JOURNALISM STUDIES
2022, VOL. 23, NO. 2, 224–244
https://doi.org/10.1080/1461670X.2021.2017793

Routledge
Taylor & Francis Group

"The Boundaries are Blurry…": How Comment Moderators in Germany See and Respond to Hate Comments

Sünje Paasch-Colberg [a,b] and Christian Strippel [a,c]

[a]Institute for Media and Communication Studies, Freie Universität Berlin, Berlin, Germany; [b]DeZIM. German Centre for Integration and Migration Research, Berlin, Germany; [c]Weizenbaum Institute for the Networked Society, Berlin, Germany

ABSTRACT

Aggressive and hateful user comments on news sites and social media threaten discussions on the Internet and pose a difficult challenge for content regulation. Previous research has mainly focused on the analysis of moderation strategies in dealing with such comments. In contrast, little attention was paid to the issue of which comments are considered problematic by content moderators in the first place. The answer to this question has more than theoretical relevance, but practical significance against the backdrop of increasing efforts to automate the detection of hate speech or toxicity in user comments. Based on 20 interviews, this paper explores what comment moderators in Germany consider to be hate comments, how they moderate them, and how differences in moderation practices can be explained. Our findings show strong agreement regarding extreme cases of hate comments, whereby there is overlap with the theoretical concept of hate speech, but also forms of incivility. Moreover, the interviews revealed differences in the perception and handling of hate comments, which can be linked to explanatory factors at the levels of the individual, professional routines, and the organization.

KEYWORDS

Comment moderation; community management; hate speech; incivility; interviews; social media; user comments

Introduction

Aggressive and hateful user comments are a growing problem for discussions on news sites or social media (Coe, Kenski, and Rains 2014; Su et al. 2018) and pose a serious challenge for content regulation. In moderating user comments, journalists and community managers have to balance journalistic principles, economic interests, and legal requirements (Braun and Gillespie 2011, 384). At the same time, these professionals themselves are increasingly exposed to intimidation, insults, and hate speech that target individual journalists or journalism as an institution (Chen et al. 2020). Those developments affect journalistic working routines and can lead to self-censorship, for example, when journalists avoid reporting on certain topics or groups because of such threats (Binns 2017).

CONTACT Sünje Paasch-Colberg s.colberg@fu-berlin.de
Supplemental data for this article can be accessed https://doi.org/10.1080/1461670X.2021.2017793.

Another consequence is that news organizations in various countries have shut down their comment sections (Harlow 2015; Quandt 2018).

With the new opportunities for audience participation and the challenges they raised, a new journalistic role has emerged in newsrooms that enforces journalistic gatekeeping in these discussion environments: comment moderators who filter, edit, moderate, and encourage user comments (Bakker 2014). A growing body of literature has examined their working routines and moderation strategies, predominantly focusing on the moderation of deviant user comments (e.g., Chen and Pain 2017). In contrast, only a few studies have addressed the question of *why* certain strategies are used. For example, Wintterlin et al. (2020) analyzed the role of organizational factors, while Frischlich, Boberg, and Quandt (2019) focused on individual characteristics of content moderators. To our knowledge, however, no study has yet investigated what forms of user comments are actually considered problematic by comment moderators, that is, which working definitions of hate comments guide moderation decisions in newsrooms.

Answering this question is not only of theoretical relevance for researching gatekeeping and comment moderation, but also has practical significance against the background of increasing efforts by both scholars and the industry to develop approaches for automated detection of "hate speech" (Fortuna and Nunes 2018), "incivility" (Su et al. 2018), or "toxicity" (Jigsaw 2021) in user comments. In order to contribute to these two lines of research, we conducted 20 interviews with community managers and comment moderators working for a broad range of German-language news media. In doing so, we focused on what comment moderators understand by hate comments and which factors explain differences in their responses to them. Since the task of comment moderation is integrated differently into newsrooms and assigned to both journalists and non-journalists (e.g., designated community managers that do not necessarily have journalistic training), our selection procedure focused on actual experience with moderation and included different professional roles (see Section 4). For simplicity, we will refer to these intereviewees as "comment moderators" in the following.

As the conceptional framework guiding our study, we introduce a model of "moderation factors" in Section 2. This model structures explanatory factors for gatekeeping practices in comment moderation at five different levels, ranging from micro to macro. To contextualize our findings on the comment moderators' understandings of hate comments, we discuss various terms and definitions that are prominently used in the public debate and the academic literature on the phenomenon in Section 3.

Comment Moderation as Gatekeeping

Commenting on articles on news websites and social media is a popular form of user participation in journalism (Ziegele, Jost et al. 2018). Published in a journalistic context, such comments have the potential to enhance democratic deliberation (Springer, Engelmann, and Pfaffinger 2015). However, these discussion environments also allow for deviant user participation (Quandt 2018), including the spread of aggressive and hateful content. In response, many newsrooms have adopted some form of community management, often including the moderation of user comments. In this context, comment moderation can be defined as "any kind of institutional engagement aimed at the organization or regulation of the process or content of online discussions" (Ziegele and Jost 2020, 894).

Early studies consistently revealed ambivalent attitudes among journalists towards user-generated content in general. For user comments specifically, their high volume and overall poor quality were identified as the main challenges for newsrooms (Braun and Gillespie 2011; Reich 2011). To address these challenges, most news organizations have implemented a variety of measures such as etiquette guidelines, registration procedures, automated filters, pre- and post-moderation strategies, as well as rating systems or other design elements that reward constructive comments (Reich 2011). In the past years, such practices of monitoring and moderation have normalized as journalistic working routines (Chen and Pain 2017), enforcing the journalistic gatekeeping role.

Various moderation strategies in particular have received attention in the literature: "Interactive moderation," for example, refers to an active and visible style of moderation and includes responses to user comments or questions, whereas deleting comments or blocking users are examples for "non-interactive moderation" (e.g., Boberg et al. 2018). Finally, "collaborative moderation" involves the community of users who are allowed to evaluate or flag other users' comments (Ziegele and Jost 2020). Some studies in the field have worked with more differentiated taxonomies and have described moderation strategies using the dimensions of "interactivity" and "authority," for example (Frischlich, Boberg, and Quandt 2019; Wintterlin et al. 2020).

To explain why certain moderation practices are used, some authors (Frischlich, Boberg, and Quandt 2019; Wintterlin et al. 2020) refer to the hierarchical influences model by Shoemaker and Reese (2014). This model accounts for the multiple influences within a media organization and beyond "that simultaneously impinge on media and suggests how influence at one level may interact with that at another" (1). In particular, it asserts that the shaping of media content needs to be analyzed from different theoretical perspectives and at different analytical levels. For this purpose, it considers factors at five different levels, namely the: (1) individual journalist, (2) working routines, (3) media organization, (4) social institutions, and (5) social system (4–9). In this way, the model aims to offer a larger theoretical framework to organize and integrate otherwise unrelated observations, to interpret findings on one level against the constraints of others, and to encourage cross-national studies (1–15).

In the context of comment moderation, some studies used this model to relate explanatory factors on different levels and account for mediation effects. For example, it has been used to examine how the editorial leaning of a newspaper, the affordances of its comment sections, the prevalence of "dark" user participation, and the moderation practices interact (Wintterlin et al. 2020). In previous work, we used the hierarchy of influences framework to systematize the (often unrelated and partly contradictory) empirical findings and hypotheses regarding factors that influence comment moderation and argued for a corresponding model of "moderation factors" (Paasch-Colberg et al. 2020). According to this model, moderation decisions can be explained by the complex interplay of different forces, including (1) the attitudes and characteristics of an individual comment moderator, (2) emerging working routines, (3) the funding and the editorial line as organizational influences, (4) constraints from social institutions such as legislation, ethical values of the profession, and technological infrastructure, and (5) the cultural and historical context of a social system (see Figure 1). This model can thus be used to systematically describe, compare, and explain the moderation practices of different newsrooms in the same

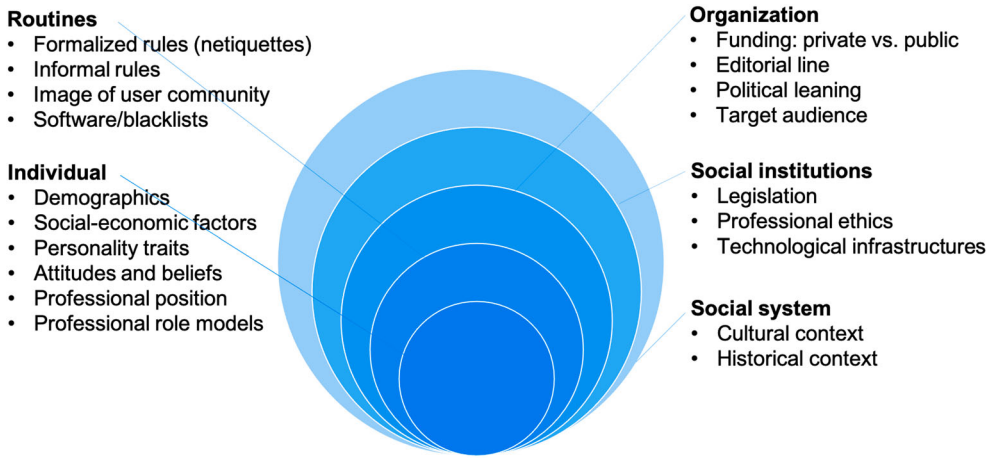

Figure 1. Explanatory factors for comment moderation.

national context as well as in different countries. Accordingly, we can also use it to discuss and interpret the literature on comment moderation in the following.

To begin with, it seems obvious that the actual content of a user comment is decisive for how it is processed in content moderation and, moreover, that *professional routines* are developed in this regard. Indeed, some studies show that certain types of deviant user comments are dealt with in a similar way across different media organizations. For example, Ziegele, Jost et al. (2018) have shown that public-level incivility (i.e., violations of democratic norms) in user comments on 15 Facebook pages of German news sites was associated with an increase in interactive moderation, whereas personal-level incivility (i.e., violations of interpersonal politeness norms) was not (543). In addition, the findings of Wintterlin et al. (2020) suggest that news sites use authoritative moderation practices when confronted with racist hate speech, regardless of their editorial leaning (916).

However, the findings of two other studies are somewhat contradictory, indicating that professional routines are not always consistent. Moderation decisions here appear to have also been affected by other factors. Analyzing more than 600,000 user comments from the popular German news website *Spiegel Online*, Boberg et al. (2018) showed that the prevalence of swearing is not significantly higher in the group of deleted comments than in the non-deleted comments. On the other hand, in a similar study using over 9 million user comments posted to the *New York Times* website, Muddiman and Stroud (2017) found that comments containing swear words actually were significantly more likely to be rejected for publication. In the same study, a similar but weaker effect was found for the use of uncivil language beyond swearing. In the light of these finding, it can be argued that incivility as a broader concept leaves more room for subjective moderation decisions and, furthermore, that cultural and social differences can affect how user comments are handled in different countries.

In fact, several studies support the assumption that factors at the *individual level* can affect moderation decisions. For instance, the work of Chen and Pain (2017) suggests that the journalist's self-image can influence their moderation of uncivil user comments.

In this study, journalists who identify themselves primarily as information distributors (and less as community builders) referred to comment sections as public rather than journalistic spaces, and tended not to engage in user discussions (884). In another study, Frischlich, Boberg, and Quandt (2019) analyze how journalists deal with different forms of "dark participation," understood as deviant, norms-transgressing commenting behavior. Their findings show that journalists differ in their notion of such behavior and, thus, in their moderation strategies. They identify four types of moderators: (1) "unconcerned gatekeepers" who tend to see deviant comments as unproblematic and use authoritative, non-interactive moderation practices, (2) "relaxed gate-watchers" who consider such comments a normal aspect of user participation and moderate in a more participatory way, (3) "alarmed guards" who report a high prevalence of dark participation and tend to use authoritative practices in moderation, and, finally, (4) "struggling fighters" who associate dark participation with the potential of negative consequences and mostly apply non-interactive and authoritative moderation (2023–2027).

In addition, there is initial empirical evidence of the influence of *organizational factors*. According to a quantitative survey of German journalists by Wintterlin et al. (2020), for example, the use of discursive-interactive moderation increased with the number of channels for user participation. Further, this study showed that left-wing media in particular tend to be more discursive in their moderation than other media (915–916).

Moreover, a cross-country study found differences in how comment sections are regulated in Finland, Sweden, the Netherlands, and the UK. The authors attribute these disparities to different legal and media ethics frameworks, that is, the *social institutions* level, but also to the political climate at the *social system level* (Pöyhtäri 2014, 517–519). At the same time, moderation practices are further shaped within this national framework by organizational factors, for example, the media brand and its policies (520).

Ferrucci and Wolfgang (2021) emphasize that the management of user comments on news sites increasingly relies on third-party platforms and technology companies, leading to less transparency in moderation and a loss of journalistic control (1020). In terms of the hierarchy of influences, this means that actions and decisions of non-journalistic *social institutions* increasingly shape comment moderation. The authors findings show that while "outside moderation" can handle large communities more efficiently, journalists are concerned that this approach leads to a decrease of the quality of user comments and community engagement. For example, when moderation is directed to an outside platform or automated tool, the comment removal decisions focus solely on specific swear words and terms. By contrast, study participants indicated that misinformation or boarderline cases of hate speech were largely neglected (1019–1021).

Capturing the Unwanted: Hate Speech, Incivility, and Toxicity

Following our model of moderation factors, we assume that factors at different analytical levels shape the comment moderators' understandings of hate comments, which, in turn, guide their moderation decisions. For example, legal definitions at the social institutions level are expected to influence what type of content comment moderators consider problematic and how they handle such comments (Pöyhtäri 2014). On the other hand, current debates about the benefits of diversity in newsrooms (Harlow 2015, 23–24) point to how a journalist's individual background might also affect their moderation

decisions. There are different concepts and terms at each level that are relevant to the handling of problematic content in user comments. As they are linked to distinct (theoretical) assumptions and stem from diverse contexts, we will briefly outline them in the following. This map will be used in Section 5.1 and 5.2 to contextualize our findings.

At the *social system level*, it is primarily Germany's cultural and historical context that likely shapes comment moderators' perceptions of certain political issues and understanding of certain terms. These contexts become visible, for example, in public debates around racism, anti-Semitism, sexism, or homophobia, which refer to specific forms of historic discrimination against certain minorities. In the context of user-generated content, the term "racism," for instance, refers in particular to communication that "seeks to denigrate or discriminate against individuals (by denying equal rights, freedom, and opportunities) or groups because of their race or ethnicity" (Bliuc et al. 2018, 76). However, such forms of group-focused enmity are also increasingly addressed both in public and research with overarching concepts such as "hate speech." This term was coined by critical race theorist Matsuda (1989), who called for legal sanctions against racist speech targeting "historically oppressed" groups (2357). The term then migrated from a legal context to other disciplines and ultimately entered the public discourse. According to Brown (2017), this development may have resulted in the term losing its meaning and becoming "merely an empty vessel" (427) incapable of denoting specific phenomena.

According to most definitions, hate speech is directed against certain groups based on specific characteristics. However, the characteristics that define a group can vary widely. Some definitions center on characteristics such as religion, race, or gender (United Nations 2020), while others stress that hate speech can target any possible group (Parekh 2006). Moreover, the definitions of scholars, social media platforms, and legislators typically emphasize different aspects of hate speech. These can be broadly differentiated into three strands: approaches that (1) emphasize the intention to cause harm, (2) address potential harms, or (3) define certain content characteristics (Sellars 2016, 14–18). Such content-based definitions usually understand hate speech as statements that negatively stereotype, demonize, or dehumanize the (presumed) members of a group, or that refer to violence or death threats against them (Bahador and Kerchner 2019).

A second overarching concept in the public debate, and thus at the social system level, is "incivility." