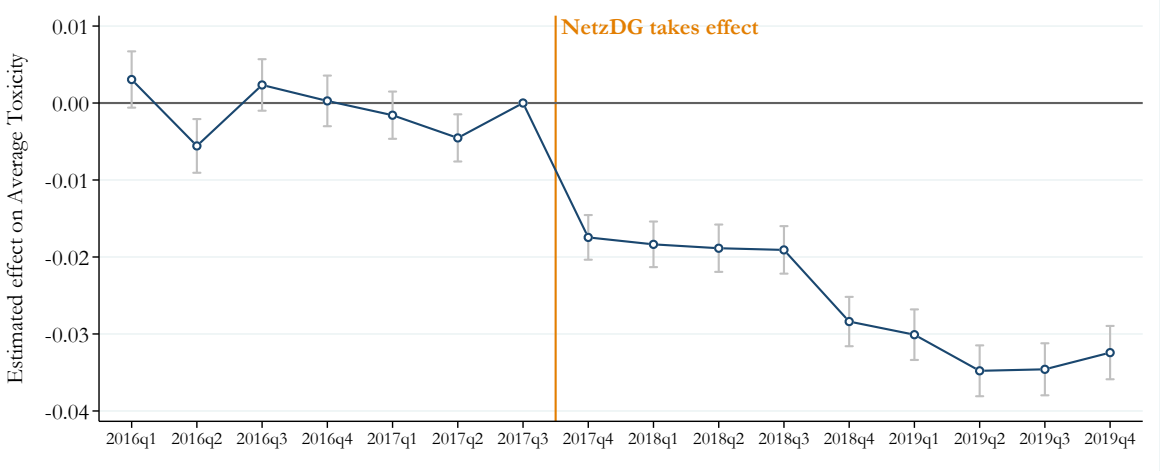
Table C.3: NetzDG and Refugee-related Twitter Activity

	Dep. var.: Asinh(Nr. Refugee Tweets)			
	(1)	(2)	(3)	(4)
Highly Toxic Users × Post	0.107***	0.071***		
	(0.012)	(0.026)		
AfD followers × Post			0.103***	0.357***
			(0.018)	(0.031)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Linear Time Trend		Yes		Yes
Observations	94,121	94,121	94,121	94,121
Pre-Period Mean of DV	1.35	1.35	1.35	1.35
R^2	0.63	0.73	0.63	0.73

Notes: This table presents the results of estimating from two-way fixed effect regression in a balanced panel of Twitter users, where the dependent variable is the inverse hyperbolic sine of the number of tweets containing the word "Flüchtling" (refugee) send by user i in quarter t. In columns (1) and (2), $Toxic\,User_i$ is an indicator variable equal to 1 if a user's tweets before the NetzDG were, on average, above the 75th percentile. In columns (3) and (4), $AfDfollowers_i$ is an indicator variable that is equal to 1 if a Twitter user follows the AfD's account. All regressions control for user and quarter fixed effects. Columns (2) and (4) additionally control for user-specific linear time trends. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

C.2. Additional Results for the Toxicity of All Tweets

Figure C.2: NetzDG and Overall Online Toxicity



Notes: The figure plots the coefficients from event study versions of Equation (1). The dependent variable is the average toxicity of all tweets sent by the users from our main analysis. The omitted category is the 3rd quarter of 2017, the quarter before the passing of the NetzDG (indicated with the vertical line). The whiskers indicate 95% confidence intervals based on standard errors clustered by user.

Table C.4: Regression Estimates: NetzDG and Overall Online Toxicity

	Dep. var.:			
	Toxicity		Asinh(Nr. Tweets)	
	(1)	(2)	(3)	(4)
Highly Toxic User × Post	-0.024***	-0.013***	0.077***	0.006
	(0.001)	(0.001)	(0.014)	(0.015)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Linear Time Trend		Yes		Yes
Observations	681,339	681,339	681,339	681,339
Pre-Period Mean of DV	0.17	0.17	0.17	0.17
R^2	0.52	0.61	0.54	0.73

Notes: This table presents the results of estimating from two-way fixed effect regression in a balanced panel of Twitter users, where the dependent variable is either the average toxicity of tweets (columns 1 and 2) or the inverse hyperbolic sine of the number of tweets (columns 3 and 4) send by user i in quarter t. $HighlyToxic\,User_i$ is an indicator variable equal to 1 if a user's tweets before the NetzDG were, on average, above the 75th percentile. All regressions control for user and quarter fixed effects. Columns (2) and (4) additionally control for user-specific linear time trends. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table C.5: Robustness: Threshold of Pre-Period Toxicity – All Tweets

	Dep. var.: Toxicity Measures			
	(1)	(2)	(3)	(4)
Pre-Period Tox ≥ 50pct × Post	-0.020***			
	(0.001)			
Pre-Period Tox ≥ 75pct × Post		-0.024***		
		(0.001)		
Pre-Period Tox ≥ 90pct × Post			-0.036***	
			(0.001)	
Pre-Period Tox ≥ 95pct × Post				-0.046***
				(0.002)
User FE	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes
Observations	681,339	681,339	681,339	681,339
Pre-Period Mean of DV	0.17	0.17	0.17	0.17
R^2	0.52	0.52	0.52	0.52

Notes: This table presents the results of estimating from two-way fixed effect regression in a balanced panel of Twitter users, where the dependent variable is the average toxicity of tweets send by user i in quarter t. Each column presents the estimate for different definitions of toxic users. In each case, toxic users are defined as an indicator variable equal to 1 if a user's tweets before the NetzDG were, on average, above the indicated percentile. All regressions control for user and quarter fixed effects. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table C.6: Robustness: Toxicity Measures – All Tweets

