Article

SM+S
social media + society

Social Media + Society
October-December 2021: 1–14
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20563051211052906
journals.sagepub.com/home/sms
SAGE

Assessing the Extent and Types of Hate Speech in Fringe Communities: A Case Study of Alt-Right Communities on 8chan, 4chan, and Reddit

Diana Rieger[1], Anna Sophie Kümpel[2], Maximilian Wich[3], Toni Kiening[1], and Georg Groh[3]

Abstract

Recent right-wing extremist terrorists were active in online fringe communities connected to the alt-right movement. Although these are commonly considered as distinctly hateful, racist, and misogynistic, the prevalence of hate speech in these communities has not been comprehensively investigated yet, particularly regarding more implicit and covert forms of hate. This study exploratively investigates the extent, nature, and clusters of different forms of hate speech in political fringe communities on *Reddit*, *4chan*, and *8chan*. To do so, a manual quantitative content analysis of user comments ($N = 6{,}000$) was combined with an automated topic modeling approach. The findings of the study not only show that hate is prevalent in all three communities (24% of comments contained explicit or implicit hate speech), but also provide insights into common types of hate speech expression, targets, and differences between the studied communities.

Keywords

hate speech, alt-right, fringe communities, Reddit, 4chan, 8chan, content analysis, topic modeling

On 15 March 2019, a right-wing extremist terrorist killed more than 50 people in mosques in Christchurch, New Zealand, and wounded numerous others—livestreaming his crimes on Facebook. Only 6 weeks later, on 27 April, another right-wing extremist attack occurred in a synagogue in Poway near San Diego, in which one person was killed and three more injured. The perpetrators were active in an online community within the imageboard *8chan*, which is considered as particularly hateful and rife with right-wing extremist, misanthropic, and White-supremacist ideas. Moreover, both the San Diego and Christchurch shooters used 8chan to post their manifestos, providing insights into their White nationalist hatred (Stewart, 2019). Following the attack in New Zealand, Internet service providers in Australia and New Zealand have temporarily blocked access to 8chan and the similar—albeit less extreme—imageboard *4chan* (Brodkin, 2019). After yet another shooting in El Paso was linked to activities on 8chan, the platform was removed[1] from the Clearnet entirely, with one of 8chan's network infrastructure providers claiming the unique lawlessness of the site that "has contributed to multiple horrific tragedies" as the main reason for this decision (Prince, 2019).

Whether the perpetrators' activities on 8chan and 4chan actually contributed to their radicalization or motivation can hardly be determined. However, especially the platforms' politics boards (8chan/pol/ and 4chan/pol/, respectively) have repeatedly been linked to the so-called alt-right movement, "exhibiting characteristics of xenophobia, social conservatism, racism, and, generally speaking, hate" (Hine et al., 2017, p. 92; see also Hawley, 2017; Tuters & Hagen, 2020). 4chan/pol/, in particular, has attracted the broader public's attention during Donald Trump's 2016 presidential campaign, often being the birthplace of conservative or even outright hateful and racist memes that circulated during the campaign. In addition to the mentioned communities on 4chan and 8chan, the controversial subreddit "The_Donald" is often referenced as a popular and more

[1]LMU Munich, Germany
[2]TU Dresden, Germany
[3]Technical University of Munich (TUM), Germany

Corresponding Author:
Diana Rieger, Department of Media and Communication, LMU Munich, Oettingenstrasse 67, 80538 Munich, Germany.
Email: diana.rieger@ifkw.lmu.de

"mainstreamy" outlet for alt-right ideas as well (e.g., Heikkilä, 2017).

Although these political fringe communities are considered as particularly hateful in the public debate, only few studies (Hine et al., 2017; Mittos, Zannettou, Blackburn, & De Cristofaro, 2019) have investigated these communities with regard to the extent of hate speech. Moreover, the mentioned studies are exclusively built on automated dictionary-based approaches focusing on explicit "hate terms," thus being unable to account for more subtle or covert forms of hate. To better understand the different types of hate speech in these communities, it also seems advisable to cluster comments in which hate speech occurs.

Addressing these research gaps, we (a) provide a systematic investigation of the extent and nature of hate speech in alt-right fringe communities, (b) examine both explicit and implicit forms of hate speech, and (c) merge manual coding of hate speech with automated approaches. By combining a manual quantitative content analysis of user comments (*N*=6,000) and unsupervised machine learning in the form of topic modeling, this study aims at understanding the extent and nature of different types of hate speech as well as the thematic clusters these occur in. We first investigate the extent and target groups of different forms of hate speech in the three mentioned alt-right fringe communities on Reddit (r/The_Donald), 4chan (4chan/pol/), and 8chan (8chan/pol/). Subsequently, by means of a topic modeling approach, the clusters in which hate speech occurs are analyzed in more detail.

Hate Speech in Online Environments

Hate speech was certainly not invented with the Internet. Being situated "in a complex nexus with freedom of expression, individual, group, and minority rights, as well as concepts of dignity, liberty, and equality" (Gagliardone, Gal, Alves, & Martínez, 2015, p. 10), it has been in the center of legislative discussion in many countries for many years. Hate speech is considered to be an elusive term, with extant definitions oscillating between strictly legal rationales and generic understandings that include almost all instances of incivility or expressions of anger (Gagliardone et al., 2015). For the context of this study, we deem both the content and the targets as crucial for conceptualizing hate speech. Accordingly, hate speech is defined here as the expression of "hatred or degrading attitudes toward a *collective*" (Hawdon, Oksanen, & Räsänen, 2017, p. 254), with people being devalued not based on individual traits, but on account of their race, ethnicity, religion, sexual orientation, or other group-defining characteristics (Hawdon et al., 2017, see also Kümpel & Rieger, 2019).

There are a number of factors—resulting from the overarching characteristics of online information environments—suggesting that hate speech is particularly problematic on the Internet. First, there is the problem of permanence (Gagliardone et al., 2015). Especially fringe communities are heavily centered on promoting users' freedom of expression, making it unlikely that hate speech will be removed by moderators or platform operators. But even if hateful content is removed, it might have already been circulated to other platforms, or it could be reposted to the same site again shortly after deletion (Jardine, 2019). Second, the shareability and ease of disseminating content in online environments further facilitates the visibility of hate speech (Kümpel & Rieger, 2019). During the 2016 Trump campaign, hateful anti-immigration and anti-establishment memes were often spread beyond the borders of fringe communities, surfacing to mainstream social media and influencing discussions on these platforms (Heikkilä, 2017). Third, the (actual or perceived) anonymity in online environments can encourage people to "be more outrageous, obnoxious, or hateful in what they say" (Brown, 2018, p. 298), because they feel disinhibited and less accountable for their actions. Moreover, anonymity can also change the relative salience of one's personal and social identity, thereby increasing conformity to perceived group norms (Reicher, Spears, & Postmes, 1995). Indeed, research has found that exposure to online comments with ethnic prejudices leads other users to post more prejudiced comments themselves (Hsueh, Yogeeswaran, & Malinen, 2015), suggesting that the communication behavior of others also influences one's own behavior. Fourth, and closely related to anonymity, there is the problem of the full or partial invisibility of other users (Brown, 2018; Lapidot-Lefler & Barak, 2012): The absence of facial expressions and other visibility originated interpersonal communication cues makes hate speech appear less hurtful or damaging in an online setting, thus increasing inhibitions to discriminate others. Last, one has to consider the community-building aspects that are particularly distinctive for online hate speech (Brown, 2018; McNamee, Peterson, & Peña, 2010). Not least in alt-right fringe communities, hate is often "meme-ified" and mixed with humor and domain-specific slang, creating a situation in which the use of hate speech can play a crucial role in strengthening bonds among members of the community and distinguishing one's group from clueless outsiders (Tuters & Hagen, 2020). Taken together, the mentioned factors facilitate not only the creation and use of hate speech in online environments, but also its wider dissemination and visibility.

