Journal of Quantitative Description: Digital Media 2(2022), 1–37 10.51685/jqd.2022.009

Hate speech's double damage: A semi-automated approach toward direct and indirect targets

MARIO HAIM[1]
LMU Munich, Germany

ELISA HOVEN
University of Leipzig, Germany

Democracies around the world have been facing increasing challenges with hate speech online as it contributes to a tense and thus less discursive public sphere. In that, hate speech online targets free speech both directly and indirectly, through harassments and explicit harm as well as by informing a vicious environment of irrationality, misrepresentation, or disrespect. Consequently, platforms have implemented varying means of comment-moderation techniques, depending both on policy regulations and on the quantity and quality of hate speech online. This study seeks to provide descriptive measures between direct and indirect targets in light of different incentives and practices of moderation on both social media and news outlets. Based on three distinct samples from German Twitter, YouTube, and a set of four news outlets, it applies semi-automated content analyses using a set of five cross-sample classifiers. Thereby, the largest amounts of visible hate speech online depict rather implicit devaluations of ideas or behavior. More explicit forms of hate speech online, such as insult, slander,

Mario Haim: haim@ifkw.lmu.de
Elisa Hoven: elisa.hoven@uni-leipzig.de
Date submitted: 2021-06-14

[1] The authors thank the anonymous reviewers as well as, in alphabetical order, Leonhard Eckwert, Hannah Heuser, Inga Jahn, Berit Kruse, Sophie Menner, and Alexandra Witting.

or vulgarity, are only rarely observable and accumulate around certain events (Twitter) or single videos (YouTube). Moreover, while hate speech on Twitter and YouTube tends to target particular groups or individuals, hate speech below news articles shows a stronger focus on debates. Potential reasons and implications are discussed in light of political and legal efforts in Germany.

Keywords: hate speech online, user comments, comment moderation, platforms, Twitter, YouTube, online journalism, Netzwerkdurchsetzungsgesetz, NetzDG, online public sphere

For democracies around the world, hate speech online has been an immanent challenge for several years. Recent research alone has brought forward various definitions (e.g., Siegel, 2020) and multi-dimensional conceptualizations (e.g., Coe et al., 2014), has pointed out its wide distribution among user comments across both social media (e.g., Matamoros-Fernández & Farkas, 2021) and online news outlets (e.g., Schabus et al., 2017), has improved its automated detection (e.g., Stoll et al., 2020) and subsequent moderation (e.g., Boberg et al., 2018), and has investigated potential impacts on both direct (e.g., Chen et al., 2020) and indirect (e.g., Prochazka et al., 2018) targets.

Importantly, hate speech's "unnecessarily disrespectful tone toward the discussion" (Boberg et al., 2018, p. 59) as well as its intentional design "to attack someone or something and, in doing so, incite anger or exasperation" (Ksiazek et al., 2015, p. 854) contributes to a disparaging public sphere in two ways. First, in addressing people or groups of people directly, hate speech online is apt to discourage individuals from participating in the discourse (Gelber, 2019). This has become evident particularly for the discourse around publicly visible politicians (Kalsnes & Ihlebæk, 2020). Seminally, the 2017 Network Enforcement Act ("Netzwerkdurchsetzungsgesetz" or "NetzDG") was installed in Germany to aid prosecution of hate speech online and thereby take providers of social media more into obligation to moderate users' negative comments (Zurth, 2020). Second, in breaking with established norms of communicative practice via channels of public communication, hate speech online is also apt to indirect damage, in that it informs a vicious environment of irrationality, misrepresentation, and disrespect (Papacharissi, 2004; Stroud et al., 2015). As such, hate speech inhibits a democracy's necessary public discourse as it hinders a collective communicative mode of viewpoint negotiation (Wessler, 2018). Especially news outlets perceiving user comments as "vehicles for accomplishing deliberative ideals" (Reich, 2011, p. 102) thus feel required to moderate their forums not only to keep off negative comments but also to encourage constructive discourse with and among users (Loosen et al., 2017).

This study takes on this dual perspective of hate speech between direct and indirect targets in light of different practices of moderation on both social media and news outlets. Focusing on the case of Germany, it asks for three exemplary datasets, (1) which forms of hate speech online are most prominent and (2) which targets are mentioned to what extent?

The manuscript starts by discussing different conceptualizations of hate speech online before comparing different modes of moderation. In that, the well-known case of the NetzDG in Germany is closely examined before investigating it empirically through descriptive measures of various forms of hate speech online by means of a semi-automated content analysis of three very distinct datasets (Twitter, YouTube, news media). Finally, the exploratory and descriptive results are discussed with regard to each platform's direct and indirect targets.

Hate Speech Online

There is no scholarly consensus on whether hate speech online is a well-defined concept, let alone agreement on a specific definition (Gagliardone et al., 2015). For example, while George (2015, p. 1) sees "no standard definition of hate speech," he concurs that the term usually covers "forms of expression aimed at persecuting people by vilifying their racial, ethnic, or other identities." Cohen-Almagor (2011, p. 1) also includes intention when defining hate speech as "bias-motivated, hostile, malicious speech" and Pereira-Kohatsu and colleagues (2019, p. 4654) define hate speech as "speech that denigrates a person or multiple persons based on their membership to a group, usually defined by race, ethnicity, sexual orientation, gender identity, disability, religion, political affiliation, or views." As such, Siegel (2020, p. 59) describes a current state of "definitional ambiguity" in which Davidson and colleagues (2017, p. 512) state that "no formal definition exists but there is a consensus that [hate speech online] is speech that targets disadvantaged social groups in a manner that is potentially harmful to them." Arguably, this also includes "statements that attack, intimidate or denigrate others […] based on their fulfillment of a certain role" (Obermaier et al., 2018, p. 503).

At a bare minimum, then, available understandings of hate speech online align in three regards. First, hate speech is an act of expressing degradation against a target, such as another person or a group of people. Second, hate speech is usually defined to require a qualification as to which premises seemingly justify the posed degradation—common denominations include references to disability, ethnicity, the fulfillment of a certain role, gender, race, religion, sexual orientation, political affiliation, or certain views. Third, while the majority of definitions implicitly refers to a certain level of intentionality insofar as that hate speech online is intentionally employed to attack, denigrate, intimidate, persecute, promote hatred, or target in potentially harmful ways, most definitions explicitly focus on manifest expressions.

For example, in a prominent study Coe and colleagues (2014) manually coded five specific "features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topics" (2014, p. 660) within English user comments posted in late 2011 to the U.S.-based journalistic outlet Arizona Daily Star. They found incivility in 22 percent of their sample of 6,535 user comments, most of which utilized exactly one of the five forms. Most prevalent were disparaging words, which appeared in 14 percent of the whole sample and thus accounted for more than half of uncivil comments. In addition, disparaging remarks about the way in which another person communicates as well as implying that an idea, plan, or policy was disingenuous accounted for 2 percent of comments each.

Given this manifest operationalization, subsequently, several approaches were taken to (semi-)automate the identification of hate speech online. For example, Ozalp and colleagues (2020) employed an automated content analysis to investigate antisemitic hate speech online on Twitter, identifying 0.7 percent of a total of 1,232,744 analyzed English tweets as antagonistic toward Jews and Jewish identity. Methodologically, the authors converted collected tweets into unigram tokens, noun phrases, and vector-based dependency representations before feeding this accumulated set of features to supervised machine-learning algorithms. In that, their findings are very much in line with Williams and Burnap (2016) who found 1 percent of 210,807 English tweets in 2013 to contain ethical or religious hate speech. Methodologically, Williams and Burnap (2015) also employed a supervised machine-learning setup with features on unigram tokens, noun phrases, and vector-based dependency representations.

Crucially, these studies entail a certain discrepancy between their definition of hate speech online and their empirical investigation. That is, while their definitions usually subsume intentionally expressed degradations against targets, empirical investigations typically build on data collected after seminal steps of moderation. Between publication of a comment and collection of its data for analysis, then, directly attacked targets might have taken legal steps to have respective hate speech removed. Thus, analyzed hate speech online might represent the legally less relevant remainder of commentary. Yet, this is by no means socially less relevant as such hate speech likely fosters a viciously irrational environment of misrepresentation and disrespect (Papacharissi, 2004; Stroud et al., 2015). As "societies evolve through contestation and disagreement" (Gagliardone et al., 2015, p. 15), post-moderation hate speech online indirectly affects democratic discourse by countering collective communicative modes of viewpoint negotiation (Wessler, 2018).

Moderating Hate Speech Online

To conform with legal requirements as well as to counter such discouraging discourse to some extent, providers of online public spheres have implemented varying degrees of comment moderation. Yet, the lack of definition effectively requires providers to define hate speech for themselves, through their own community guidelines. Among the most influential platform providers, Twitter (2021) states that one "may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease." It thus focuses primarily on targeted individuals. In slight contrast, Facebook (2021) claims to "remove language that incites or facilitates serious violence" while also "prohibit[ing] people from facilitating, organizing, promoting, or admitting to certain criminal or harmful activities targeted at people, businesses, property or animals." Facebook thus omits listing qualifiers to posed degradation. As a third example, YouTube (Google, 2021) claims to remove "content promoting violence or hatred against individuals or groups based on any of the following attributes: age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender, sexual orientation, victims of a major violent event and their kin, veteran status."

In contrast to social media platform providers, various news outlets seek to moderate more constructively to encourage civil public discourse (Loosen et al., 2017). Expressed through their own policies and guidelines (Ksiazek, 2015), news outlets' motivation for moderation is thereby different to platform providers' motivation, and so is their practical approach between authority and discourse (Wintterlin et al., 2020). News outlets which adhere to discursive ideals seek to form users' "reasoned opinion expression on a social or political issue" (Stromer-Galley, 2007, p. 3) through requesting, where applicable, argumentation on topic, respectful tone, evidence for claims, and consideration of others' opinions (Dahlberg, 2001; Stromer-Galley, 2007). They employ an active moderation style, sometimes including participation in discussions (Wintterlin et al., 2020). Conversely, news outlets adhering to an authoritative approach maintain a hierarchy between them and their users where moderation often manifests in guarding and, in cases of misconduct, pointing users to community guidelines (Wintterlin et al., 2020).

The Case of Germany

In addition to theoretical definitions, empirical operationalizations, and providers' interpretations, criminal law holds various offences applicable to hate speech online (e.g., European Court of Human Rights, 2020; No Hate Speech Movement Deutschland, 2021).