	Dep. var.: Toxicity measured by:					
	Toxicity	Severe Toxicity	Identity Attack	Insult	Profanity	Threat
	(1)	(2)	(3)	(4)	(5)	(6)
Highly Toxic User × Post	-0.024***	-0.017***	-0.016***	-0.022***	-0.015***	-0.011***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	681,339	681,339	681,339	681,339	681,339	681,339
Pre-Period Mean of DV	0.17	0.10	0.14	0.15	0.11	0.16
R^2	0.52	0.49	0.55	0.55	0.44	0.55

Notes: This table presents the results of estimating from two-way fixed effect regression in a balanced panel of Twitter users, where the dependent variable is the average toxicity of tweets send by user i in quarter t. Toxicity is measured based on the toxicity dimension indicated at the top of each column. $HighlyToxic\,User_i$ is an indicator variable equal to 1 if a user's tweets before the NetzDG were, on average, above the 75th percentile. All regressions control for user and quarter fixed effects. Robust standard errors in parentheses are clustered by user. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

C.3. Additional Results for Content Changes

Figure C.3: Changes in Word Frequencies – All Tweets

(a) Decreasing Frequency

(b) Increasing Frequency

Notes: This figure shows word clouds that visualize the relative word frequency changes for toxic compared to non-toxic users after the NetzDG among all tweets. Panel (a) shows words with decreasing relative frequency, while panel (b) shows words with increasing relative frequency. The size of the words is proportional to the frequency change.

Table C.7: Overview Topic Model: Falling Topics

Topic Nr.	Topic Label	Topic Words
67	Online Commentary	positive, positive, positive, positive, positive, negative, negative, negative, video, video evidence, negative, optimistic, optimism, strachevideo, commented, videos, youtube, music video, feedback, comments, comment, minus, comments, comment, hate comments, youtubers, youtuber, commentators, reaction, reactions, criticized, stability, review, commentator, criticize, reacts, comments, response, neutrality, react, satisfaction, neutral, ratings, rating, clip, clipmedianews, rated, viral, test, neutral
21	Turkey & Refugees	erdogan, erdogan, turkish, turkish, turk, turkish, turkish, turkey, turchen, turkey, turk, turkish, istanbul, turks, jihadists, gundogan, syrian, syrian, armenia, iranian, syrian, iran, kurdish, iranian, iranian, kurdish, iranian, kurds, taliban, iraqi, islamist, islamophobia, pakistan, refugees, islamists, syrians, asylum policy, islamist, asylum crisis, croatia, islamic, islamic, right-wing extremist, islamic, islamist, islamist, islamization, neo-nazis, constantine, refugees
15	Public Holidays & Event	new year reception, april, new year cleaning, new year, november, https, october, september, june, mars, previous year, april fool's joke, january, february, year of life, july, this year, august, anno, year, this year, decades, election night, decades, spring, the other day, start of the week, new beginning, oktoberfest, election year, today show, decades, zuckerberg, new elections, valentine's day, tomorrow, decade, berlin election, morning post, federal election, quarter finals, tomorrow, elections, inauguration, to begin, started, contemporary, debut, Reichstag, marcus
0	Informal Opinions	stabbing, princess, mess, for my sake, brewery, district office, amok, nogroko, me, imo, moi, nintendo, imam, sam, ergo, imho, lk, hmmm, yo, hahah, ahhh, pforzheim, ohhh, take, mi, hmm, ios, eu, bim, ahh, I, mine, kerstin, ohh, ohm, oh well, uff, ehm, hah, diego, uncomfortable, hahaha, hahahaha, my, hh, mia, haha, imm, unpleasant, ypg
41	Foreign Politics	obama, federal president, presidential election, obamas, republicans, president, wikileaks, candidate for chancellor, president, president, assange, anti-democrats, goes to the elections, chair, bolsonaro, chairman, democrat, americans, election night, democrats, liberal, new elections, women's suffrage, liberals, tweet, america, chairman, america, social democrat, erdogan, twitter, chairman, twitterer, zuckerberg, spd chairmanship, erdogan, protest voter, neoliberal, liberal, america, twitter account, tweet, neo-nazis, federal republic, american, tweet, neoliberalism, retweet, neoliberal, american
4	Sport	women's football, football, football fans, football game, handball, soccer, world cup, footballer, eurosport, national player, fcbayern, schweinsteiger, ltwbayern, hopesheim, women's football, ancelotti, champions league, sports director, stadium, league, cup final, olympics, sports show, hockey, teams, superbowl, ice hockey, European champions, semi-finals, Upper Bavaria, derby, sports, Lower Bavaria, round of 16, sporty, basketball, Olympics, sport, playing field, coach, athlete, esports, the team, DFB Cup, sports studio, club, team, playoffs, Bierhoff
62	Terrorism	terrorist, terrorist, terrorist attack, terrorism, terrorists, terrorist attacks, terrorist, terrorist group, terrorist militia, suspected terrorism, terror threat, terror, bomb attack, terrors, bombed, extremists, world war bomb, aerial bomb, bomb, bomb threat, extremism, right-wing terror, bombs, jihadists, assassin, assassination, right-wing extremist, mass murderer, bomber, atomic bomb, islamophobia, explosion, islamists, explode, explosions, mass murder, islamist, islamist, islamist, killer, death threats, deep, islamism, islamist, murder, murder, violent, attacks, buffet, murder case
34	Elections	elections, new elections, local elections, deselections, parliamentary election, presidential election, local election, select, protest voters, state elections, women's suffrage, election night, re-election, European elections, state election, voting, Berlin election, Bavarian election, word choice, votes, postal vote, eu election, candidate for chancellor, new election, vote, vote, run-off election, european election, non-voter, hessian election, selected, elected, voted out, referendum, selection, elected, choose, selected, voted, federal election, elected, democracy, candidates, democracies, democratic, democratic, undemocratic, run for office, free voter, democratic
19	Local Events	augsburg, wurzburg, harburg, freiburg, aschaffenburg, tecklenburg, stronghold, petersburg, neubrandenburg, stauffenberg, flensburg, homburg, wolfsburg, poggenburg, ravensburg, strasbourg, mecklenburg, wurttemberg, heidelberg, regensburg, ludwigsburg, luneburg, hamburg, brandenburg, magdeburg, oldenburg, hopenheim, salzburg, charlottenburg, lindenberg, duisburg, brandenburger, gretathunberg, wittenberg, vorarlberg, luxembourg, nurnberg, reinickendorf, hamburg, marburg, hambacherwald, train stations, ingolstadt, luxemburg, nurnberger, capital, hellersdorf, recklinghausen, meinfeldkirch, refugee home
17	News	today today the day, afternoon, everyday, afternoon, Saturday morning, daily, good morning, church day, morning, if possible, everyday, morning, enable, the day after tomorrow, state media, Friday, Monday, hour day, fridayforfuture, happybirthday, matchday, noon, Fridays, Valentine's Day, impossible

Notes: This table lists the most important topic words for the topics generated by the top2vec topic model (Angelov, 2020).

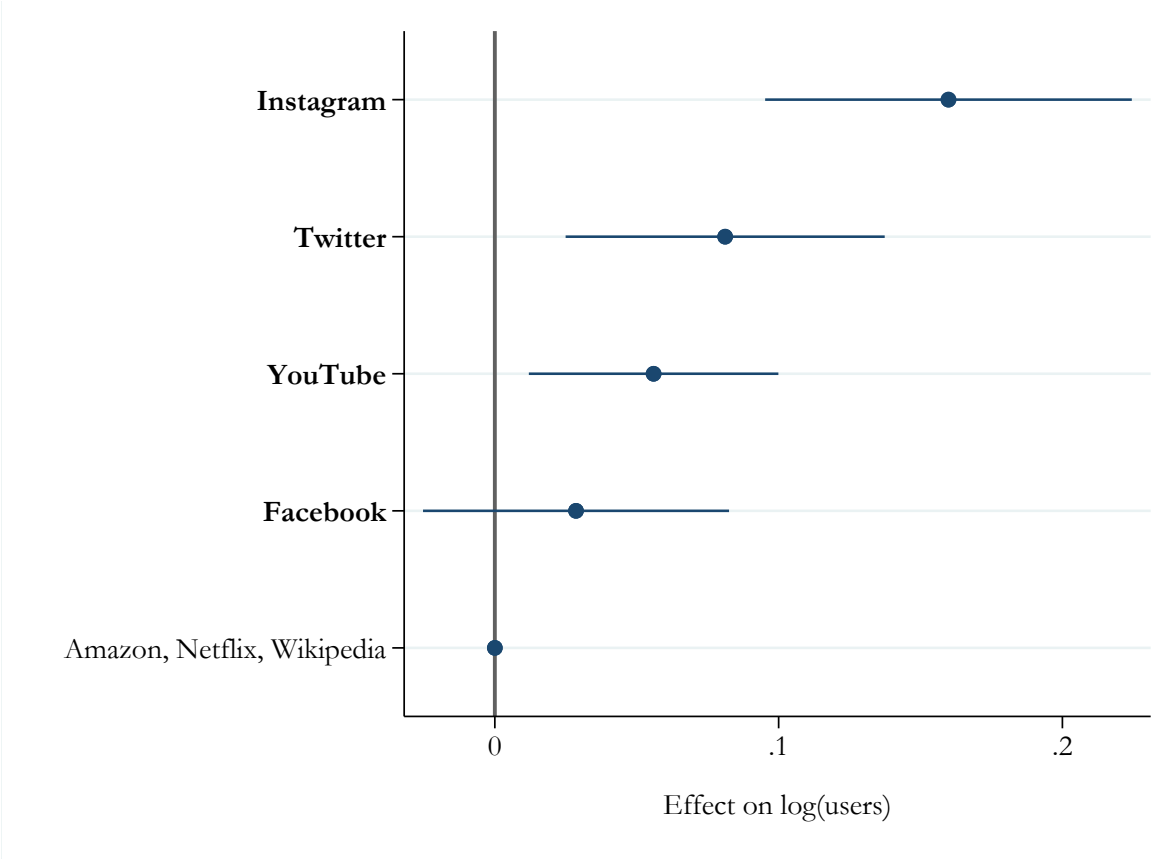
Table C.7: Overview Topic Model: Rising Topics

Topic Nr.	Topic Label	Topic Words
3	Germany	german, german turks, germans, germany, german, german, german, german, germany, deutschlandfunk, deutschebahn, germanywide, tedesco, germany, germania, deutschlan, deutschl, nazi march, deutsch, deutschla, niemehrcsu, east german, deutschebank, east german, nazis, east germany, north germany, niemehrcdu, berlin election, nazisraus, berlinale, berlin, west german, nazi, neo-nazis, south german, berlin, austria, neo-nazi, holocaust, hollande, berlindirekt, deutschrap, niemehrspd, wurttemberg, nationalists, ltwbayern, nationalism, migrants, amsterdam
20	Neo-nazis	nazi march, neo-nazis, nazis, nazisraus, neo-nazi, nazi, holocaust, fascist, anti-fascists, right-wing extremist, anti-fascism, fascist, anti-semitism, anti-semitic, anti-fascist, anti-semitic, anti-semitic, nationalists, anti-semitic, anti-semitic, german turks, extremists, nationalism, fascists, anti-semitic, deutschebahn, extremism, north germany, deutschlandfunk, fascism, neoliberalism, german, mass murderer, german, german, germany, germany, civil war, left-wing fascists, communists, german, german, neoliberal, socialist, socialist, german, berliner, berlinale, right-wing radical, jihadists
18	Feminism	feminists, feminists, feminist, feminist, feminism, feminist, feminist, sexism, sexist, sexist, women, female, patriarchy, female, ladies, gender, woman, female, muller, genders, gender, international women's day, politicians, lady, female, candidates, participants, cleaning lady, hannelore, ladies, masculinity, masculinity, comrades, teachers, girl, journalists, lesbian, sexual
45	Anti-semitism	anti-Semitism, anti-Semites, anti-Semitic, anti-Semitic, anti-Semitic, anti-Semitic, anti-Semitic, Jews, Jew-hatred, Jewish, Jewish, Israeli, Israeli, Israel, Jewish, Israelis, Jewish, Zionists, Israel, Jewish, Jew, holocaust, synagogues, neo-Nazis, synagogue, nazi march, neo-nazi, judaism, anti-fascists, judith, nazis, netanyahu, nazisraus, anti-fascism, jerusalem, jihadists, right-wing extremist, nazi, anti-fascist, dusseldorf, duesseldorf, hellersdorf, dusseldorfer, islamophobia, extremists, extremism, fascist, zehlendorf, fascist, islamists
7	Politics (Far-right/left)	people's party, left-wing party, anti-democrats, the party, people's parties, democrats, republicans, pirate party, democrat, right-wing extremist, old parties, fascist, anti-fascists, social democrat, state party conference, liberals, protest voters, liberal, federal party conference, workers' party, fascist, liberals, fascists, anti-fascist, women's suffrage, democratic, left-wing fascists, democratic, anti-fascism, parties, right-wing radical, right-wing radical, democratic, party, democratic, right-wing radical, liberal, extremists, liberalism, conservative, democratic, undemocratic, conservatives, political, parliament, political, party donations, parliamentary election, fascism, demonstrators
157	Local Events	Clausnitz, Mecklenburg, Chemnitz, Copenhagen, Heidenheim, Hoffenheim, Meinfeldkirch, Recklinghausen, Connewitz, Hellersdorf, Tecklenburg, Oberhausen, Hildesheim, Reinickendorf, Weinheim, Dusseldorf, capital, Duesseldorf, Naziaufmarsch, neo -Nazis, Holocaust, Russelladt, Ingolstadt, Dusseldorfer,, mulheim, holstein, berlin, zehlendorf, berlin election, berlinale, neustadt, sweden, magnitz, neo-nazi, switzerland, darmstadt, anti-semitism, stockholm, alexanderplatz, kimmich, stauffenberg, schanzenviertel, swedish, nordstadt, hessen election, heidelberg, wikileaks, nazis, demonstrators
72	Christmas	christmas time, christmas party, christmas festival, christmas, christmas, christmas, christmas tree, christmas money, santa claus, christmas market, christmas eve, new year's reception, holidays, holidays, holiday, natalie, new year's cleaning, new year's day, celebrates, valentine's day, christchurch, christ child, snowing, halloween, christopher, winter break, snowden, christiane, christian, christian, celebrate, christian, winter, christine, christianity, christ, christin, kristina, christian, christoph, winter time, january, christ, christina, gifts, winterthur, christian, celebrated, november, february
37	Politics (Conservatism/Liberalism)	people's party, the party, left party, people's parties, parliament, pkk, parliaments, parliaments, social democrat, republican, pirate party, old parties, parliamentary election, parliamentarians, socialist, socialists, socialism, socialist, democrats, politician, anti-democrats, politicians, local elections, goes elections, federal minister, democrat, new elections, berlin election, workers' party, liberal, politician, liberals, political, politician, political, conservative, politics, political, candidate for chancellor, conservatives, political, neoliberalism, politicians, political, political, protest voter, foreign minister, liberal, party donations
38	Justice System	legal, legal, constitutional state, legal, legal, legal, constitutional state, legal situation, legal, lawyer, constitutional state, legal, international law, legislation, law, unlawful, legal, basic law, legal, basic law, lawyer, legal system, lawyers, laws, legal, legislator, police law, court, criminal law, public prosecutor, prevention, bill, betting, court, law, lawyer, prosecutors, legal, legal committee, right-wing, law, criminal, asylum law, legalization, sued, arbitrator, just, civil rights, court of auditors, justice
22	Journalism	journalist, journalists, journalist, journalism, clipmedianews, journalist, journalists, media report, media, newsticker, daily newspaper, reporter, newsbasel, reporter, anonymousnews, newsflash, press conference, newsletter, news, media library, newsroom, srfnews, propaganda, freedom of the press, pers, press mirror, journal, press spokesman, koran, medial, medial, reports, social media, press, press club, mediale, magazine, press, reportage, news, media, printed, publication, publish, wikileaks, multimedia, liegenpresse, magazine, print out, tv

Notes: This table lists the most important topic words for the topics generated by the top2vec topic model (Angelov, 2020).

C.4. Additional Results for Platform Usage

Figure C.4: The Effect of the NetzDG on Platform Usage by Platform



Notes: This figure plots coefficients from a version of Equation (2) which replaces the dummy for treated platforms with dummies for individual platforms. The dependent variable is the log number of users. The omitted category is the set of untreated platforms. The whiskers indicate 95% confidence intervals based on standard errors clustered by country.

D Additional Results on Offline Effects

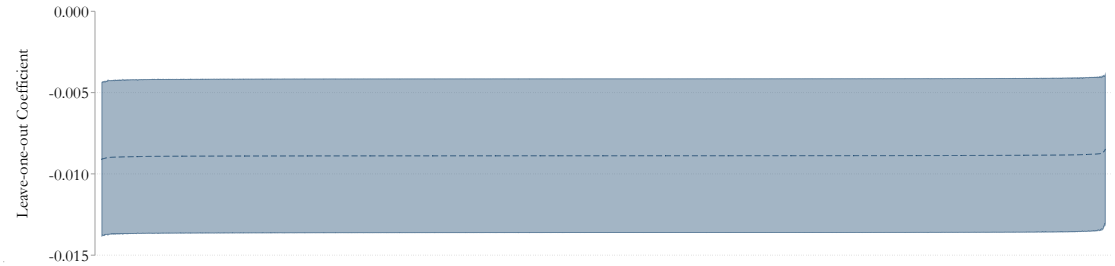
D.1. Additional Results for Hate Crimes

Table D.1: Robustness: Type of Hate Crime Incident

	Dep. var.: Type of Anti-refuge Hate Crime					
	All	Arson	Assault	Demonstration	Other	Suspect. Cases
	(1)	(2)	(3)	(4)	(5)	(6)
AfD Facebook users p.c. (std) × Post	-0.009***	-0.000	-0.003**	-0.001**	-0.008***	-0.001**
	(0.002)	(0.000)	(0.001)	(0.000)	(0.002)	(0.000)
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Ln(Pop/) × Post	Yes	Yes	Yes	Yes	Yes	Yes
AfD vote share × Post	Yes	Yes	Yes	Yes	Yes	Yes
Election Controls × Post	Yes	Yes	Yes	Yes	Yes	Yes
Facebook users p.c × Post	Yes	Yes	Yes	Yes	Yes	Yes
Broadband internet × Post	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71,008	71,008	71,008	71,008	71,008	71,008
Pre-Period Mean of DV	0.12	0.00	0.02	0.00	0.10	0.01
R^2	0.44	0.09	0.38	0.15	0.40	0.16

Notes: This table presents the results of estimating municipality-quarter-level regressions as in Equation (3) where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes of a specific type (indicated in the top row). *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. All regressions include municipality and quarter fixed effects, as well as controls for the logarithm of population, the AfD vote share, Facebook users per capita, and broadband internet access, all interacted with *Post*. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure D.1: Leave-one-out Estimates



Notes: This figures shows the estimates of a leave-one-out exercise, where we estimate Equation (1) omitting one municipality at a time. The figure plots a total of 4,466 estimates sorted by size. The dashed line are the point estimate and the shading indicates 95% confidence intervals.

Table D.2: Robustness: Specification

	Dep. var.: Anti-Refugee Hate Crime					
	Asinh	Count	Ln(p.c.)	Asinh	Count	Ln(p.c.)
	(1)	(2)	(3)	(4)	(5)	(6)
AfD Facebook users p.c. (std) × Post	-0.009***	-0.022***	-0.007***			
	(0.002)	(0.005)	(0.002)			
High AfD Usage × Post				-0.023***	-0.065***	-0.018***
				(0.007)	(0.018)	(0.005)
Municipality FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Ln(Pop/) × Post	Yes	Yes	Yes	Yes	Yes	Yes
AfD vote share × Post	Yes	Yes	Yes	Yes	Yes	Yes
Election Controls × Post	Yes	Yes	Yes	Yes	Yes	Yes
Facebook users p.c × Post	Yes	Yes	Yes	Yes	Yes	Yes
Broadband internet × Post	Yes	Yes	Yes	Yes	Yes	Yes
Observations	71,008	71,008	71,008	71,008	71,008	71,008
Pre-Period Mean of DV	0.12	0.19	-9.06	0.12	0.19	-9.06
R^2	0.44	0.63	0.95	0.44	0.63	0.95

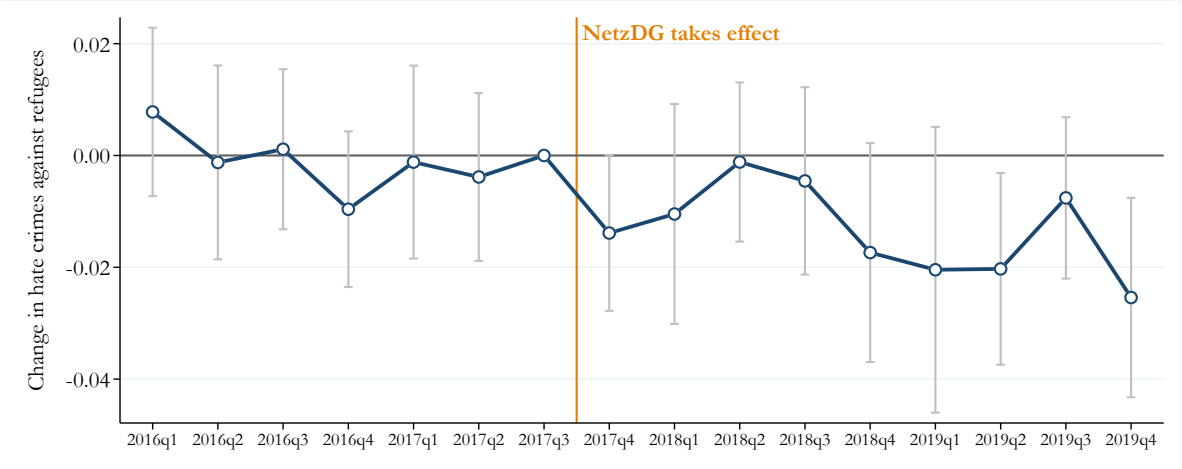
Notes: This table presents the results of estimating municipality-quarter-level regressions as in Equation (3) where the dependent variable is the transformation of anti-refugee hate crimes indicated at the top of the table. *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. *High AfD Usage* is an indicator equal to 1 for municipalities with an above-median number of AfD Facebook followers per capita. All regressions include municipality and quarter fixed effects, and controls for the logarithm of population, the AfD vote share, Facebook users per capita, and broadband internet access, all interacted with *Post*. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table D.3: Robustness: Standard Errors

	Standard Errors Clustered by:			
	County	County & Quarter	Municipality	Municipality & Quarter
	(1)	(2)	(3)	(4)
AfD Facebook users p.c. (std) × Post	-0.009***	-0.009***	-0.009***	-0.009***
	(0.002)	(0.002)	(0.002)	(0.002)
Municipality FE	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes
Ln(Pop/) × Post	Yes	Yes	Yes	Yes
AfD vote share × Post	Yes	Yes	Yes	Yes
Election Controls × Post	Yes	Yes	Yes	Yes
Facebook users p.c × Post	Yes	Yes	Yes	Yes
Broadband internet × Post	Yes	Yes	Yes	Yes
Observations	71,008	71,008	71,008	71,008
Pre-Period Mean of DV	0.12	0.12	0.12	0.12
R^2	0.44	0.44	0.44	0.44

Notes: This table presents the results of estimating municipality-quarter-level regressions as in Equation (3) where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes. *AfD Facebook users p.c. (std)* is the number of AfD Facebook followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. All regressions include municipality and quarter fixed effects, as well as controls for the logarithm of population, the AfD vote share, Facebook users per capita, and broadband internet access, all interacted with *Post*. Robust standard errors in parentheses are clustered at the level indicated at the top of the table. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure D.2: Event Study Hate Crime (Twitter Exposure)



Notes: This figure plots the coefficients from running an event study version of regression Equation (3). The dependent variable is the inverse hyperbolic sine of the number of anti-refugee incidents. Exposure is measured based on the number of AfD Twitter followers per capita in each municipality. The omitted category is the 3rd quarter of 2017, the quarter before the passing of the NetzDG (indicated with the vertical line). The whiskers indicate 95% confidence intervals based on standard errors clustered by county.

Table D.4: Robustness: Social Media Exposure measured with Twitter

	Dep. var.: Asinh(Anti-Refugee Hate Crimes)				
	(1)	(2)	(3)	(4)	(5)
AfD Twitter Followers p.c. (std) × Post	-0.012**	-0.010**	-0.011**	-0.010**	-0.011**
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
AfD vote share (std) × Post		0.033***	0.032***	0.034***	0.029**
		(0.012)	(0.012)	(0.012)	(0.012)
Facebook users p.c (std) × Post			0.002	0.002	0.003
			(0.002)	(0.002)	(0.002)
Broadband internet (std) × Post				0.004	0.000
				(0.004)	(0.004)
Municipality FE	Yes	Yes	Yes	Yes	Yes
Year-Quarter FE	Yes	Yes	Yes	Yes	Yes
Ln(Pop.) × Post	Yes	Yes	Yes	Yes	Yes
Election Controls × Post		Yes	Yes	Yes	Yes
All Controls (19) × Post					Yes
Observations	71,456	71,008	71,008	71,008	68,736
Pre-Period Mean of DV	0.12	0.12	0.12	0.12	0.12
R^2	0.44	0.44	0.44	0.44	0.45

Notes: This table presents the results of estimating Equation (3), where the dependent variable is the inverse hyperbolic sine of the number of anti-refugee hate crimes in a municipality in a given quarter. *AfD Twitter Followers p.c. (std)* is the number of AfD Twitter Followers per capita, standardized to have a mean of 0 and a standard deviation of 1 to ease interpretation. All regressions include municipality and quarter fixed effects as well as a control for the natural logarithm of population, interacted with *Post*. See text for a detailed description of the additional control variables. Robust standard errors in parentheses are clustered by county. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

D.2. Additional Results on Refugee Attitudes

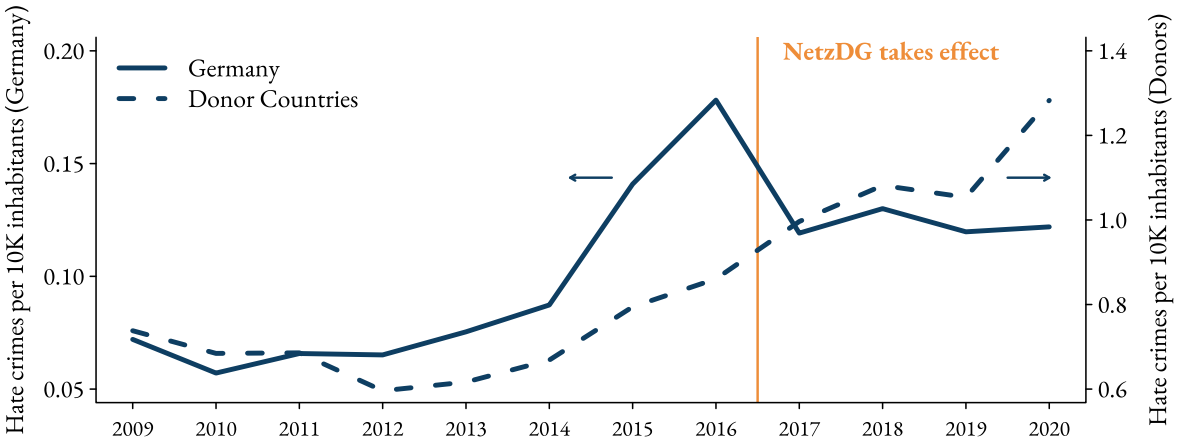
Table D.5: Changes in Action Towards Refugees

	Dep. var.: Helped Refugees by ...						
		Donated		Volunteered		Demonstrated	
	Index	Last Year	Future	Last Year	Future	Last Year	Future
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel (a): All Respondents							
Social Media User × Post	-0.005	-0.004	-0.008	-0.006	-0.011*	-0.006	0.005
	(0.004)	(0.008)	(0.008)	(0.005)	(0.006)	(0.004)	(0.006)
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Survey Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	36,786	36,560	35,586	36,388	35,404	36,336	35,360
Pre-Period Mean of DV	0.17	0.31	0.36	0.07	0.13	0.05	0.09
R^2	0.78	0.73	0.73	0.71	0.68	0.69	0.69
Panel (b): AfD Voters							
Social Media User × Post	-0.002	0.018	0.002	0.010	-0.005	-0.014	-0.013
	(0.011)	(0.020)	(0.021)	(0.014)	(0.015)	(0.015)	(0.024)
Individual FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Survey Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,554	2,544	2,506	2,544	2,502	2,534	2,488
Pre-Period Mean of DV	0.07	0.10	0.10	0.02	0.03	0.05	0.10
R^2	0.68	0.68	0.72	0.69	0.62	0.66	0.62

Notes: This table presents the results of estimating Equation (4), where the dependent variables are different measures for positive actions towards refugees. *Social Media Users* is an indicator for respondents who use social media at least once a week. All regressions include individual and survey year fixed effects as well as a control for the respondent's gender and age, interacted with *Post*. See the text for a detailed description of the variables. Robust standard errors in parentheses are clustered by individual. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

D.3. Additional Synthetic Control Results

Figure D.3: Evolution of Hate Crimes in Germany vs. Donor Countries



Notes: This figure compares hate crimes per 10K inhabitants in Germany vs. the unweighted average in the donor countries in 2009-2020.

Table D.6: Country Weights in the Synthetic Germany

Country	Weight
Austria	0.09
Belgium	0.01
Bosnia and Herzegovina	0
Bulgaria	0
Croatia	0
Cyprus	0
Czech Republic	0
Denmark	0.01
Finland	0
Italy	0.1
Lithuania	0.55
Moldova	0.07
Poland	0.12
Portugal	0
Slovakia	0
Spain	0
Switzerland	0
Turkey	0.03
UK	0
Ukraine	0
US	0

Notes: This table presents the country weights used to generate the synthetic version of Germany for the synthetic control estimates.

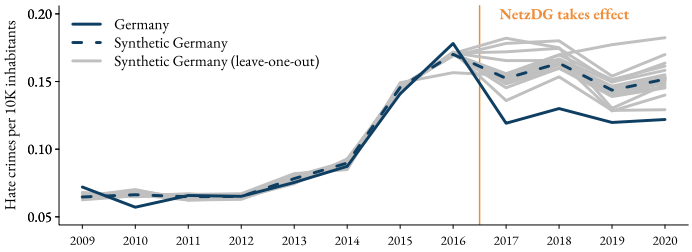
Table D.7: Hate Crimes Predictor Means

Variable	Germany		Donors	OECD	OSCE
	Real	Synthetic			
Hate crimes per 10K inhabitants 2009	0.07	0.06	0.74	1.02	0.71
Hate crimes per 10K inhabitants 2010	0.06	0.07	0.68	0.93	0.66
Hate crimes per 10K inhabitants 2011	0.07	0.06	0.69	0.92	0.66
Hate crimes per 10K inhabitants 2012	0.07	0.07	0.6	0.75	0.57
Hate crimes per 10K inhabitants 2013	0.08	0.08	0.62	0.73	0.59
Hate crimes per 10K inhabitants 2014	0.09	0.09	0.67	0.83	0.64
Hate crimes per 10K inhabitants 2015	0.14	0.15	0.79	1.03	0.76
Hate crimes per 10K inhabitants 2016	0.18	0.17	0.86	1.18	0.83
Measure change 2009-2016	0	0.11	0.11	0.11	0.11

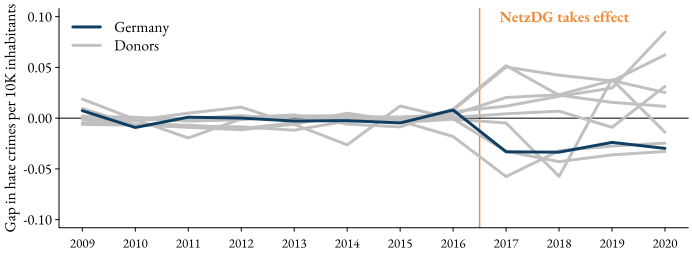
Notes: This table presents the means of the predictor variables for Germany and the synthetic Germany, as well as the simple mean among the donor, OECD, and OSCE countries.

Figure D.4: Leave-One-Out and In-Space Placebos

(a) Germany vs. Synthetic Control

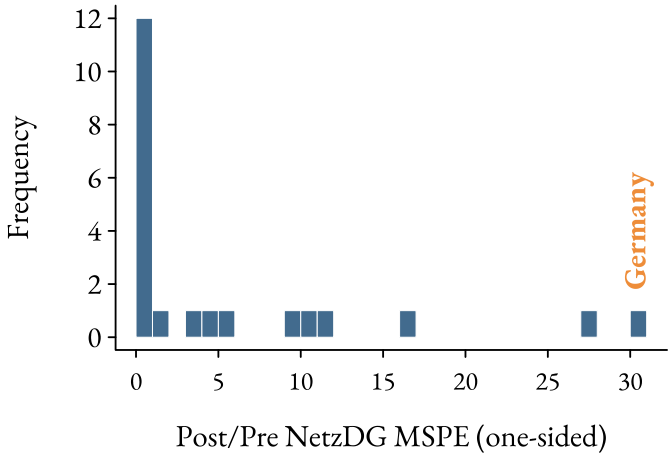


(b) Gaps Between Observed and Synthetic Hate Crimes



Notes: Panel (a) compares hate crimes per 10K inhabitants in Germany vs. the synthetic Germany and a synthetic Germany built by dropping each of the donor countries. Panel (b) shows the gaps between observed and synthetic values for Germany and each of the donor countries acting as a "placebo" treated country. As in Abadie et al. (2010), we drop countries with a pre-NetzDG MSPE higher than 5 times the one of Germany to improve the visibility of the graph.

Figure D.5: Mean Squared Prediction Error Ratios (One-Sided)



Notes: This graph plots the histogram of the ratio between the MSPE post-NetzDG and the MSPE pre-NetzDG. One-sided MSPE are calculated as in Abadie (2021).

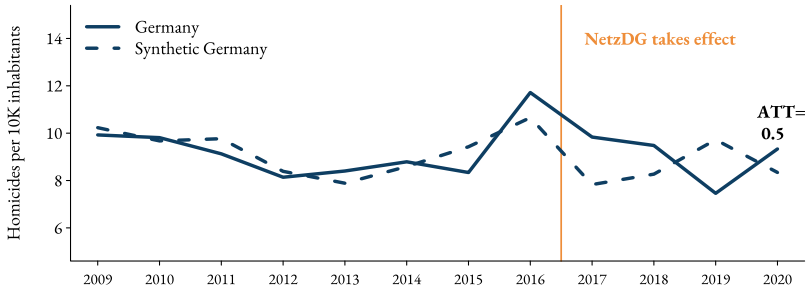
Table D.8: Robustness to Alternative Specifications

Specification	ATT	p-value (one-sided)	p-value (two-sided)	Donors	Pre-NetzDG RMSPE
Baseline	**-0.03**	**0.045**	**0.227**	**21**	**0.005**
Alternative interpolation					
Interpolation dummy	-0.03	0.045	0.182	21	0.005
No interpolation	-0.048	0.167	0.167	11	0.01
Alternative outcomes					
Log	-0.097	0.05	0.1	19	0.005
Levels	-0.047	0.136	0.318	21	0.012
Alternative periods					
Period 2009-2019	-0.051	0.042	0.042	23	0.005
Period 2009-2021	-0.086	0.056	0.111	17	0.007
Alternative donors					
Leave-one-out (max ATT)	-0.014	0.19	0.524	20	0.006
Leave-one-out (min ATT)	-0.048	0.048	0.238	20	0.009
OECD	-0.067	0.067	0.133	14	0.007
No Lithuania	-0.036	0.048	0.143	20	0.006

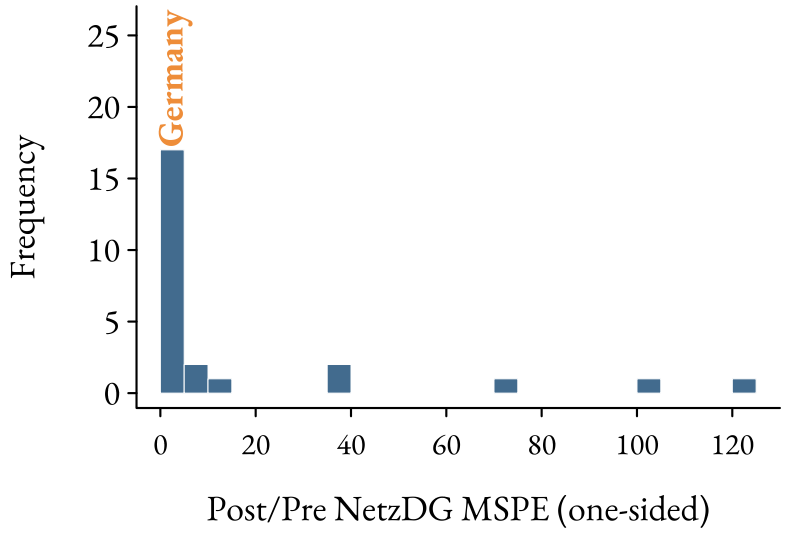
Notes: This table presents estimates of the average treatment effect post-NetzDG, its one- and two-sided p-values, the number of donors and the pre-NetzDG root mean squared prediction error. Note that the ATT and the RMSPE are expressed in hate crimes per 10K inhabitants, to facilitate comparison between specifications. Inference is based on the permutation method of Abadie et al. (2010); see Abadie (2021) for how to compute one-sided p-values. "Interpolation dummy" adds as predictor the pre-NetzDG average of a dummy indicating observations that were linearly interpolated. "No interpolation" keeps only countries without missing values during the period of study.

Figure D.6: Placebo Outcome: Homicides

(a) Synthetic Control Estimates



(b) Mean Squared Prediction Error Ratios (One-Sided)



Notes: Panel (a) shows synthetic control estimates. The figure presents the evolution of homicides per 10,000 inhabitants in Germany and the synthetic Germany. The synthetic control uses all lagged outcomes as predictors, as well as the average of a dummy variable indicating whether there were measurement changes in the hate-crime series in the pre-period. Panel (b) plots the histogram of the ratio between the MSPE post-NetzDG and the MSPE pre-NetzDG. One-sided MSPE are calculated as in Abadie (2021).