Implicit Forms of Hate Speech

While many types of online hate speech are relatively straightforward and "in your face" (Borgeson & Valeri, 2004), hate can also be expressed in a more implicit or covert form (see Ben-David & Matamoros-Fernández, 2016 ; Benikova, Wojatzki, & Zesch, 2018; ElSherief, Kulkarni, Nguyen, Wang, & Belding, 2018; Magu & Luo, 2018; Matamoros-Fernández, 2017)—for example, by spreading negative stereotypes or strategically elevating one's ingroup.

Implicit hate speech shares characteristics with what Buyse (2014, p. 785) has labeled fear speech, which is "aimed at instilling (existential) fear of another group" by highlighting harmful actions the target group has allegedly engaged in or speculations about their goals to "take over and dominate in the future" (Saha, Mathew, Garimella, & Mukherjee, 2021, p. 1111). Indeed, one variety of implicit hate speech can be seen in the intentional spreading of "fake news," in which deliberate false statements or conspiracy theories about social groups are circulated to marginalize them (Hajok & Selg, 2018). This could be observed in connection with the European migrant crisis during which online disinformation often focused on the degradation of immigrants, for example, through associating them with crime and delinquency (Hajok & Selg, 2018, see also Humprecht, 2019).

Implicitness is a major problem for the automated detection of hate speech, as it "is invisible to automatic classifiers" (Benikova et al., 2018, p. 177). Using such implicit forms of hate speech is a common strategy to even avoid automatic detection systems and to cloak prejudices and resentments in "ordinary" statements (e.g., "My cleaning lady is really good, even though she is Turkish," see Meibauer, 2013). Thus, implicit hate speech points to the importance of acknowledging the wider context of hate speech instead of just focusing on the occurrence of single (and often ambiguous) hate terms.

Extent of Hate Speech

Considering the mentioned problems with the (automated) detection of hate speech, it is hard to determine the overall prevalence of hate speech in online environments. To account for individual experiences, extant studies have often relied on surveys to estimate hate speech exposure. Across different populations around the globe, such self-reported exposure to online hate speech ranges from about 28% (New Zealanders 18+, see Pacheco & Melhuish, 2018), to 64% (13- to 17-year-old US Americans, see Common Sense, 2018), and up to 85% (14- to 24-year-old Germans, see Landesanstalt für Medien NRW, 2018). In studies focusing both on younger and older online users (Landesanstalt für Medien NRW, 2018; Pacheco & Melhuish, 2018), exposure to online hate was more commonly reported by younger age groups, which might be explained by different usage patterns and/or perceptual differences. However, while these survey figures suggest that many online users seem to have been exposed to hateful comments, they tell us only little about the overall amount of hate speech in online environments. In fact, even a single highly visible hate comment could be responsible for survey participants responding affirmatively to questions about their exposure to online hate. Thus, to determine the actual extent of hate speech, content analyses are needed—although the results are equally hard to generalize. Indeed, the amount of content labeled as hate speech seems to differ considerably, depending on the studied platforms and (sub-)communities, the topic of discussions, or the lexical resources and dictionaries used to determine what qualifies as hate speech (ElSherief et al., 2018; Hine et al., 2017; Meza, 2016). Considering our focus on alt-right fringe communities, we will thus aim our attention at the presumed and actual hatefulness of these discussion spaces.

The "Alt-Right" Movement and Fringe Communities

What Is the Alt-Right?

The alt-right (=abbreviated form of alternative right) is a rather loosely connected and largely online-based political movement, whose ideology centers around ideas of White supremacy, anti-establishmentarianism, and anti-immigration (see Hawley, 2017; Heikkilä, 2017; Nagle, 2017). Gaining momentum during Donald Trump's 2016 presidential campaign, the alt-right "took an active role in cheerleading his candidacy and several of his controversial policy positions" (Forscher & Kteily, 2020, p. 90), particularly on the mentioned message boards on Reddit (r/The_Donald), 4chan, and 8chan (/pol/ on both platforms). Similar to other online communities, the alt-right uses a distinct verbal and visual language that is characterized by the use of memes, subcultural terms, and references to the wider web culture (Hawley, 2017; Tuters & Hagen, 2020; Wendling, 2018). Another common theme is "the cultivation of a position that sees white male identity as threatened" (Heikkilä, 2017, p. 4), which is connected both to strongly opposing policies related to "political correctness" (e.g., affirmative action) and to condemning social groups that are perceived to be profiting from these policies (Phillips & Yi, 2018). Openly expressing these ideas often culminates in the use of hate speech, particularly against people of color and women. However, while discussion spaces linked to the alt-right are routinely described as hateful, there is little published data on the quantitative amount of hate speech in these fringe communities.

Hate Speech in Alt-Right Fringe Communities

To our knowledge, empirical studies addressing the extent of hate speech in alt-right fringe communities have exclusively relied on automated dictionary-based approaches, estimating the amount of hate speech by identifying posts that contain hateful terms (Hine et al., 2017; Mittos et al., 2019). Focusing on 4chan/pol/, Hine and colleagues (2017) use the hatebase dictionary to assess the prevalence of hate speech in the "Politically Incorrect" board. They find that 12% of posts on 4chan/pol/ contain hateful terms, thus revealing a substantially higher share than the two examined "baseline" boards 4chan/sp/ (focusing on sports) with 6.3% and 4chan/int/ (focusing on international cultures/languages) with 7.3%. However, 4chan generally seems to be more hateful than other social media platforms: Analyzing a sample of Twitter posts for comparison, the authors find that only 2.2% of the analyzed tweets contained hateful terms. Looking at the most "popular" hate terms used in 4chan/pol/, it is also possible to draw cautious conclusions about the (main) target groups of hate speech. The hate terms appearing most—"nigger," "faggot," and "retard"—are indicative of racist, homophobic, and ableist sentiments and suggest that people of color, the lesbian, gay, bisexual, transgender and queer or questioning (LGBTQ) community, and people with disabilities might be recurrent victims of hate speech.

Utilizing a similar analytical approach, but exclusively focusing on discussions about genetic testing, Mittos and colleagues (2019) investigate both Reddit and 4chan/pol/ with regard to their levels of hate. For Reddit, their analysis shows that the most hateful subreddits alluding to the topic of genetic testing are associated with the alt-right (e.g., r/altright, r/TheDonald, r/DebateAltRight), with posts displaying "clear racist connotations, and of groups of users using genetic testing to push racist agendas" (Mittos et al., 2019, p. 9). These tendencies are even more amplified on 4chan/pol/ where discussion about genetic testing are routinely combined with content exhibiting racial and anti-Semitic hate speech. Reflecting the findings of Hine and colleagues (2017), racial and ethnic slurs are prevalent and illustrate the boards' close association with White-supremacist ideologies.

While these studies offer some valuable insights into the hatefulness of alt-right fringe communities, the dictionary-based approaches are unable to account for more veiled and implicit forms of hate speech. Moreover, although the most "popular" terms hint at the targets of hate speech, a systematic investigation of the addressed social groups is missing. Based on the literature review and theoretical considerations, our study thus sought to answer three overarching research questions:

> *Research Question 1*. What percentage of user comments in the three fringe communities contains explicit or implicit hate speech?
>
> *Research Question 2*. (a) In which way is hate speech expressed and (b) against which persons/groups is it directed?
>
> *Research Question 3*. What is the topical structure of the coded user comments?

Method

Our empirical analysis of alt-right fringe communities focuses on three discussion boards within the platforms Reddit (r/The_Donald), 4chan (4chan/pol/), and 8chan (8chan/pol/), thus spanning from central and highly used to more peripheral and less frequented communities. While Reddit, the self-proclaimed "front page of the Internet," routinely ranks among the 20 most popular websites worldwide, 4chan and 8chan have (or had) considerably less reach. However, due to their connection with the perpetrators of Christchurch, Poway, and El Paso, 4chan and 8chan are nevertheless of high relevance for this investigation. All three platforms follow a similar structure and are divided into a number of different subforums (called "subreddits" on Reddit and "boards" on 4chan/8chan). While Reddit requires users to register to post or comment, both 4chan and 8chan do not have a registration system, thus allowing everyone to contribute anonymously. The specific discussion boards—r/The_Donald, 4chan/pol/, and 8chan/pol/—were chosen due to their association with alt-right ideas as well as their relative centrality within the three platforms. Moreover, all three boards have previously been discussed as important outlets of right-wing extremists' online activities (Conway, Macnair, & Scrivens, 2019).

In the following sections, we will first describe the data collection process and then outline the two methodological/analytical approaches used in this study: (a) a manual quantitative content analysis of user comments in the three discussion boards and (b) an automated topic modeling approach. While 4chan and 8chan are indeed imageboards, (textual) comments play an important role on these platforms as well. On Reddit, pictures can easily be incorporated in the original post that constitutes the beginning of a thread, but comments are by default bound to text. Due to our two-pronged strategy, the nature of these communities, and to ensure comparability between the discussion boards, we focused our analyses on the textual content of comments and did not consider (audio-)visual materials such as images or videos. However, we refer to their importance in the context of hate speech in the discussion.

Data Collection

Since accessing and collecting content from the three discussion boards varies in complexity, we relied on different sampling strategies. Comments from r/The_Donald were obtained by querying the Pushshift Reddit data set (Baumgartner, Zannettou, Keegan, Squire, & Blackburn, 2020) via *redditsearch.io*. Between 21 April and 27 April 2019, we downloaded a total of 70,000 comments, of which 66,617 could be kept in the data set after removing duplicates and deleted/removed comments. Comments from 4chan/pol/ were obtained by using the independent archive page *4plebs.org* and a web scraper. Between 14 April and 29 April 2019, a total of 16,000 comments were obtained, of which 15,407 remained after the cleaning process.[2] Finally, comments from 8chan/pol/ were obtained by directly scraping the platform: All comments in threads that were active on 24 April 2019 were downloaded, resulting in a data set of 63,504 comments for this community. For the manual quantitative content analysis, 2,000 comments were randomly sampled from the data set of each of the three communities, thus leading to a combined sample size of 6,000 comments.

Approach I: Manual Quantitative Content Analysis

As our first main category, we coded explicit hate speech in accordance with recurrent conceptualizations in the literature. Within this category, we defined insults (attacks to individuals/groups on the basis of their group-defining characteristics, e.g., Erjavec & Kovačič, 2012) as offensive, derogatory, or degrading expressions, including the use of ethnophaulisms (Kinney, 2008). Instead of coding insults in general, we distinguished between personal insults (i.e., attacks of a specific individual) and general insults (i.e., attacks of a collective), also coding the reference point of personal insults and the target of general insults. The specific reference points [(a) Ethnicity, (b) Religion, (c) Country of Origin, (d) Gender, (e) Gender Identity, (f) Sexual Orientation, (g) Disabilities, (h) Political Views/Attitudes] or targets [(a) Black People, (b) Muslims, (c) Jews, (d) LGBTQ, (e) Migrants, (f) People with Disabilities, (g) Social Elites/Media, (h) Political Opponents, (i) Latin Americans*, (j) Women, (k) Criminals*, (l) Asians) were compiled on the basis of research on frequently marginalized groups (Burn, Kadlec, & Rexer, 2005; Mondal, Silva, Correa, & Benevenuto, 2018), and inductively extended (targets marked with *) during the coding process. Furthermore, we have coded violence threats as a form of explicit hate speech (Erjavec & Kovačič, 2012; Gagliardone et al., 2015), including both concrete threats of physical, psychological, or other types of violence and calls for violence to be inflicted on specific individuals or groups.

As our second main category, we coded implicit hate speech. To distinguish different subcategories of this type of hate speech, we relied more strongly on an explorative approach by focusing on communication forms that have been described in the literature as devices to cloak hate (see section "Implicit Forms of Hate Speech"). The first subcategory of implicit hate speech is labeled negative stereotyping and was coded when users expressed overly generalized and simplified beliefs about (negative) characteristics or behaviors of different target groups. The second subcategory—disinformation/conspiracy theories—reflects both "simple" disinformation and false statements about target groups and "advanced" conspiracy theories that represent target groups as maliciously working together toward greater ideological, political, or financial power (e.g., "the Jew media controls everything"). A third subcategory was labeled ingroup elevation and was coded when statements elevated or accentuated belonging to a certain (racial, demographic, etc.) group, oftentimes implicitly excluding and devaluing other groups. The last subcategory of implicit hate speech was labeled inhuman ideology. Here, it was coded whether a user comment supported or glorified hateful ideologies such as National Socialism or White supremacy, including the worshiping of prominent representatives of such ideologies.

In addition, a category spam was added to exclude comments containing irrelevant content such as random character combinations or advertisements. The entire coding scheme as well as an overview of the main content categories described in the previous paragraphs can be accessed via an open science framework (OSF) repository[3].

The manual quantitative content analysis was conducted by two independent coders. Both coders coded the same subsample of 10% from the full sample of comments to calculate inter-rater reliability with the help of the R package "tidycomm" (Unkel, 2021). Using both percent agreement and Brennan and Prediger's Kappa, all reliability values were satisfactory ($\kappa \geqslant 0.83$, see also Table 1). Prior to the analyses, all comments coded as spam were removed, leading to a final sample size of 5,981 comments.

Approach II: Topic Modeling

Topic modeling is an unsupervised machine learning approach to identify topics within a collection of documents and to classify these documents into distinct topics. Günther and Domahidi (2017) generally describe a topic as "what is being talked/written about" (p. 3057). Each topic would thus be represented in a cluster. Consequently, each cluster is assigned a set of words that are representative of the comments within the cluster. For our analysis, we first generated a topic model (TM_1) for all 5,981 comments to gain an understanding of the topics within the entire data set. Combined with the manual coding, these results provide insights on which topics are more hateful than others. Second, another topic model (TM_2) was created only for the comments identified as hateful (n=1,438) to examine the clusters of the comments in which hate speech occurs. To do so, TM_1 and TM_2 were compared by investigating the transitions between the models. In addition, TM_2 was also combined with the manually coded data, allowing to establish a connection between the cluster, type, and targets of hate speech.

CluWords was selected as the topic model algorithm—a state-of-the-art short-text topic modeling technique (Viegas et al., 2019). The reason for not choosing a more conventional technique such as Latent Dirichlet Allocation (LDA) is that these do not perform well on shorter texts because they rely on word co-occurrences (Campbell, Hindle, & Stroulia, 2003; Cheng, Yan, Lan, & Guo, 2014; Quan, Kit, Ge, & Pan, 2015). CluWords overcomes this issue by combining non-probabilistic matrix factorization and pre-trained word-embeddings (Viegas et al., 2019). Especially the latter allows enriching the comments with "syntactic and semantic information" (Viegas et al., 2019, p. 754). For this article, the fastText word vectors pre-trained on the English Common Crawl dataset were used because it is trained on web data and thus an appropriate basis (Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2019).

Table 1. Inter-Rater Reliability for Coded Categories.

Category	Percentage agreement	Brennan and Prediger's kappa
Source of the comment	1	1
Spam	0.99	0.99
Personal insult	0.93	0.87
Target of the personal insult	0.92	0.9
Reference point of the personal insult	0.92	0.91
Second target of the personal insults[a]	0.99	0.98
Second reference point of the personal insult[a]	0.99	0.99
General insult to a group of people	0.92	0.84
Group reference of the insults	0.91	0.91
Second group reference of the insults[a]	0.98	0.98
Violence threats	0.96	0.94
Target of the violence threat(s)	0.94	0.94
Negative stereotyping	0.92	0.91
Second negative stereotyping[a]	0.97	0.97
Disinformation/Conspiracy theories	0.87	0.83
Reference point of the disinformation/conspiracy theory	0.87	0.86
Ingroup elevation	0.93	0.85
Inhumane ideology	0.96	0.94

Note. $N=590$, two coders, all categories were nominal.
[a] If present, more than one target (or group of targets) could be coded.

One challenge of topic modeling is to find a meaningful number of clusters. Since topic modeling is an unsupervised learning approach, there is no single right solution. To cope with this problem, the following five criteria have been used to determine an appropriate number of clusters: (a) the same number of topics for TM_1 and TM_2, (b) a meaningful and manageable number of topics, (c) comprehensibility of the topics, (d) standard deviation of the topics' sizes, and (e) (normalized) pointwise mutual information.

Results

Results of Manual Quantitative Content Analysis

Addressing RQ1 (extent of explicit/implicit hate speech), we found that almost a quarter (24%, $n=1{,}438$) of the analyzed 5,981 comments contained at least one instance of explicit or implicit hate speech (see Table 2). In 821 of the comments (13.7%), forms of explicit hate speech were identified (i.e., at least one of the categories personal insult, general insult, or violence threat was coded). Implicit hate speech (i.e., negative stereotyping, disinformation/conspiracy theories, ingroup elevation, and inhuman ideologies) occurred slightly more often and was observed in 928 comments (15.5%).

Table 2. Number of Comments Containing Hate Speech.

Comments contained . . .	Absolute	Relative[a] (%)
. . . no hate speech	4,543	76.0
. . . hate speech of at least one type[b]	1,438	24.0
. . . explicit hate speech	821	13.7
. . . implicit hate speech	928	15.5

[a] $n=5{,}981$.
[b] Due to the fact that explicit and implicit hate speech can occur in the same comment, numbers of explicit and implicit hate speech do not add up to the overall numbers.

Focusing on RQ2a (forms of hate speech), general insults were the most common form of hate speech and observed in 570 comments: they were included in almost every 10th comment of the entire sample (9.5%) and in more than one-third of all identified hateful comments (39.6%). Disinformation and conspiracy theories followed next and made up 31.8% of all comments with hate speech ($n=458$). Within this category, conspiracy theories ($n=294$) were observed almost twice as often as mere disinformation ($n=164$). In over a quarter of all hateful comments (25.7%), inhuman ideologies were referenced or expressed ($n=369$), with 10.8% relating to National Socialism and 14.9% to White-supremacist ideologies. Violence threats were observed in 221 comments (3.7% total; 15.4% of hateful comments), negative stereotyping in 192 comments (3.2% total, 13.4% of hateful comments), and ingroup elevation was coded for 303 comments (5.1% total, 21.1% of hateful comments), Within our sample, personal insults emerged as the least common form of hate speech ($n=139$), making up only 2.3% of all comments and 9.7% of all hateful comments.

Nevertheless, to answer RQ2b (reference points/targets of hate speech), we analyzed the reference points of these personal insults in more detail. Most personal insults attacked an individual's sexual orientation (32.1%), their ethnicity (27%), their political attitude (10.9%), or referred to an actual or alleged disability (10.2%). Personal insults referring to one's religion, country of origin, gender, or gender identity could only rarely be observed. For the categories general insults, violence threat, negative stereotyping, and disinformation/conspiracy theories, we further analyzed which groups were targeted with hateful sentiments (see Table 3). Jews were by far the most affected group and targets of explicit or implicit hate speech in 478 comments. When Jews were targeted, this happened most often in the context of disinformation/conspiracy theories and general insults. Black people were the second most targeted group in the sample (targeted in 277 comments), with attacks occurring primarily in the context of general insults. Other frequent targets were political opponents (targeted in 238 comments), Muslims (targeted in 148 comments), and the LGBTQ community (targeted in 127 comments).

To identify differences between the three fringe communities, we also conducted the analyses separately for r/

Table 3. Targets of Hate Speech Across Different Types of Hate Speech.

Group	General insult	Violence threat	Negative stereotyping	Disinformation/ conspiracies	Total
Black people	197	19	39	22	277
Muslims	42	26	34	46	148
Jews	182	41	44	211	478
LGBTQ	99	10	7	11	127
Migrants	7	5	3	4	19
People with disabilities	—	—	—	—	—
Social elites/media	8	3	4	35	50
Political opponents	38	38	52	110	238
Latin Americans	19	2	7	4	32
Women	21	10	18	6	55
Criminals	—	6	—	—	6
Asians	9	—	3	1	13
Rest/undefined	13	58	6	8	85
Total	635	218	217	458	1,528

LBGTQ: lesbian, gay, bisexual, transgender and queer or questioning.

Table 4. Amount of Hate Speech on the Studied Communities across Different Types of Hate Speech.

Extent of	r/The_Donald/		4chan/pol/		8chan/pol/	
	n = 1,998		n = 1,992		n = 1,991	
	Absolute	Relative (%)	Absolute	Relative (%)	Absolute	Relative (%)
Hate speech total	275	**13.8**	478	**24.0**	685	**34.4**
Explicit hate speech	99	**5.0**	329	**16.5**	393	**19.7**
Personal insult	11	0.6	71	3.6	57	2.9
General insult	40	2.0	238	11.9	292	14.7
Violence threat	52	2.6	67	3.4	102	5.1
Implicit hate speech	207	**10.4**	247	**12.4**	474	**23.8**
Negative stereotyping	68	3.4	50	2.5	74	3.7
Disinformation/Conspiracy theory	114	5.7	125	6.3	219	11.0
Ingroup elevation	98	4.9	74	3.7	131	6.6
Inhumane ideology	12	0.6	98	4.9	259	13.0

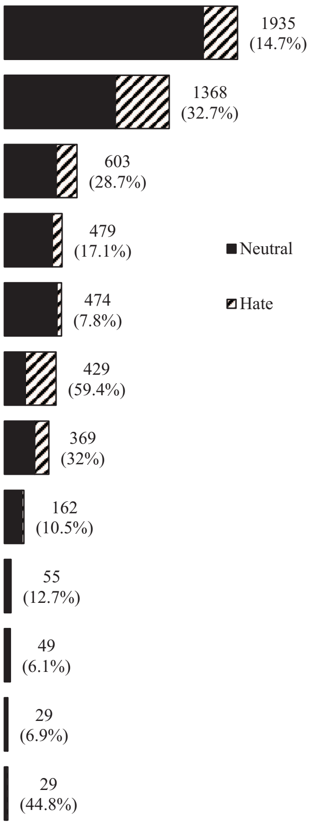
The_Donald, 4chan/pol/, and 8chan/pol/. Moving from the more "mainstreamy" r/The_Donald to the outermost 8chan/pol/, the amount of hate speech increases steadily: While 13.8% of all analyzed comments on r/The_Donald included at least one form of hate speech, we identified 24% of comments on 4chan/pol/ and even 34.4% of comments on 8chan/pol/ as containing hate speech. As can be inferred from Table 4, the amount of explicit and implicit hate speech also differed between the three communities: Particularly striking here is the low amount of explicit hate speech on r/The_Donald, which is mainly due to the fact that general insults are much less common than on 4chan/pol/ and 8chan/pol/. Looking more closely at implicit hate speech, we see that 8chan/pol/ emerged as the community with the highest share of such indirect, more veiled forms of hate speech, resulting mainly from the relatively high amount of comments featuring disinformation/conspiracy theories and inhuman ideologies.

Results of Topic Modeling

To answer RQ3 (topical structure of the coded comments), two topic models (TM_1 and TM_2) were generated and combined with the results of the manual quantitative content analysis. TM_1 focuses of the entire data set, while TM_2 is restricted to the comments that were identified as containing hate speech. Table 5 shows the topics of TM_1, their relative distribution between the sources, the absolute number of comments, and the proportion of hate speech. After the evaluation of different numbers of topics, 12 topics turned out to be most appropriate. Overall, the topics can be considered meaningful, and their content meets the expectations for these fringe communities (e.g., focus on political affairs, conspiracy theories, anti-Semitism)[4]. A2–A8 have a thematic focus, while A9, A11, and A12 bundle foreign-language comments. As A9–A12 are relatively small compared to the total number of comments (and

Table 5. Topics From TM_1 and Their Frequency Distribution.

Topics of TM_1	r/The_Donald (%)	4chan/pol (%)	8chan/pol (%)	Absolute (hate speech share)
(A1) Really actually think know something never want certainly obviously though	37.9	30.1	32.0	1935 (14.7%)
(A2) Shit fucking damn dipshit asshole faggot bitch motherfucker dumbass goddamn	27.6	40.4	32.0	1368 (32.7%)
(A3) Government political society ideology people democratic nation economic citizens morality	47.3	28.4	24.4	603 (28.7%)
(A4) John Robert David James Michael Chris Richard Ryan Todd George	45.5	29.0	25.5	479 (17.1%)
(A5) Foods protein nutrient fats diet hormone cholesterol meat vitamins veggies	21.5	40.9	37.6	474 (7.8%)
(A6) Jews Muslims Zionists Arabs Judaism Christians Gentiles Kikes Semitic Goyim	22.4	30.8	46.9	429 (59.4%)
(A7) Poland Germany Europe France British Finland Sweden Russia Italy American	26.0	38.2	35.8	369 (32%)
(A8) Wikileaks FBI CIA FOIA Intel Mossad NWO files gov leaks	39.5	25.3	35.2	162 (10.5%)
(A9) ett drar och handlar speciellt samtliga framtida liknande tror sluta	20.0	29.1	50.9	55 (12.7%)
(A10) xt torrent urn magnet tn ut hd ui ii aws	26.5	20.4	53.1	49 (6.1%)
(A11) erfolg muessen vorausgesetzt betroffenen natuerlich dortigen verbreiten einzigen wahres skeptisch	0	10.3	89.7	29 (6.9%)
(A12) een voor wordt uit het niet gaat zijn krijg terugkeer	10.3	31.0	58.6	29 (44.8%)



consequently less meaningful), they will be excluded from the following analyses.

In general, each topic is equally distributed across the three sources with some noticeable exceptions: 47.3% and 45.5% of the comments from the political topics A3 and A4 originate from r/The_Donald. Topic A2—consisting exclusively of swear words—can mostly be allocated to 4chan/pol (40.4%) and 8chan/pol (32.0%), which is in line with the results from the manual content analysis. The topic with a focus on anti-Semitism and Islam (A6) also exhibits an unequal distribution: r/The_Donald/'s share is only 22.4%, while 4chan/pol's share is 30.8% and 8chan/pol's is 46.9%. In light of the observed hatefulness of 4chan/pol and 8chan/pol, it is remarkable that both are the main origin of the identified topic focusing on nutrition (A5), which might be explained by their broader scope. Focusing on the occurrence of hate speech, the topics A2 (32.7%), A6 (59.4%), and A7 (32.0%) have to be highlighted due to their higher-than-average share of hate. This is not surprising, as the keywords from A2 only contain swear words, A6 covers (anti-)Semitic and Islamic comments, and A7 refers to foreign countries which are often the target of hate due to the alt-rights' nationalist orientation.

To better understand the clusters/topics in which hate speech occurs, a second topic model (TM_2) was generated based on the 1,438 hateful comments only (see Tables 6 and 7). Both models show a similar topical structure and some topics from TM_1 are reflected in TM_2 as well: A1 is similar to H3 (generic topic), A2 to H1 (swear words), A6 to H2 (largely anti-Semitic), and A7 to H4 (foreign affairs). On the contrary, other topics emerged as more fine-grained when only considering hate speech–related comments (TM_2). A good example is topic A3, which focuses on the government, politics, and society. Hateful comments from this topic can be found, among others, in the topics about US democrats and republicans (H5), political ideology (H9), and finances and taxes (H10).

Tables 6 and 7 depict the topics of TM_2 in combination with the manual analysis to get a deeper understanding of thematic clusters in which the different types of hate speech occur: The first one distinguishes between the different forms of explicit and implicit hate speech, the second one between the different targets of hate speech. Concerning the forms of hate speech, the comments from the topic with swear words (H1) tend to be explicit hate speech, particularly general insults (238 out of 398). In contrast to that, all other topics contain more implicit hate speech—a difference that should not be surprising due to the nature of the topics. What is interesting is the difference between the two (anti-)

Table 6. Topics of TM_2 Combined With Forms of Hate Speech From Manual Coding.

Topics of TM_2	# Comments	Explicit			Implicit			
		Personal insult	General insult	Violence threat	Negative stereotyping	Disinformation/ Conspiracy theory	Inhumane ideology	Ingroup elevation
(H1) Shit fucking bitch asshole motherfucker faggot dipshit dumbass damn dick	396	88	238	54	38	58	50	53
(H2) Jews Goyim Kikes Semitic Ashkenazic Gentiles Zionists Arabs Yids Africans	265	12	138	37	39	110	87	45
(H3) Really actually think want know something going never obviously certainly	250	17	77	40	40	85	65	68
(H4) Poland Germany Europe Russian France British Austria Berlin Soviet American	96	4	29	13	6	27	51	8
(H5) Democrats republicans voters conservatives liberals people electorate government citizens socialists	86	1	21	10	18	40	19	51
(H6) David NWO Donald Hilary FBI CIA NBC James Clinton Kennedy	86	8	16	9	7	35	28	12
(H7) Islam Muslims Allah Quran Koran Christians Mohammedan Infidels Sunni Jihad	80	2	19	15	25	39	12	17
(H8) Murderers terror killing enemy innocents crimes violence deadly civilians horrific	79	2	17	30	11	25	24	22
(H9) Ideology worldview morality political societal dialectics nationalism liberalism dogma religion	43	1	6	3	7	20	18	15
(H10) Tax pay costs government price tariffs economic amount money considerable	41	4	7	8	2	13	8	9
(H11) muessen erfolg verbreiten natuerlich anfang wahres wieso vorausgesetzt irgend gebrauchen	11	0	0	1	0	4	6	3
(H12) een uit gaat voor wordt niet het krijg zijn deze	5	0	2	1	0	2	1	0
Total	1,438	139	570	221	193	458	369	303

Table 7. Topics of TM_2 Combined With Targets of Hate Speech From Manual Coding.

Topics of TM_2	# Comments	Black people	Muslims	Jews	LGBTQ	Migrants	People with disabilities	Elites/ Media	Political opponents	Latin Americans	Women	Criminals	Asians	Rest
(H1) Shit fucking bitch asshole motherfucker faggot dipshit dumbass damn dick	396	104	16	68	87	9	—	11	49	6	27	1	4	28
(H2) Jews Goyim Kikes Semitic Ashkenazic Gentiles Zionists Arabs Yids Africans	265	60	22	230	3	1	—	4	18	8	5	—	2	4
(H3) Really actually think want know something going never obviously certainly	250	48	21	71	19	2	—	12	37	11	13	2	3	19
(H4) Poland Germany Europe Russian France British Austria Berlin Soviet American	96	18	10	22	3	3	—	1	11	3	—	—	—	6
(H5) Democrats republicans voters conservatives liberals people electorate government citizens socialists	86	13	2	7	—	1	—	8	56	2	3	—	—	1
(H6) David NWO Donald Hilary FBI CIA NBC James Clinton Kennedy	86	6	2	25	3	1	—	7	24	—	—	—	1	1
(H7) Islam Muslims Allah Quran Koran Christians Mohammedan Infidels Sunni Jihad	80	7	61	10	6	—	—	2	10	—	—	—	—	6
(H8) Murderers terror killing enemy innocents crimes violence deadly civilians horrific	79	14	6	14	5	—	—	2	20	2	3	2	3	14
(H9) Ideology worldview morality political societal dialectics nationalism liberalism dogma religion	43	2	4	11	1	—	—	3	9	—	4	—	—	2
(H10) Tax pay costs government price tariffs economic amount money considerable	41	3	3	15	—	2	—	—	2	—	—	1	—	4
(H11) muessen erfolg verbreiten natuerlich anfang wahres wieso vorausgesetzt irgend gebrauchen	11	—	1	3	—	—	—	—	1	—	—	—	—	—
(H12) een uit gaat voor wordt niet het krijg zijn deze	5	2	—	2	—	—	—	—	1	—	—	—	—	—
Total	1,438	279	148	480	127	19	0	50	239	32	55	6	13	85

LBGTQ: lesbian, gay, bisexual, transgender and queer or questioning.

religious topics H2 ([anti-]Semitism) and H7 ([anti-]Islam). While the first one contains many explicit general insults (138 out 265), the second one has a stronger focus on implicit hate speech, in particular on disinformation (39 out of 80) and negative stereotyping (25 out of 80). Beyond that, H4 and H5 have to be mentioned. H4, the topic about foreign affairs, has its maximum in the category inhuman ideologies (51 out of 96). The topic about US democrats and republicans (H5) exhibits a relatively large number of ingroup elevation (51 out of 86) and disinformation (40 out of 86).

Concerning the targets of hate speech, the automatically generated topics are in line with the manual coding, as shown in Table 7. The (anti-)Semitic and Islamic topic have their maximum in the respective target groups (230 out of 265; 61 out of 80). H4, the topic about US democrats and republicans, mainly contains comments targeting political opponents (56 out of 86). The two more generic topics (H1) and (H3) target a wider range of groups and their distribution is in line with the overall distribution of all topics.

Discussion

Building on ongoing public debates about alt-right fringe communities—that have been described as "the home of some of the most vitriolic content on the Internet" (Stewart, 2019)—this study investigates whether these public perceptions withstand empirical scrutiny. Focusing on three central alt-right fringe communities on Reddit (r/The_Donald), 4chan (4chan/pol/), and 8chan (8chan/pol/), we provide a systematic investigation of the extent and nature of both explicit and implicit hate speech in these communities. To do so, we combine a manual quantitative content analysis of user comments (N=6,000) with an automated topic modeling approach that offers additional insights into the clusters in which hate speech occurs.

The most obvious finding to emerge from our analysis is that hate speech is prevalent in all three studied communities: In almost a quarter of the sample (24%), at least one instance of explicit or implicit hate speech could be observed. Reflecting results from an automated dictionary-based approach by Hine and colleagues (2017)—who identified 12% of comments on 4chan/pol/ to contain (explicitly) hateful terms—we found that 13.7% of all analyzed comments featured explicit hate speech. However, our manual quantitative content analysis allowed us to also examine the extent of more veiled, indirect forms of hate speech, which was found in 15.5% of all comments. Differences between platforms are in line with the expectations one might have when moving from the more moderate to the more extreme communities: Comparatively, r/The_Donald featured the lowest amount of hate speech, followed by 4chan/pol/, and 8chan/pol/, suggesting that the "fringier" communities are distinctly more hateful.

Looking more closely at hate speech expression and common targets of hate speech, the results show that general insults of groups, referencing, or spreading disinformation/conspiracy theories, as well as the expression or glorification of inhuman ideologies such as National Socialism or White supremacy occurred most frequently. The reason for the high incidence of general insults might partly result from including ethnophaulisms and other derogatory terms such as "newfag" and "oldfag" that are regularly used on 4chan and 8chan to refer to new versus experienced users. The observed prevalence of disinformation and conspiracy theories might thus be even more alarming than the use of "plain" insults.

With regard to the social groups affected by hate speech in alt-right fringe communities, our analysis shows that Jews were targeted most often, followed by Black people and political opponents. While Jews were similarly observed as being targets of general insults, they were most often referenced in the context of disinformation and conspiracy theories, which chimes in with the observed extent of National socialist and White-supremacist ideologies in the studied communities. Political opponents are most often referenced within disinformation and conspiracy theories as well, thus reflecting the communities' close connection to populist attitudes that are associated with the demonization of institutions and political others (see Fawzi, 2019).

The topic models generated on the basis of the sampled user comments are in line with the results of the manual quantitative content analysis and provide additional insights into discussion topics that are likely to feature hate speech. They reflect the extent of (group-related) insults, anti-Semitic and anti-Islamic sentiments, and the strong nationalist orientation of the studied communities. Furthermore, the analysis shows that hate speech—although this might come as no surprise considering our focus on political fringe communities—often occurs in discussions about the government, the (US) political system, religious and political ideologies, or foreign affairs. Subsequent (computational) analyses could take these insights as a starting point to use specific contexts (=topics) for hate speech detection and artificial intelligence (AI) training sets.

Taking a look into potential directions for future studies, hate and antidemocratic content is not only conveyed through text: In an analysis of German hate memes, Schmitt and colleagues (2020) found that memes often display symbols, persons, or slogans known from National Socialism and the Nazi regime. Relatedly, Askanius (2021) traced an adaptation of stylistic strategies and visual aesthetics of the alt-right in the online communication of a Swedish militant neo-Nazi organization. Considering "that the visual form is increasingly used for strategically masking bigoted and problematic arguments and messages" (Lobinger, Krämer, Venema, & Benecchi, 2020, p. 347), and that images and videos tend to develop more virality than mere text (Ling et al., 2021), future studies should focus more strongly on such visual hate speech, which would also more adequately reflect the communication routines of the studied alt-right fringe communities.

Under the guise of "insider jokes," humor, or memes, it is possible that hate speech is not recognized as such or is perceived as less harmful. Oftentimes, it cannot be judged as unequivocally criminal and is thus not deleted by platforms. Content that—due to this "milder" perception—also finds favor in groups that do not in principle share the hostile ideas behind it is thus increasingly becoming the norm (Fang & Woodhouse, 2017). Accordingly, it can be assumed that the frequent confrontation with hate speech is loosening the boundaries of what can be said and thought, even among initially uninvolved Internet users. This mainstreaming process is described, for example, by Whitney Phillips (2015), who notes the historical transition of hateful, racist memes from fringe communities on the Internet to an increasingly broader public. Sanchez (2020), therefore, warns against a normalization of the "dark humor" that occurs in viral hate memes and calls for critical consideration and research of a possible desensitization to hate and incitement as a consequence. This study adds to this body of literature by providing first evidence that implicit hate speech is as prevalent as explicit hate speech and should thus be considered when analyzing both the extent as well as the potential harm of online hate. In addition, future studies should emphasize the long-term perspective and potential dangers of this development in which mainstreaming would contribute to hate becoming more and more "normal."

This work has limitations that warrant discussion. First, due to difficulties with the data collection, the initial number of comments on the analyzed communities varied, with 4chan/pol/ having a considerably smaller base of comments to sample from than r/The_Donald and 8chan/pol/. Moreover, all comments were scraped in April 2019, which might have influenced the results due to specific (political) topics being more or less obtrusive during that time period, possibly also influencing the general amount of hate speech. Second, it should be noted that we did not explicitly exclude hate terms that are part of typical communication norms within the studied communities. Terms such as the mentioned "newfag" were coded as hate speech although they may simply reflect 4chan jargon and are not used with malicious intentions. Nevertheless, we intentionally decided to code it as hate speech as even "normalized" or unintended hate speech can have negative effects (e.g., Burn et al., 2005). Third, our methodology and analysis were focused on textual hate speech, which is why we are unable to account for the amount of hate speech that is transmitted via shared pictures, (visual) memes, or videos. As we have outlined above, it is nevertheless an important endeavor to include the analysis of visual hate speech for which the results of our study might provide a fruitful starting point.

Notwithstanding its limitations, this study provides a first systematic investigation of the extent and nature of hate speech in alt-right fringe communities and shows how widespread verbal hate is on these discussion boards. Further research is needed to confirm and validate our findings, explore the effects of distinct forms of explicit and implicit hate speech on users, and assess the risks of virtual hate turning into real-life violence.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Diana Rieger https://orcid.org/0000-0002-2417-0480
Anna Sophie Kümpel https://orcid.org/0000-0001-7184-4057

Notes

1. In autumn 2019, 8chan was relaunched as *8kun*, which can be accessed from the Clearnet again. However, the original creator of 8chan, Fredrick Brennan, has not only publicly claimed to regret his creation but also vocally opposed the relaunch of 8chan (Roose, 2019).
2. Due to rate limits and technical hurdles, we were only able to scrape 1,000 comments per day from *4plebs.org*, which is why 4chan/pol/ has (a) overall the smallest initial data set and (b) the longest span of data collection.
3. https://osf.io/yfxzw/
4. Exceptions are A1 and A10. A1 is a generic topic containing comments that the algorithm could not assign to more meaningful classes. A10 is the result of comments containing links to file-sharing platforms.

References

Askanius, T. (2021). On frogs, monkeys, and execution memes: Exploring the humor-hate nexus at the intersection of neo-nazi and alt-right movements in Sweden. *Television & New Media*, *22*(2), 147–165.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit dataset. ArXiv:2001.08435 [Cs]. http://arxiv.org/abs/2001.08435

Ben-David, A., & Matamoros-Fernández, A. (2016). Hate Speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, *10*, 1167–1193.

Benikova, D., Wojatzki, M., & Zesch, T. (2018). What does this imply? Examining the impact of implicitness on the perception of hate speech. In G. Rehm & T. Declerck (Eds.), *Language Technologies for the Challenges of the Digital Age* (pp. 171–179). Springer. https://doi.org/10.1007/978-3-319-73706-5_14

Borgeson, K., & Valeri, R. (2004). Faces of hate. *Journal of Applied Sociology*, *21*(2), 99–111.

Brodkin, J. (2019, March 20). 4chan, 8chan blocked by Australian and NZ ISPs for hosting shooting video. *Ars Technica*. https://arstechnica.com/tech-policy/2019/03/australian-and-nz-isps-blocked-dozens-of-sites-that-host-nz-shooting-video/

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities*, *18*(3), 297–326.

Burn, S. M., Kadlec, K., & Rexer, R. (2005). Effects of subtle heterosexism on gays, lesbians, and bisexuals. *Journal of Homosexuality*, *49*(2), 23–38.

Buyse, A. (2014). Words of violence: Fear speech, or how violent conflict escalation relates to the freedom of expression. *Human Rights Quarterly*, *36*(4), 779–797.

Campbell, J. C., Hindle, A., & Stroulia, E. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, *26*(12), 2928–2941.

Common Sense. (2018). *Social media, social life. Teens reveal their experiences*. https://www.commonsensemedia.org/sites/default/files/uploads/research/2018_cs_socialmediasociallife_fullreport-final-release_2_lowres.pdf

Conway, M., Macnair, L., & Scrivens, R. (2019). *Right-wing extremists' persistent online presence: History and contemporary trends* (pp. 1–24). International Centre for Counter-Terrorism (ICCT). https://icct.nl/app/uploads/2019/11/RWEXOnline-1.pdf

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Twelfth International AAAI Conference on Web and Social Media* (pp. 42–51). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17910

Erjavec, K., & Kovačič, M. P. (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society*, *15*(6), 899–920.

Fang, L., & Woodhouse, L. A. (2017, August 25). How white nationalism became normal online. *The Intercept*. https://theintercept.com/2017/08/25/video-how-white-nationalism-became-normal-online/

Fawzi, N. (2019). Untrustworthy news and the media as "enemy of the people?" How a populist worldview shapes recipients' attitudes toward the media. *The International Journal of Press/Politics*, *24*(2), 146–164.

Forscher, P. S., & Kteily, N. S. (2020). A psychological profile of the alt-right. *Perspectives on Psychological Science*, *15*(1), 90–116.

Gagliardone, I., Gal, D., Alves, T., & Martínez, G. (2015). *Countering online hate speech*. UNESCO.

Günther, E., & Domahidi, E. (2017). What communication scholars write about: An analysis of 80 years of research in high-impact journals. *International Journal of Communication*, *11*, 3051–3071.

Hajok, D., & Selg, O. (2018). Kommunikation auf Abwegen? Fake news und hate speech in kritischer Betrachtung. *Jugend Medien Schutz-Report*, *41*(4), 2–6.

Hawdon, J., Oksanen, A., & Räsänen, P. (2017). Exposure to online hate in four nations: A cross-national consideration. *Deviant Behavior*, *38*(3), 254–266.

Hawley, G. (2017). *Making sense of the alt-right*. Columbia University Press.

Heikkilä, N. (2017). Online antagonism of the alt-right in the 2016 election. *European Journal of American Studies*, *12*(2). https://doi.org/10.4000/ejas.12140

Hine, G., Onaolapo, J., Cristofaro, E. D., Kourtellis, N., Leontiadis, I., Samaras, R., Stringhini, G., & Blackburn, J. (2017). Kek, cucks, and god emperor Trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In S. Gonzalez-Bailon, A. Marwick, & W. Mason (Eds.), *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)* (pp. 92–101). Association for the Advancement of Artificial Intelligence (AAAI).

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research*, *41*(4), 557–576.

Humprecht, E. (2019). Where "fake news" flourishes: A comparison across four Western democracies. *Information, Communication & Society*, *22*(13), 1973–1988.

Jardine, E. (2019). Online content moderation and the Dark Web: Policy responses to radicalizing hate speech and malicious content on the Darknet. *First Monday*, *24*(12). https://doi.org/10.5210/fm.v24i12.10266

Kinney, T. A. (2008). Hate speech and ethnophaulisms. In W. Donsbach (Ed.), *The international encyclopedia of communication*. Wiley. https://doi.org/10.1002/9781405186407.wbiech004

Kümpel, A. S., & Rieger, D. (2019). *Wandel der Sprach- und Debattenkultur in sozialen Online-Medien. Ein Literaturüberblick zu Ursachen und Wirkungen von inziviler Kommunikation* [The Changing Culture of Language and Debate on Social Media: A Literature Review of the Causes and Effects of Incivil Communication]. Konrad-Adenauer-Stiftung.

Landesanstalt für Medien NRW. (2018). *Hate speech und Diskussionsbeteiligung im Internet*.

Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, *28*(2), 434–443.

Ling, C., AbuHilal, I., Blackburn, J., De Cristofaro, E., Zannettou, S., & Stringhini, G. (2021). Dissecting the meme magic: Understanding indicators of virality in image memes. ArXiv:2101.06535 [Cs]. http://arxiv.org/abs/2101.06535

Lobinger, K., Krämer, B., Venema, R., & Benecchi, E. (2020). Pepe—just a funny frog? A visual meme caught between innocent humor, far-right ideology, and fandom. In B. Krämer & C. Holtz-Bacha (Eds.), *Perspectives on populism and the media* (pp. 333–352). Nomos. https://doi.org/10.5771/9783845297392-333

Magu, R., & Luo, J. (2018). *Determining code words in euphemistic hate speech using word embedding networks* [Conference session]. Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Brussels, Belgium.

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, *20*(6), 930–946.

McNamee, L. G., Peterson, B. L., & Peña, J. (2010). A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. *Communication Monographs*, *77*(2), 257–280.

Meibauer, J. (2013). Hassrede—Von der Sprache zur Politik [Hate Speech—From Language to Politics]. In J. Meibauer (Ed.), *Hassrede/Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion* (pp. 1–16). Gießener Elektronische Bibliothek.

Meza, R. (2016). Hate-speech in the Romanian online media. *Journal of Media Research*, *9*(26), 55–77.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2019). *Advances in pre-training distributed word representations* [Conference session]. LREC 2018 - 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan.

Mittos, A., Zannettou, S., Blackburn, J., & De Cristofaro, E. (2019, October 4). *"And we will fight for our race!" A measurement study of genetic testing conversations on Reddit and 4chan* [Conference session]. Proceedings of the 14th International AAAI Conference on Web and Social Media. ICWSM 2020, Atlanta, GA. http://arxiv.org/abs/1901.09735

Mondal, M., Silva, L. A., Correa, D., & Benevenuto, F. (2018). Characterizing usage of explicit hate expressions in social media. *New Review of Hypermedia and Multimedia*, *24*(2), 110–130.

Nagle, A. (2017). *Kill all normies: The online culture wars from Tumblr and 4chan to the alt-right and Trump*. Zero Books.

Pacheco, E., & Melhuish, N. (2018). *Online hate speech: A survey on personal experiences and exposure among adult New Zealanders*. Netsafe. https://www.netsafe.org.nz/wp-content/uploads/2019/11/onlinehatespeech-survey-2018.pdf

Phillips, J., & Yi, J. (2018). Charlottesville paradox: The 'liberalizing' alt-right, 'authoritarian' left, and politics of dialogue. *Society*, *55*(3), 221–228.

Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.

Prince, M. (2019, August 5). Terminating service for 8chan. *The Cloudflare Blog*. https://blog.cloudflare.com/terminating-service-for-8chan/

Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015). *Short and sparse text topic modeling via self-aggregation* [Conference session]. IJCAI international joint conference on artificial intelligence, Palo Alto, CA, United States.

Reicher, S. D., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology*, *6*(1), 161–198.

Roose, K. (2019). 'Shut the site down,' says the creator of 8chan, a megaphone for gunmen. *The New York Times*. https://www.nytimes.com/2019/08/04/technology/8chan-shooting-manifesto.html

Saha, P., Mathew, B., Garimella, K., & Mukherjee, A. (2021). *"Short is the road that leads from fear to hate": Fear speech in Indian WhatsApp groups* [Conference session]. Proceedings of the Web Conference 2021

Sanchez, B. C. (2020). Internet memes and desensitization. *Pathways: A Journal of Humanistic and Social Inquiry*, *1*(2), 1–11.

Schmitt, J. B., Harles, D., & Rieger, D. (2020). Themen, motive und mainstreaming in rechtsextremen online-memes. *Medien & Kommunikationswissenschaft*, *68*(1–2), 73–93.

Stewart, E. (2019, May 3). 8chan, a nexus of radicalization, explained. *Vox*. https://www.vox.com/recode/2019/5/3/18527214/8chan-walmart-el-paso-shooting-cloudflare-white-nationalism

Tuters, M., & Hagen, S. (2020). (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society*, *22*(12), 2218–2237.

Unkel, J. (2021). *tidycomm: Data modification and analysis for communication research* (Version 0.2.1) [Computer software]. https://CRAN.R-project.org/package=tidycomm

Viegas, F., Luiz, W., Canuto, S., Rosa, T., Gomes, C., Ribas, S., Rocha, L., & Gonçalves, M. A. (2019). *Cluwords: Exploiting semantic word clustering representation for enhanced topic modeling* [Conference session]. WSDM 2019—Proceedings of the 12th ACM international conference on web search and data mining, Melbourne, VIC, Australia.

Wendling, M. (2018). *Alt-right: From 4chan to the White House*. Pluto Press.

Author Biographies

Diana Rieger (PhD, University of Cologne) is a Professor of Communication Science at the Department of Media and Communication at LMU Munich. Her current work addresses characteristics and effects of hate speech, extremist online communication, and counter voices (e.g., counter narratives and counter speech).

Anna Sophie Kümpel (Dr. rer. soc., LMU Munich) is an Assistant Professor of communication at the Institute of Media and Communication at TU Dresden. Her research interests are focused on media uses and effects, particularly in the context of social media, (incidental exposure to) online news, and media entertainment.

Maximilian Wich (MSc, University of Mannheim) is a PhD student at the Technical University of Munich. His research interests include hate speech detection with machine learning (ML) and explainable artificial intelligence (XAI).

Toni Kiening (BA, LMU Munich) is a graduate from the communication science curriculum at the Department of Media and Communication at LMU Munich.

Georg Groh (Dr. rer. nat., TU Munich) heads the Social Computing research group at the Department of Informatics of the Technical University of Munich. His research focuses on modeling and inferring social context, for example, via ML-based natural language processing.