In Germany, among others, this includes insult, defamation, and slander (StGB § 166, 185, 186, 187), threat (StGB § 126, 241), approval of and incitement to violence (StGB § 111, 140) as well as incitement to hatred (StGB § 130). And while these statutes do not explicitly refer to hate speech online, Germany's 2017 Network Enforcement Act ("Netzwerkdurchsetzungsgesetz" or "NetzDG") as well as the European Union's Convention on Cybercrime and the 2021 German law to combat right-wing extremism and hate crime particularly aim at adhering to the specifics of the online environment (Banks, 2010). The NetzDG and its subsequent adjustments require social media platform providers to maintain transparent means for complaints, reporting, and deletion of unlawful content. In aiming at a more feasible prosecution of hate speech within highly fragmented and personalized online environments by means of heavy financial punishment, the NetzDG has gained global attention in serving as debatable blueprint (Brown, 2018; Heldt, 2019; Zurth, 2020).

The NetzDG is very specific in requiring means for complaints which need to be guaranteed "a reaction on 'manifestly unlawful content' within 24 hours" (Heldt, 2019, p. 6) where unlawfulness explicitly refers to, among others, insult, defamation, slander, threat, or incitement to violence and hatred (i.e., StGB §§ 86, 86a, 89a, 91, 100a, 111, 126, 129-129b, 130-131, 140, 166, 184b, 185-187, 201a, 241, 269). The combination of these requirements and the volume of usage on said platforms "effectively necessitates their use of automated systems to detect illegal or otherwise problematic material" (Gorwa et al., 2020, p. 2). In turn, this has raised major criticism in that "requiring Internet providers to engage in content moderation by law is referred to as collateral censorship" (Zurth, 2020, p. 31).

Moreover, critics of the NetzDG have raised concerns about obliging commercial and mostly U.S.-based companies to act as preliminary legal judges which, in turn, may nudge them toward even stricter community guidelines that may compromise on others' freedom of expression (Dias Oliva, 2020; Kasakowskij et al., 2020). A long-standing principle to fundamental rights, freedom of expression has thus significantly shaped the policy discourse around the NetzDG (He, 2020). Indeed, the seminal transparency reports, which the providers are required to publish under the NetzDG, "show that social media platforms tend to moderate content on the grounds of their own community guidelines more than on the basis of national criminal law" (Heldt, 2019, p. 8). Summing up 2018 and 2019, the three most influential providers in Germany (Newman et al., 2020), Facebook, Twitter, and YouTube, have dealt with a total of 2,921,553 complaints, 28 percent of which resulted in blocking (Zurth, 2020, p. 21f.).

News outlets, while not included in the NetzDG, are equally required to take down unlawful content as a result from general legal principles (Zurth, 2020, p. 18f.). Without the requirement for transparent means for complaints and reporting, news outlets are, however, typically more inclined toward maintaining a constructive atmosphere for public discourse. In that, they usually restrict user commentary to certain topics or articles whereby some outlets police deviating comments (i.e., authoritative approach to moderation) vis-à-vis others who more actively engage in discussions to encourage viewpoint negotiation (i.e., discursive approach to moderation).

Research Questions

The diverse landscape of definitions on hate speech online makes it difficult to discuss and compare empirical research on this matter (Matamoros-Fernández & Farkas, 2021). Moreover, a lack of definition from legislators inhibits adequate policy debates on empirical grounds. Instead, platform providers have been identified as key actors in this regard as they have subsequently defined online hate regulations for themselves to implement automated comment moderation. Facing similar hate speech issues, news outlets as another key actor for public discourse, however, have been shown to employ different ideals and practices for manual or semi-automated comment moderation. This study thus sets out to provide empirical grounds of post-moderation hate speech online and its multitude of definitions. It considers different types of platforms with their respective moderation styles (i.e., social media and news outlets) but focuses on publicly visible user commentary (i.e., leaving out Facebook) in three very distinct datasets. Hence, the first research question asks:

RQ1: Which forms of hate speech online are most prominent in German user comments on (a) Twitter, (b) YouTube, and (c) news outlets?

As different forms of hate speech online have been shown to affect both direct and indirect targets, it is important to investigate particularities in greater detail. As such, the case of politicians has been mentioned repeatedly as prominently and directly affected "elitist" figures (e.g., Boberg et al., 2018; Kalsnes & Ihlebæk, 2020). In commenting publicly to threads aimed at or opened by politicians, one could assume that the perceived status of a politician's elitism affects the forms and prominence of hate speech online, for example in that greater influence raises stronger forms of hate speech online. Likewise, commenting to news articles from an outlet with a specific political slant might also be reflective of various levels of hate speech online. Yet, in lacking stronger indications, this study generally asks:

RQ2: Which forms of hate speech online are most prominent in German user comments addressing various (a) state-level politicians, (b) communal-level politicians, and (c) news outlets?

Furthermore, as hate speech online can both directly and indirectly contribute to a vicious online public sphere, the prominence of various forms of hate speech online (RQ1) needs to be put into perspective of its particular targets. Hence, this study finally asks:

RQ3: What are main targets of German user comments with prominent forms of hate speech online on (a) Twitter, (b) YouTube, (c) and news outlets?

Method

This study seeks to generate descriptive measures on three distinct datasets through a semi-automated quantitative content analysis of user comments. The three employed datasets are hardly comparable in their compilation but individually allow for quantitative insights into forms of hate speech online with a focus on three different levels of political issues. This most-different systems approach thus subsumes various platform logics, moderation styles, and norms of user commentary in Germany. A sample of each of the three datasets was coded manually for different forms of hate speech online and was subsequently used as training data for deep-learning techniques to come up with cross-dataset classifiers to classify various forms of hate speech online in the remainder of the data. Then, comments entailing various forms of hate speech online were clustered and analyzed using structural topic modelling to identify main targets.

Models, scripts, and document-term matrices have been shared via the Open Science Framework under https://osf.io/dgztn/.

Data

First, given that Twitter is built around follower networks, Twitter data is based on 1,004 out of 1,868 German members of state parliaments (MSP) who maintain accounts for others to be followed. The lists of accounts for all 16 German states were taken from "@wahl_beobachter," a prominent advisor for digital political communication who maintains up-to-date directories of Twitter handles for German state-level politicians. For each politician's account, then, all tweets authored by or addressed at it were collected via the Twitter API on a daily basis during the full month of February 2020. In total, this procedure yielded $N_{Twitter}$ = 153,761 tweets, authored by a total of 33,685 individual users. Importantly, the 16 states did neither yield equal nor proportionally adequate numbers of tweets. That is, when compared to the numbers of MSP, the states of Berlin (18% of all tweets in the dataset vis-à-vis 9% of all of Germany's MSP), Schleswig-Holstein (24% of tweets, 4% of MSP), and Thüringen (24% of tweets, 5% of MSP) are largely overrepresented in the Twitter dataset whereas all other states are slightly underrepresented with Niedersachsen holding the smallest number of tweets in the dataset (1% of tweets, 7% of MSP). This bias, however, can be attributed to the varying numbers of MSP who maintain Twitter accounts—the amounts of MSP on Twitter and the amounts of tweets collected per state correlate highly (Pearson's $r = .87$; $df = 14$; $p < .001$).

Second, YouTube data is based on all 339 German mayors of large cities ("Oberbürgermeister:innen") as listed by Wikipedia in June 2020. For each mayor, suffixed by the respective city, the YouTube search function was used via its API to query for videos (e.g., "Michael Müller Berlin"). Per search query, the first ten found videos were collected and, for each found video, all user comments were collected. Similar to Twitter's follower structure, also this procedure should loosely reflect user behavior in that YouTube is less built around follower networks but more about search and recommendation; that is, users interested in local politics likely consult the platform's search function to follow recommended videos. Data collection took place in June 2020 and yielded $N_{YouTube}$ = 16,973 user comments, authored by a total of 11,911 individual users, yet adhering to only 830 videos. That is, while all but two YouTube search queries for city mayors yielded videos (*M* = 9.5; *SD* = 1.77), a majority of 2,356 out of 3,186 videos did not contain any user commentary. The remaining 830 videos, then, received user comments following a long-tail distribution (*M* = 20.4; *SD* = 162.00) without systematic differences between either female/male mayors (*t* = -0.34; *df* = 335; *p* = .735) or states (*F* = 0.54; *df* = 15, 321; *p* = .918).

Third, news outlet data is based on user comments from four major German news outlets reflective of the political spectrum. That is, faz.net and welt.de were included as (somewhat) more conservative outlets while zeit.de and taz.de were included as (somewhat) more liberal outlets. According to anecdotal evidence as well as community guidelines, at the time of publication, all four news outlets engaged in mostly manual comment moderation following either a more authoritative or a more discursive approach. Again, in loosely resembling user behavior this data is based on commentary below news articles which come up after searching for a certain topic. As such, climate has been chosen due to its potential to polarize individuals' perceptions of science (Anderson & Huntington, 2017). To collect commentary, the German term for climate ("klima") was searched on Google News with filters set to each of the outlets and to the time span under investigation. All result pages on Google News were included and data collection took place in February 2020. Per outlet, then, all articles published from the beginning of February 2019 until the end of January 2020 were collected, which roughly covers the first year of the "Friday for Future" movement's global presence. Per article, it was coded whether an article was marked as an opinion piece (11%) or not (89%). Ultimately, for each article, each comment was scraped, resulting in a total of N_{news} = 455,003 comments posted and publicly visible below 3,022 out of a total of 4,686 articles. Thereby, comments per article (*M* = 151.0; *SD* = 258.00) vary slightly over time with highest values in May 2019 (global strikes), late June 2019 (congress on coal), and mid-September (UN Climate Action Summit).

It is noteworthy that all data was collected with some delay to its publishing. As such, default moderation has already been carried out and in some cases targeted or authoring users might have had some of their comments removed before ending up in the current study's datasets. Prior research indicates that such self-deletion could yield biases in that comments with negative sentiment have a slightly higher tendency of being deleted (Schatto-Eckrodt et al., 2020). As such, the current datasets are neither representative nor do they depict a wholesome array of hate speech online. Instead, the current datasets provide three distinct yet large-scale insights into forms of hate speech online which remain visible and as such are apt to negatively affect public discourse. Due to potential deletion biases, these insights likely underestimate actual shares of hate speech online.

Coding

A total of ten categories created for the purpose of this study was coded to capture hate speech online. With respect to the available plethora of definitions, from legal and providers' perspectives, a first set of categories focused on the expression of degradation. To capture a wide array of degradation, categories included (1) insult as the use of malicious vocabulary directed at people or groups of people (cf. Davidson et al., 2017), (2) devaluation as derogatory labels for ideas or behavior of other people or groups of people (cf. Obermaier et al., 2018), and (3) devaluation as derogatory labels for the communication of other people or groups of people (cf. Coe et al., 2014). A second set of categories, then, covered qualifiers as to which premises seemingly justify posed degradations. As such, degradation with references to (4) gender, including either hostile or benevolent references to gender identification or sexuality (cf. Chen et al., 2020; Eckert & Metzger-Riftkin, 2020), (5) racism, including hostile references to physical appearance and/or the belonging to minorities (cf. George, 2015; Nielsen, 2002), and (6) religion (cf. Pereira-Kohatsu et al., 2019; Rosenfeld, 2003) were included. Moreover, (7) slander was coded when bias-motivated allegations were raised to potentially disparage, degrade, or jeopardize the credit of others (cf. Cohen-Almagor, 2011). Finally, a third set of categories subsumed further forms of posed degradations. That is, (8) vulgarity was coded as unnecessarily disrespectful profanity (cf. Boberg et al., 2018; Coe et al., 2014). Lastly, (9) approval of and incitement to violence or hatred, public appeal of violent acts (StGB § 111, 130, 140), and (10) psychological and/or physical violence or other criminal acts being credibly and actively announced (StGB § 126, 241), were included as they refer to various laws in Germany; however, while incitement to violence is not explicitly mentioned by the platforms' community guidelines, threat is.

Eleven coders participated in the coding process. After an extensive coder training, they coded an intercoder reliability sample subsuming all three datasets and totaling 373 comments ($n_{news} = 163$; $n_{Twitter} = 111$; $n_{YouTube} = 99$). Due to heavily imbalanced occurrences of each category and each category's dichotomous measurement, intercoder reliability was calculated using a range of four different coefficients: First, Fleiss' κ was used as a metric particular to nominal data coded by multiple coders (Fleiss, 1971). Second, Krippendorff's α was employed as building atop κ and representing one of the most common coefficients despite its tendency to penalize for heavy imbalance (Krippendorff, 2011). Third, to overcome this tendency to be "profoundly influenced by the frequency distribution of the units being scored" (Quarfoot & Levine, 2016, p. 383) von Eye's κ_s as an extended version of Brennan and Prediger's Kappa which corrects for chance agreement was employed (Brennan & Prediger, 1981; von Eye, 2006). Fourth and in seminal vein, Fretwurst's Lotus, denominated as Lambda (λ), was used as it employs a different approach to account for imbalanced data by focusing on most-commonly coded values per coding unit (Fretwurst, 2015).

Almost all categories in the sample resembled more or less strongly imbalanced distributions of positive and negative cases. Intercoder reliability, then, was considered sufficient for the category of insult (n = 153 comments coded as insulting; $\alpha = .64$; $\kappa = .64$; $\kappa_s = .73$; $\lambda = .94$); it was mediocre yet acceptable given the number of cases for vulgarity ($n = 28$; $\alpha = .65$; $\kappa = .65$; $\kappa_s = .94$; $\lambda = .98$), devaluation of ideas or behavior ($n = 244$; $\alpha = .51$; $\kappa = .51$; $\kappa_s = .56$; $\lambda = .88$), and devaluation of communication ($n = 64$; $\alpha = .55$; $\kappa = .55$;

κ_s = .82; λ = .94). Intercoder reliability was inacceptable or even unfeasible with respect to the given numbers of cases, however, for incitement to violence (n = 4; α = 1.0; κ = 1.0; κ_s = 1.0; λ = 1.0), slander (n = 3; α =.49; κ = .49; κ_s = .77; λ = .94), threat (n = 3; α = 1.0; λ = 1.0; κ and κ_s not computable), and degradation with regard to religion (n = 4; α = 1.0; κ = 1.0; κ_s = 1.0; λ = 1.0), racism (n = 9; α = .66; κ = .66; κ_s = .97; λ = .99), and gender (n = 1; α = 1.0; λ = 1.0; κ and κ_s not computable).

The coders then manually coded a total of 11,226 comments which were randomly sampled per corpus (n_{news} = 6,226; $n_{Twitter}$ = 2500; $n_{YouTube}$ = 2500). Similar to the intercoder reliability sample, five categories did not yield noteworthy distribution within this larger sample. That is, incitement to violence (n = 16), threat (n = 15), and degradation with regard to religion (n = 17), racism (n = 50), and gender (n = 46) were each coded in less than one percent of all manually coded comments. While this is an interesting finding in itself, the small amounts of training data prohibit any subsequent supervised machine learning; these five categories were thus dropped for classification. In contrast, devaluation of ideas or behavior (n = 1,602), insult (n = 1,038), devaluation of communication (n = 260), and vulgarity (n = 133) showed feasible amounts as well as acceptable levels of intercoder reliability and were thus taken forward. Finally, slander showed less ideal intercoder reliability based on a very small amount of cases (n = 3; α =.49; κ = .49; κ_s = .77; λ = .94) but was more prominent in the larger sample (n = 475); it was thus included in the classification approach while requiring special caution.

Classification

A plethora of work has already been done on automated hate speech classification. Given previously promising results (Pereira-Kohatsu et al., 2019), five neural networks, one each per category, were trained. Each network was trained on a combined corpus of all comments from all three datasets to optimize for general instead of platform-dependent hate speech detection.

Following previous literature, every comment was reduced into seven types of features. First, the length of each comment was included as both the count of words and the count of sentences (Pereira-Kohatsu et al., 2019). Second, part-of-speech grammatical word tagging was employed through the spaCy (Honnibal et al., 2020) and spacyr (Benoit & Matsuo, 2020) packages. As such, sixteen relative shares per comment of adjectives, pre-/postpositions, adverbs, auxiliaries, coordinating and subordinating conjunctions, determiners, interjections, nouns as well as proper nouns, numerals, particles, pronouns, punctuation, verbs, and other words were included. Third, vector-based named entity recognition as provided by spaCy's "de_core_news_lg" model was employed for all four available features holding relative shares of locations, organizations, persons or families, and miscellaneous entities such as events, nationalities, products or works of art. Fourth, noun phrases—that is, a noun and its describing words—were extracted using spaCy's dependency parser (Pereira-Kohatsu et al., 2019). Noun phrases that appeared in more than half a percent of comments yet in less than half of all comments were included and transformed into term frequency-inverse document frequencies (tf-idf). This procedure yielded a total of 97 noun-phrase features. Fifth, German terms common to hate speech online as collected and categorized by Hatebase.org, a commercial database of so-called malignant public conversations, were employed (Davidson et al., 2017; Stoll et al., 2020). As such, vocabulary indicative of hate was transferred into relative-share features per comment referencing nationality, ethnicity, religion, gender, sex, disability, and class, as well as a sum totaling the seven Hatebase.org categories. In addition, the concatenated German Hatebase.org vocabulary was located in a 300-dimensional continuous bag-of-words vector space, using the fastText model implementation in spaCy. After also locating each comment in the same vector space, one additional feature was derived as a comment's cosine distance to the Hatebase.org vector (Nithyanand et al., 2017). Sixth, two dictionaries were employed for sentiment-based features (Watanabe et al., 2018). For arousal, the average per-word ratings based on the BAWL-R dictionary was used (Võ et al., 2009). For valence, positive and negative words were counted based on a recent multi-dictionary compilation (Rauh, 2018) and divided by each comment's word count. As an additional feature, the ratio between positive and negative words was included. Seventh, unigram tokens were converted to lowercase and lemmatized, again using the spaCy packages. Stop words and HTML notations as well as those features that appeared in less than half a percent of comments or in more than half of all comments were removed. This yielded another 422 features which were also transformed into term frequency-inverse document frequencies (Pereira-Kohatsu et al., 2019; Watanabe et al., 2018).

Table 1. Classification Performance Measures.

Category	**Intercoder agreement**				**Training share**	**Acc.**	**P**	**R**	**F1**	**AUC**
	α	**κ**	**κ_s**	**λ**	**[%]**					
Devaluation of ideas/behavior	.51	.51	.56	.88	14.3	.95	.84	.79	.82	.88
Insult	.64	.64	.73	.94	9.2	.96	.86	.73	.79	.86
Devaluation of communication	.55	.55	.82	.95	2.3	.99	.85	.64	.73	.82
Vulgarity	.65	.65	.94	.98	1.2	.99	.85	.45	.59	.73
Slander	.49	.49	.77	.94	4.2	.98	.85	.73	.79	.86

Note. Intercoder agreement refers to Krippendorff's α, Fleiss' κ, von Eye's κ_s as an extended version of Brennan and Prediger's Kappa, and Fretwurst's Lotus (denominated as Lambda, λ). Training share indicates the percentage of comments among the manually coded 11,226 comments that hold a respective category. Accuracy (Acc.), precision (P), recall (R), and F1 scores as well as area-under-the-ROC-curve values are all based on binary outcome decisions, coding 1 if an output probability exceeded 50 percent.

All comments along with the total of 554 features (sparsity of 94%) were then used to train each of the five neural networks on (1) the devaluation of ideas or behavior, (2) insult, (3) devaluation of communication, (4) vulgarity, and (5) slander. For each category, the 11,226 coded comments were split into a training set (75% of comments) and a test set (25%). Thereby, the training sets were compiled as weighted random samples to resemble a respective category's shares (e.g., as 133 out of 11,226 or 1.2% of comments were coded as vulgar, the training set for vulgarity was compiled to also include 1.2% of vulgar comments). All neural networks were initialized with one hidden layer containing 40 units, random weights between -.5 and +.5 along a decay parameter of .1, and a maximum number of 50,000 iterations although all five training processes converged at around 8,400 iterations. Finally, a category was coded as "1" if the respective neural network yielded an output probability of more than 50 percent. In line with similar classifiers (e.g., Pereira-Kohatsu et al., 2019; Stoll et al., 2020), this procedure yielded acceptable performance measures for four categories. Given that imbalanced training data likely distorts interpretability of the F1 score, also areas-under-the-ROC-curve values are reported. Notably, vulgarity yielded a sufficient performance score at the cost of a slightly lower recall. Taken with care throughout the discussion, though, all five categories were applied to all remaining comments as their area under the ROC curve values suggested severe improvement vis-à-vis a classification by chance (Table 1).

Clustering

To address the third research question, various structural topic models were fitted among comments coded as either form of hate speech online. Structural topic models (STM) are a method to infer structure across texts from co-occurrences of groups of words, resulting in representations of comments as a mix of topic affiliations (Günther & Quandt, 2016; Roberts et al., 2019). To account for the various selection criteria and the different platforms' affordances and to select the number of topics, a total of 93 topic models, from $k = 10$ to $k = 40$ for each corpus, was estimated. Per corpus, lemmatized and lowercase terms were included if they appeared in at least half a percent of corpus comments. Based on semantic coherence and exclusivity, we then selected the 23-topic model for Twitter (based on 238 terms), the 20-topic model for YouTube (based on 464 terms), and the 22-topic model for news commentary (based on 804 terms). Qualitative assessment then led to further reduced complexity in that overlapping topics were merged and aligned across corpora, yielding a final set of five topics resembling a focus on (1) individuals such as "Greta" or "Merkel," (2) parties or groups of individuals such as "CDU" or "Regierung" (government), (3) events such as an election or a summit, (4) debates pertaining to "Generationenvertrag" (intergenerational equity) or sustainable travelling, or (5) other. Ultimately, although STM estimates term distributions, each comment was assigned its most prominent topic only (by means of document-topic loadings theta) for easier interpretation.

Results

Among all three datasets, hate speech online and its individual forms (RQ1) are visible to a moderate extent. Overall, 21 percent of the tweets, 16 percent of the YouTube comments, and 23 percent of comments below news articles hold at least one of the five coded categories of hate speech online (Figures 1-3). These shares of user commentary containing hate speech can largely be attributed to a devaluation of ideas or behavior—a category most visible in comments below news articles (a total of 16% of such user comments fall into that category). While the devaluation of communication as well as vulgarity and slander only play marginal roles, insult accounts for 4 percent in each of the three datasets (for examples of each category, see Appendix A3). Similar to earlier U.S.-based findings by Coe and colleagues (2014), most comments with identified hate speech utilized exactly one of the five forms.

With respect to various state-level and communal-level politicians as well as individual news outlets (RQ2), distinct patterns across Twitter, YouTube, and the news become evident. In the Twitter dataset, results vary barely. The shares of idea-devaluing, insulting, or slandering hate speech is mostly constant across the MSP's originating states. Noticeable exceptions are the Saarland with 35 percent of tweets identified as hate speech and, to a smaller extent, Schleswig-Holstein where one-in-four tweets identified as hate speech. Both exceptions can be attributed to individual events, though. In the case of Saarland, its prime minister Tobias Hans prominently called for his party (CDU) to move more toward the political center, raising a lot of hate speech on February 11-14. For Schleswig-Holstein, two public speeches by MSP Ralf Stegner (SPD) articulating a need to push back right-wing extremism and populism raised major increases in detected hate speech.

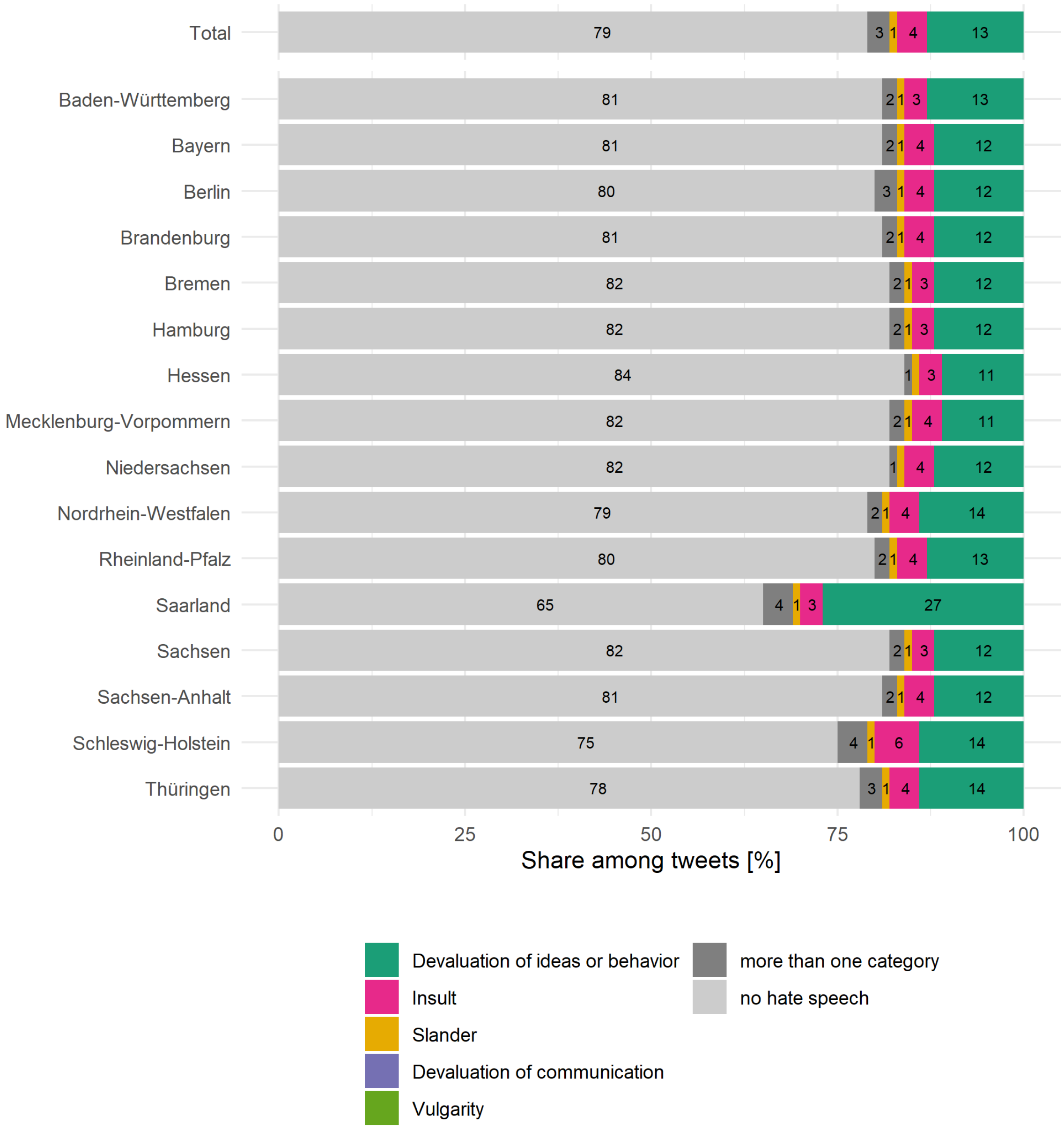

Figure 1. Share of hate speech online among tweets.
Note. Based on all tweets (*N* = 153,761).

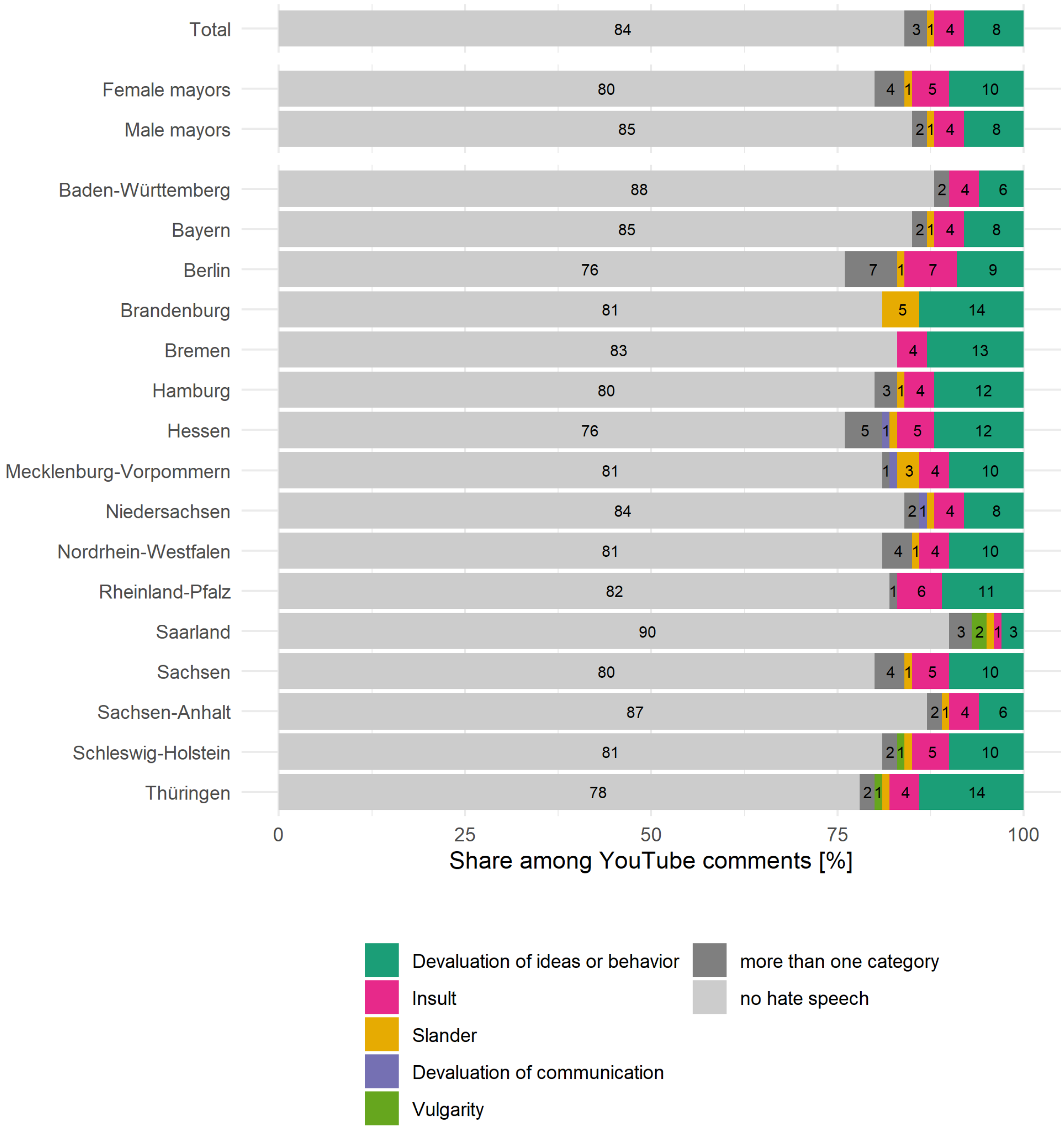

Figure 2. Share of hate speech online among YouTube comments.
Note. Based on all YouTube comments ($N = 16{,}973$).

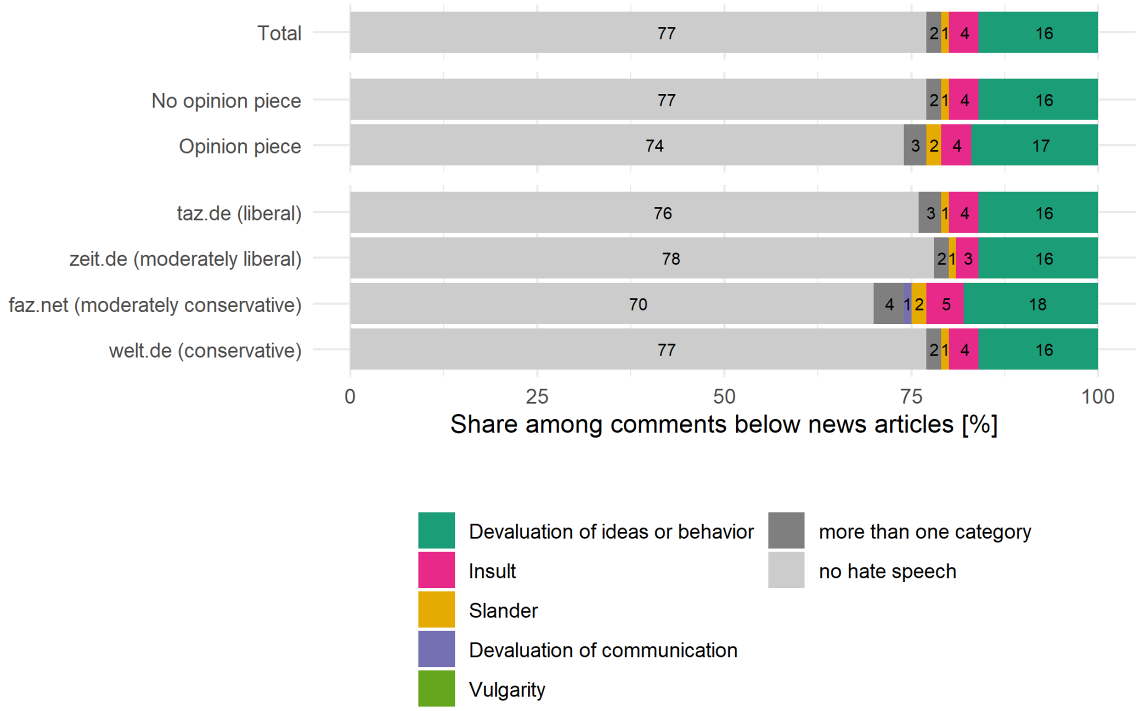

Figure 3. Share of hate speech online among comments below news articles.
Note. Based on all news articles' comments (*N* = 455,003).

On YouTube, the small number of videos renders variation across states very much as variation across single videos and thus as variation across individual mayors. For example, the 24 percent of comments containing hate speech among videos appearing for mayors in Hessen can be attributed largely to one video uploaded by Russia's state-driven German broadcaster "RT DE" in which Hanau's mayor Claus Kaminsky reacts to the racist shootings in February 2020. Similar dependencies on single videos apply to Bayern, Berlin, Brandenburg, Hamburg, Niedersachsen, Nordrhein-Westfalen, Sachsen-Anhalt, Schleswig-Holstein, and Thüringen. In contrast, the remaining six states as well as the diverging results between female and male mayors are not single-video artifacts but depict broader trends. That is, videos about female mayors, such as Pia Findeiß (from the city of Zwickau), Simone Lange (Flensburg), Barbara Ludwig (Chemnitz), Henriette Reker

(Köln), Carda Seidel (Ansbach), or Katja Wolf (Eisenach) face more-than-usual user comments containing hate speech, even after moderation.

Below news articles, comments devaluating ideas or behavior mark the largest share of hate speech observed in this study. Variation between opinion and other pieces as well as across outlets is small. The only notable exception is moderately conservative faz.net which seems to moderate commentary more loosely as the slightly higher shares in all categories are not attributable to single events or articles but are visible across the whole time span.

Targets of hate speech online (RQ3) vary notably across the three datasets whereas variation between the five coded categories of hate speech is less prominent. For Twitter, between 8 and 13 percent of tweets containing hate speech focus on individuals, making individuals the smallest share of targets in this dataset. More often, hate speech on Twitter addresses parties/groups, events, or debates. On YouTube, between 15 and up to 27 percent of hate speech addresses individuals, trumped only by the share of comments addressing parties/groups. Here, events and debates are seemingly less important drivers of hate speech. In stark contrast, up to 45 percent of hate speech below news articles addresses certain debates and another 43 to 44 percent of hate speech addresses either parties/groups or individuals. Interestingly, targets within user comments below news articles are rather clearly identifiable, with only two percent of comments falling into the residual “other” category. Vis-à-vis tweets and user comments on YouTube, this may, again, highlight the rather different approaches in comment moderation.

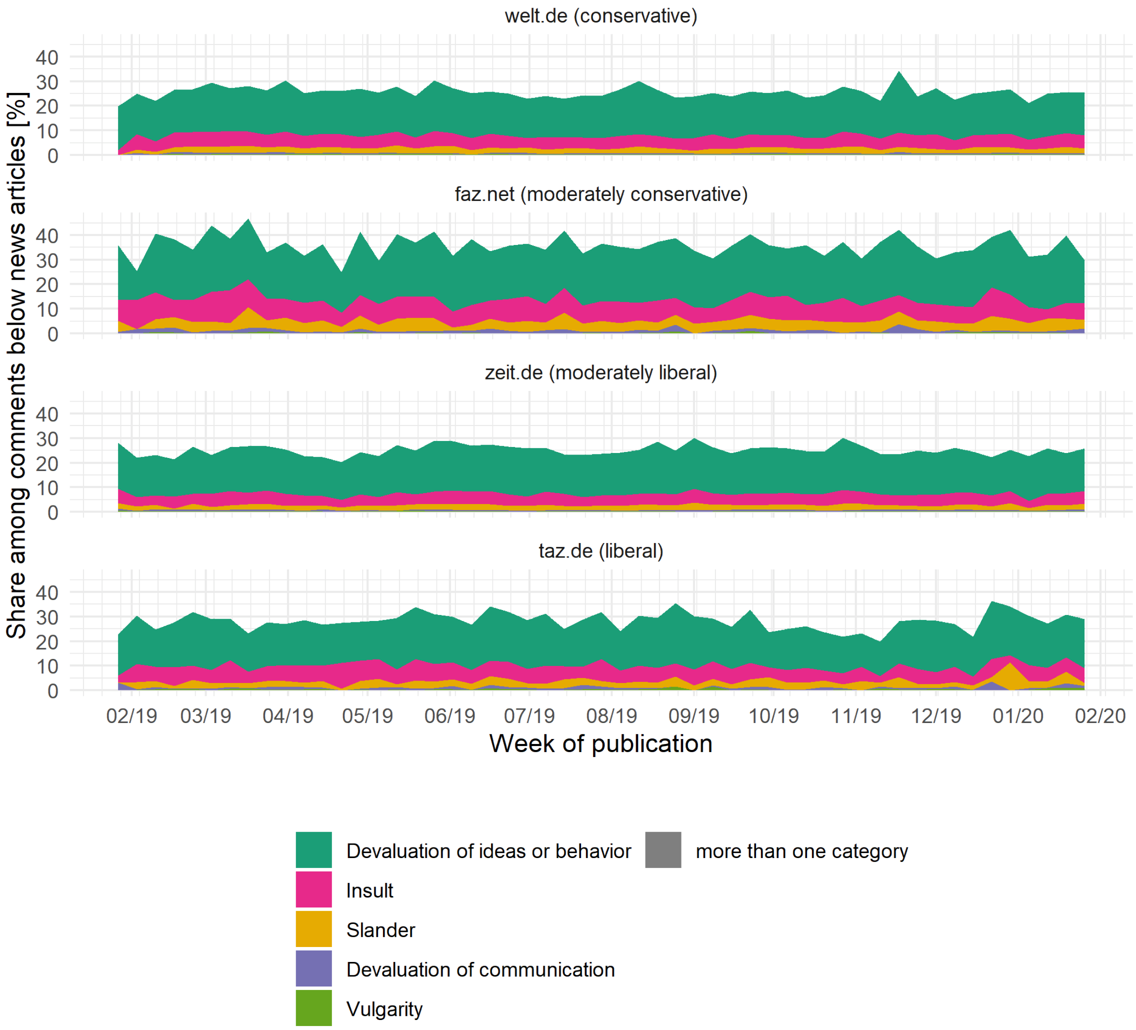

Figure 4. Share of hate speech online among comments below news articles over time.

Note. Based on all news articles' comments ($N = 455{,}003$).

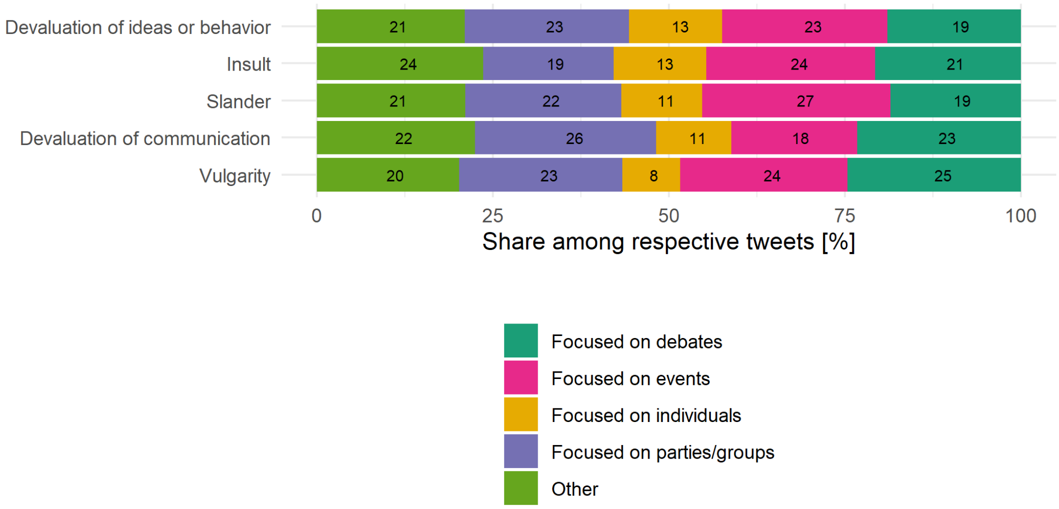

Figure 5. Share of targets among hate speech online in tweets.
Note. Based on tweets containing a respective category of hate speech online (e.g., devaluation of ideas or behavior is based on 13% of all *N* = 153,761 tweets; see Fig. 1).

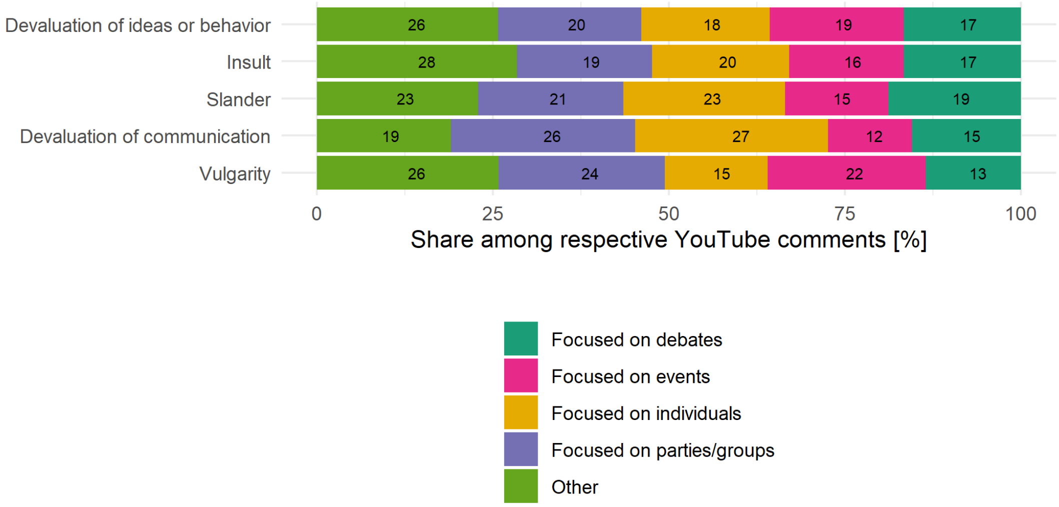

Figure 6. Share of targets among hate speech online in YouTube comments.
Note. Based on YouTube comments containing a respective category of hate speech online (e.g., insult is based on 4% of all *N* = 16,973 YouTube comments; see Fig. 2).

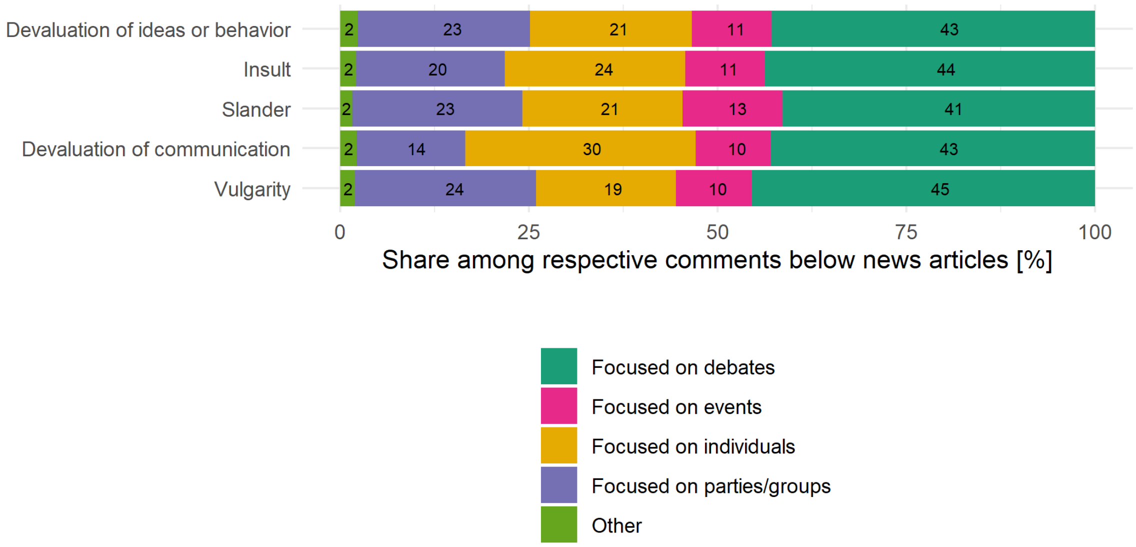

Figure 7. Share of targets among hate speech online in comments below news articles.

Note. Based on comments below news articles containing a respective category of hate speech online (e.g., slander is based on 1% of all *N* = 455,003 comments; see Fig. 3).

Discussion

Hate speech online follows different considerations and platform characteristics. Its definitions vary across stakeholders, are rather broad, and overlap only partially. On top of these inconsistencies, moderation techniques and goals also vary considerably. As such, it is little surprising that the prominence and targets of different forms of hate speech in the current study also vary across the three platforms Twitter, YouTube and comments below news articles. While the devaluation of ideas or behavior constitutes the most prominent type of hate speech in all three rather distinct datasets, the amount of other forms of visible hate speech (i.e., insult, slander, devaluation of communication, vulgarity) is comparably small and accumulates around certain events (Twitter) or single videos (YouTube). These latter findings are also roughly in line with prior research from other platforms, other cultural contexts, and other time periods (e.g., Coe et al., 2014; Ozalp et al., 2020; Williams & Burnap, 2016).

As opposed to news outlets, Twitter and YouTube can be expected to employ different techniques to reach their goals when moderating user comments. Consequently, differences in the visibility of hate speech online are to be expected—an expectation that is partly reflected in the prominence of different categories but also in the targets mentioned in the user comments. That is, while hate speech in tweets tends to focus on parties/groups and events, hate speech on YouTube comments primarily focuses on parties/groups and individuals. This is surprising also in light of these platforms' community guidelines which clearly highlight the focus on, in the case of YouTube, "individuals or groups" (Google, 2021).

User commentary on news outlets containing hate speech shows a strong focus on debates which, given the outlets' goal to promote civil public discourse, seems both consequential and problematic. On the one hand, this finding might very much depict the outlets' efforts to moderate their forums to stay on topic. In that, comments focusing on other (off-topic) aspects might have been removed by moderators or the outlets' communities have, over time, adhered to a principle of discussing and, also, devaluing ideas and behavior (i.e., the most prominent form of hate speech in the current study) on a given topic. On the other hand, the sheer amount of such a focus cannot be in the good interest of a civil discourse. A valuable viewpoint negotiation in the interest of a democracy's public discourse requires a collective communicative mode of mutual respect (Wessler, 2018). As such, hate speech online informs a less discursive public sphere and thus indirectly affects a broad set of targets turning to online discourse for orientation and an estimation of public opinion. On its most basic level, then, public discourse needs a mutual understanding that for these needs incivility and hate speech are destructive.

Whether the enforcement of such an understanding requires policy changes has to remain an open question, however. The current study provides empirical insights into three distinct datasets of public discourse which have spurred hate speech online on platforms adhering to two different kinds of moderation. In that, hate speech online not only has potentially contributed indirectly to a less discursive public sphere but also directly by mentioning or attacking targets in the comments. The current study has identified a need to focus even more on individual triggers spurring hate speech, such as interviews or prominent reports. Under the NetzDG, reports of hate speech are tied to individual comments; the current study, however, points out that it might make sense to understand accumulations of reports also as indications for trigger events which deserve further inquiry. Moreover, the current study highlights the urgent call to particularly provide direct targets with more support for dealing with hate speech online—a finding that has only recently also been echoed by survey studies (e.g., Chen et al., 2020; Der Spiegel, 2021).

The empirical grounds of these findings face several limitations, though. First, the three datasets are distinct and, to individual degrees, incomplete. They thus need to be treated with care and cannot be compared as-is. Second, both human coding and automated classification entail certain levels of uncertainty. Intercoder reliability for the nominal and imbalanced categories was sufficient using adequate coefficients yet less so when considering the commonly used Krippendorff's α which, however, is prone to penalizing for heavy imbalance. As for the classifiers, small numbers of positive cases distort interpretability of the commonly applied F1 score which becomes particularly apparent for the category of vulgarity. Therefore, also area-under-the-ROC-curve values have been reported. Moreover, the fact that vulgarity has almost never been classified in other studies echoes this circumstance and points out that the occurrences of hate speech online in the models used have been underestimated. Third, structural topic modeling (STM) as employed in this study to identify targets is one way, however certainly not the only one to identify targets. STM helps to identify topics across texts from co-occurrences of words and thus allows to cluster what—in the current study—can be considered rather homogenous texts per dataset. Qualitative inspection of all topics proved the approach applicable. That said, other approaches to identify targets might entail part-of-speech tagging and named entity recognition. Fourth and finally, the current study can only analyze commentary after moderation. Thus, the amounts of hate speech reported here depict what users can see but not what platforms seminally have to deal with. This differentiation, however, also seems crucial to respective policymaking—not only because platforms are presumably dealing with much larger shares of hate speech but also because directly targeted victims might have received much more hate speech in the first place, before moderation took action.

References

Anderson, A. A., & Huntington, H. E. (2017). Social media, science, and attack discourse: How Twitter discussions of climate change use sarcasm and incivility. *Science Communication*, *39*(5), 598–620. https://doi.org/10.1177/1075547017735113

Banks, J. (2010). Regulating hate speech online. *International Review of Law, Computers & Technology*, *24*(3), 233–239. https://doi.org/10.1080/13600869.2010.522323

Benoit, K., & Matsuo, A. (2020). *spacyr: Wrapper to the "spaCy" "NLP" Library*. https://CRAN.R-project.org/package=spacyr

Boberg, S., Schatto-Eckrodt, T., Frischlich, L., & Quandt, T. (2018). The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media and Communication*, *6*(4), 58–69. https://doi.org/10.17645/mac.v6i4.1493

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, *41*(3), 687–699. https://doi.org/10.1177/001316448104100307

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, *18*(3), 297–326. https://doi.org/10.1177/1468796817709846

Chen, G. M., Pain, P., Chen, V. Y., Mekelburg, M., Springer, N., & Troger, F. (2020). "You really have to have a thick skin": A cross-cultural perspective on how online harassment influences female journalists. *Journalism*, *21*(7), 877–895. https://doi.org/10.1177/1464884918768500

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. https://doi.org/10.1111/jcom.12104

Cohen-Almagor, R. (2011). Fighting hate and bigotry on the internet. *Policy & Internet*, *3*(3), 1–26. https://doi.org/10.2202/1944-2866.1059

Dahlberg, L. (2001). The internet and democratic discourse: Exploring the prospects of online deliberative forums extending the public sphere. *Information, Communication & Society*, *4*(4), 615–633. https://doi.org/10.1080/13691180110097030

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*, 512–515.

Der Spiegel. (2021, February 12). *Frauenfeindlichkeit im -Bundestag durch AfD gestiegen*. Der Spiegel. https://www.spiegel.de/politik/deutschland/bundestag-frauenfeindlichkeit-durch-afd-gestiegen-a-4c8c425c-6b08-4ac5-b049-61ad65d1240c

Dias Oliva, T. (2020). Content moderation technologies: Applying human rights standards to protect freedom of expression. *Human Rights Law Review*, *20*(4), 607–640. https://doi.org/10.1093/hrlr/ngaa032

Eckert, S., & Metzger-Riftkin, J. (2020). Doxxing, privacy and gendered harassment. The shock and normalization of veillance cultures. *Medien & Kommunikationswissenschaft*, *68*(3), 273–287. https://doi.org/10.5771/1615-634X-2020-3-273

European Court of Human Rights. (2020, September). *Hate speech*. https://www.echr.coe.int/Documents/FS_Hate_speech_ENG.pdf

Facebook. (2021). *Community standards*. https://www.facebook.com/communitystandards/violence_criminal_behavior

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619

Fretwurst, B. (2015). Reliabilität und Validität von Inhaltsanalysen. Mit Erläuterungen zur Berechnung des Reliabilitätskoeffizienten "Lotus" mit SPSS. In W. Wirth, K. Sommer, M. Wettstein, & J. Matthes (Eds.), *Qualitätskriterien in der Inhaltsanalyse* (pp. 176–203). Halem.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. Unesco Publishing.

Gelber, K. (2019). Differentiating hate speech: A systemic discrimination approach. *Critical Review of International Social and Political Philosophy*, 1–22. https://doi.org/10.1080/13698230.2019.1576006

George, C. (2015). Hate speech law and policy. In R. Mansell, P. H. Ang, C. Steinfield,

S. van der Graaf, P. Ballon, A. Kerr, J. D. Ivory, S. Braman, D. Kleine, & D. J. Grimshaw (Eds.), *The International Encyclopedia of Digital Communication and Society* (pp. 1–10). Wiley-Blackwell. https://doi.org/10.1002/9781118767771.wbiedcs139

Google. (2021). *Hate speech policy*. https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 1–15. https://doi.org/10.1177/2053951719897945

Günther, E., & Quandt, T. (2016). Word counts and topic models. Automated text analysis methods for digital journalism research. *Digital Journalism*, *4*(1), 75–88. https://doi.org/10.1080/21670811.2015.1093270

He, D. (2020). Governing hate content online: How the Rechtsstaat shaped the policy discourse on the NetzDG in Germany. *International Journal of Communication*, *14*, 3746–3768.

Heldt, A. (2019). Reading between the lines and the numbers: An analysis of the first NetzDG reports. *Internet Policy Review*, *8*(2), 1–18.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. https://doi.org/10.5281/zenodo.1212303

Kalsnes, B., & Ihlebæk, K. A. (2020). Hiding hate speech: Political moderation on Facebook. *Media, Culture & Society*. https://doi.org/10.1177/0163443720957562

Kasakowskij, T., Fürst, J., Fischer, J., & Fietkiewicz, K. J. (2020). Network enforcement as denunciation endorsement? A critical study on legal enforcement in social media. *Telematics and Informatics*, *46*, 101317. https://doi.org/10.1016/j.tele.2019.101317

Krippendorff, K. (2011). *Computing Krippendorff's Alpha-Reliability*. https://repository.upenn.edu/asc_papers/43

Ksiazek, T. B. (2015). Civil interactivity: How news organizations' commenting policies explain civility and hostility in user comments. *Journal of Broadcasting & Electronic Media*, *59*(4), 556–573. https://doi.org/10.1080/08838151.2015.1093487

Ksiazek, T. B., Peer, L., & Zivic, A. (2015). Discussing the news. Civility and hostility in user comments. *Digital Journalism*, *3*(6), 850–870. https://doi.org/10.1080/21670811.2014.972079

Loosen, W., Häring, M., Kurtanović, Z., Merten, L., Reimer, J., Roessel, L. van, & Maalej, W. (2017). Making sense of user comments: Identifying journalists' requirements for a comment analysis framework. *Studies in Communication and Media*, *6*(4), 333–364. https://doi.org/10.5771/2192-4007-2017-4-333

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, *22*(2), 205–224. https://doi.org/10.1177/1527476420982230

Newman, N., Fletcher, R., Schulz, A., Andı, S., & Nielsen, R. K. (2020). *Digital news report 2020*. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

Nielsen, L. B. (2002). Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech. *Journal of Social Issues*, *58*(2), 265–280. https://doi.org/10.1111/1540-4560.00260

Nithyanand, R., Schaffner, B., & Gill, P. (2017, August 14). Measuring offensive speech in online political discourse. *Proceedings of the 7th USENIX Workshop on Free and Open Communications on the Internet, FOCI 2017*. https://www.usenix.org/system/files/conference/foci17/foci17-paper-nithyanand.pdf

No Hate Speech Movement Deutschland. (2021, February 3). *What laws are there against hate speech?* https://no-hate-speech.de/en/knowledge/what-laws-are-there-against-hate-speech/

Obermaier, M., Hofbauer, M., & Reinemann, C. (2018). Journalists as targets of hate speech. How German journalists perceive the consequences for themselves and how they cope with it. *Studies in Communication and Media*, *7*(4), 499–524. https://doi.org/10.5771/2192-4007-2018-4-499

Ozalp, S., Williams, M. L., Burnap, P., Liu, H., & Mostafa, M. (2020). Antisemitism on Twitter: Collective efficacy and the role of community organisations in challenging online hate speech. *Social Media + Society*, *6*(2), 2056305120916850. https://doi.org/10.1177/2056305120916850

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–

283. https://doi.org/10.1177/1461444804041444

Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in Twitter. *Sensors*, *19*(21), 4654–4691. https://doi.org/10.3390/s19214654

Prochazka, F., Weber, P., & Schweiger, W. (2018). Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism Studies*, *19*(1), 62–78. https://doi.org/10.1080/1461670X.2016.1161497

Quarfoot, D., & Levine, R. A. (2016). How robust are multirater interrater reliability indices to changes in frequency distribution? *The American Statistician*, *70*(4), 373–384. https://doi.org/10.1080/00031305.2016.1141708

Rauh, C. (2018). Validating a sentiment dictionary for German political language—A workbench note. *Journal of Information Technology & Politics*, *15*(4), 319–343. https://doi.org/10.1080/19331681.2018.1485608

Reich, Z. (2011). User comments: The transformation of participatory space. In J. B. Singer, A. Hermida, D. Domingo, A. Heinonen, S. Paulussen, T. Quandt, Z. Reich, & M. Vujnovic (Eds.), *Participatory Journalism. Guarding open gates at online newspapers* (pp. 96–117). Wiley-Blackwell.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(2). https://doi.org/10.18637/jss.v091.i02

Rosenfeld, M. (2003). Hate speech in constitutional jurisprudence: A comparative analysis. *Cardozo Law Review*, *24*(4), 1523–1576.

Schabus, D., Skowron, M., & Trapp, M. (2017). One million posts: A data set of German online discussions. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '17*, 1241–1244. https://doi.org/10.1145/3077136.3080711

Schatto-Eckrodt, T., Boberg, S., Frischlich, L., & Quandt, T. (2020, May 24). *Hidden biases: The effects of deleted content on Twitter on sampling quality* [Paper presented at the 70th meeting of the International Communication Association]. https://player.vimeo.com/video/417213256

Siegel, A. A. (2020). Online hate speech. In N. Persily & J. A. Tucker (Eds.), *Social Media and Democracy: The State of the Field, Prospects for Reform* (pp. 56–88). University Press. https://doi.org/10.1017/9781108890960

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting incivility and impoliteness in online discussions. *Computational Communication Research*, *2*(1), 109–134. https://doi.org/10.5117/CCR2020.1.005.kath

Stromer-Galley, J. (2007). Measuring deliberation's content: A coding scheme. *Journal of Public Deliberation*, *3*(1), Art. 12.

Stroud, N. J., Scacco, J. M., Muddiman, A., & Curry, A. L. (2015). Changing deliberative norms on news organizations' Facebook sites. *Journal of Computer-Mediated Communication*, *20*(2), 188–203. https://doi.org/10.1111/jcc4.12104

Twitter, Inc. (2021). *Hateful conduct policy*. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Võ, M. L.-H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin Affective Word List Reloaded (BAWL-R). *Behavior Research Methods*, *41*, 534–538. https://doi.org/10.3758/BRM.41.2.534

von Eye, A. (2006). An alternative to Cohen's κ. *European Psychologist*, *11*(1), 12–24. https://doi.org/10.1027/1016-9040.11.1.12

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on Twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, *6*, 13825–13835. https://doi.org/10.1109/ACCESS.2018.2806394

Wessler, H. (2018). *Habermas and the Media*. Polity.

Williams, M. L., & Burnap, P. (2016). Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *The British Journal of Criminology*, *56*(2), 211–238. https://doi.org/10.1093/bjc/azv059

Wintterlin, F., Schatto-Eckrodt, T., Frischlich, L., Boberg, S., & Quandt, T. (2020). How to cope with dark participation: Moderation practices in German newsrooms. *Digital Journalism*, *8*(7), 904–924. https://doi.org/10.1080/21670811.2020.1797519

Zurth, P. (2020). The German NetzDG as role model or cautionary tale? Implications for the debate on social media liability. *Fordham Intellectual Property, Media & Entertainment Law Journal*. https://doi.org/10.2139/ssrn.3668804

Appendix

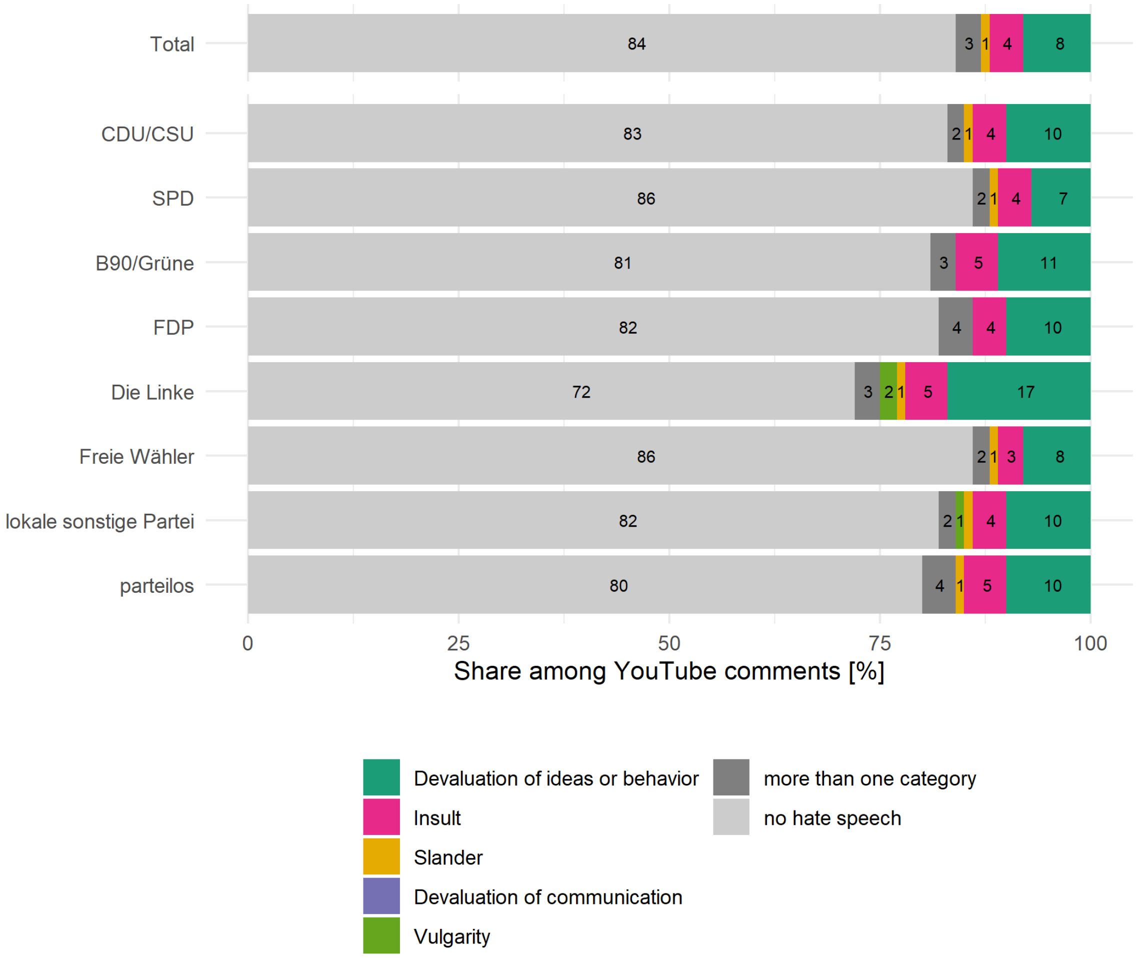

Figure A1. Share of hate speech online among YouTube comments per party.
Note. Based on all YouTube comments (*N* = 16,973).

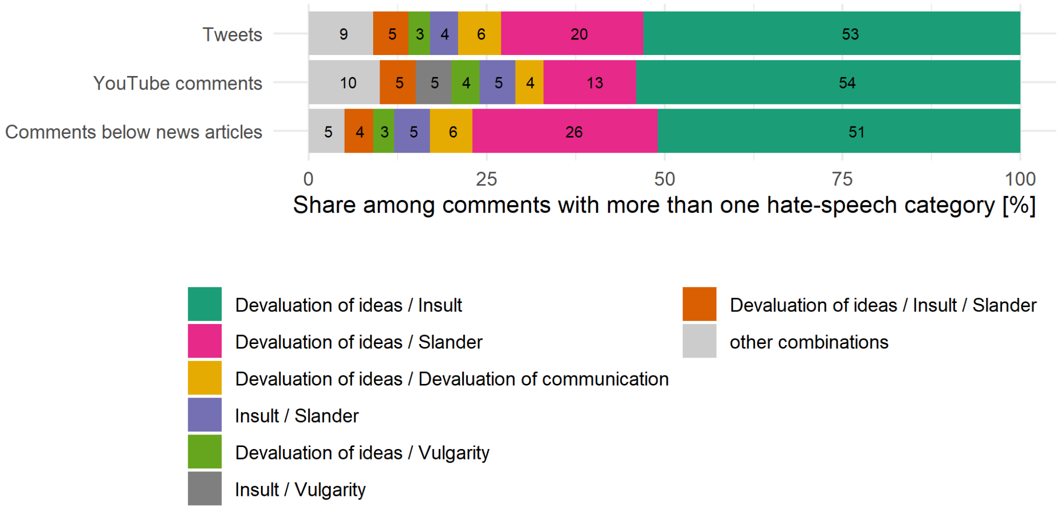

Figure A2. Share of Hate Speech Online among Comments in Multiple Categories. *Note*. Based on comments containing more than one category of hate speech (*N* = 16,329). This includes 3 percent of tweets (*n* = 4,553), 3 percent of YouTube comments (*n* = 428), and 2 percent of comments below news articles (*n* = 11,348).

Table A3. Automatically Classified Examples of Hate Speech Online.

Category **(not exclusive)**	**Data**	**Original Comment**	**Translation**
Devaluation of ideas/behavior	Twitter	@BjoernHoecke Ach Bernd... warum denkst du ständig, was du machst wäre von Relevanz. Du Fratz, du.	@BjoernHoecke Oh Bernd... why do you keep thinking what you do is relevant. You rascal, you.
	YouTube	Die Kinder hätten nicht so werden müssen, da haben alle Ämter über Jahrzehnte versagt !	The children did not have to become like this, all the offices have failed for decades
	News outlets	Noch besser!! Und jetzt? Meinen Sie damit irgendwen überzeugen zu können von Ihren Argumenten? Ich denke nein.	Even better!! And now? Do you think you can convince anyone of your arguments? I think no.
Insult	Twitter	Heute wählt Hamburg eine neue Bürgerschaft!Alle Stimmen für den guten Bürgermeister @TschenPe und die @spdhh !Geht wählen! Wählt (sozial)demokratisch!Sorgt dafür, dass die rechtsradikale AfD aus dem Parlament verschwindet - erst in Hamburg und dann überall in Deutschland!	Today Hamburg elects a new parliament! All votes for the good mayor @TschenPe and the @spdhh!Go vote! Vote (social) democratic!Make sure that the right-wing radical AfD disappears from parliament - first in Hamburg and then everywhere in Germany!
	YouTube	Verurteilung dieser Tat ohne jede Frage !!! Denken wir jedoch an die mindestens 2000 Vergewaltigten , Erstochenen, Erschlagen Menschen durch die Invasoren seit der illegalen	Condemnation of this act without question !!! However, let's think of the at least 2000 raped , stabbed , beaten to death people by the invaders since the illegal opening of

		Grenzöffnung durch Merkel ! Hat es da ein Ansprache gegeben, Alles nur Einzelfälle! Schuld trägt allein dieses politische System von Merkel und Co!	the border by Merkel ! Has there been a speech, All only individual cases! Blame carries alone this political system of Merkel and Co!
	News outlets	Ja. Jedes Jahr das ungenutzt verstrichen ist, macht sich in heute fälligen Rekord Kosten für die Abfederung der jährlich schlimmeren Folgen dieses Nichtstuns bemerkbar. Dazu kommen auch jetzt gleichzeitig die Kosten, die eine Umstellung auf klimaneutrales Leben und Wirtschaften verursachen werden, die unvernünftigerweise nicht über die vergangenen drei Jahrzehnte gestreckt investiert wurden. Das aus den laufenden Einnahmen machen zu wollen - Stichwort schwarze null - ist einfach nur möglich, wenn man zusätzliche Einnahmen von Bürgern kassiert. Angesichts von 20 % kinderarmut, von steigender Altersarmut, von steigenden Mieten und Energie- und mobilitarsskosten. Sowas nenne ich asozial.	Yes. Every year that has passed unused, makes itself felt in today due record costs for the cushioning of the annually worse consequences of this doing nothing. In addition, there are also now simultaneously the costs that will cause a conversion to climate-neutral life and economy, which have not been invested over the past three decades. Trying to do this out of current revenues - the keyword being black zero - is simply only possible if you collect additional revenues from citizens. In the face of 20% child poverty, rising old-age poverty, rising rents and energy and mobility costs. I call this asocial.
Devaluation of communication	Twitter	@Ralf_Stegner Sowas hat man mit 15 Jahren bei Facebook gepostet...Interessiert niemanden!	@Ralf_Stegner Such a thing was posted on Facebook when you were 15 years old...Nobody cares!
	YouTube	Einfach ne Glasfaserleitung von Gebäude A nach Gebäude B zu legen und den Aktenaustausch sowie die Kommunikation Digital über ein internes Netzwerk zu erledigen wäre wohl zu einfach gewesen. Wobei der Altersdurchschnitt in diesem Stadtrat der Optik nach wohl ziemlich hoch ist, da wird Netzwerktechnik und Digitalisierung noch Neuland sein.	Simply laying a fiber optic line from building A to building B and exchanging files and communication digitally via an internal network would have been too easy. The average age in this city council is probably quite high, so network technology and digitalization will still be new territory.
	News outlets	Hilfe, wir sind im Klimawahn! Jeden Tag sondert irgendjemand einen anderen Dünnpfiff ab. Ist es denn mal möglich in Gesamtzusammenhängen zu denken, statt unabgestimmtes und nicht umsetzbares Klein-Klein zu machen?	Help, we are in climate madness! Every day, someone spouts off a different load of tripe. Is it possible to think in overall contexts instead of making uncoordinated and not realizable small details?
Vulgarity	Twitter	@Ralf_Stegner Von der "Sozial-Demokratie zur Asozial-Demokratie" Linksradikal, Hass und Hetze, danke der SPD für die Spaltung der Gesellschaft....Ihr seid Asozial!!! Und ich wähle nicht die AfD!!!!	@Ralf_Stegner From "social democracy to asocial democracy" Left-wing radical, hate and agitation, thank the SPD for the division of society....You are asocial!!!! And I do not vote for the AfD!!!!
	YouTube	1880 halt, ohne telefon oder skype ... and shit, ohne tunnel müsste man quasi nochmal 10 beamte einstellen?	Just 1880, without phone or skype ... and shit, without tunnel you would have to hire another 10 civil servants?
	News outlets	Wir müssen vor allem weniger Dumme haben! Allein mir fehlt der Glaube.	Above all, we must have fewer stupid people! I alone lack faith.
Slander	Twitter	RT @Dani92K: @Ralf_Stegner Für mich ist die SPD eine linksradikale+demokratiefeindliche	RT @Dani92K: @Ralf_Stegner For me, the SPD is a left-wing radical+anti-democratic party that

		Partei die mit der Antifa Nazimethoden benutzt um p…	uses Nazi methods with the Antifa to p...
	YouTube	Merkel macht das nicht, sondern die Hintermänner ...und Merkel ist dafür geeignet die Schläge einzustecken.	Merkel doesn't do that, the backers do ...and Merkel is suited for taking the hits.
	News outlets	„Dinge als Fakten zu bezeichnen, die landläufig nicht als solche gehandelt werden. Also statt objektiv zu sein, ein Gefühl zum Faktum zu erklären.“ Wird dieser Ansatz nicht in Fächern wie Genderstudies schon längst praktiziert? Das neue ist wohl nur, dass nun auch Rechte auf diesen Zug aufspringen.	“Calling things facts that are not commonly traded as such. So instead of being objective, declaring a feeling to be fact.” Hasn't this approach been practiced in subjects like gender studies for a long time? The only new thing is that now also right-wingers are jumping on this bandwagon.

Note. Random samples automatically coded (also) as the given category.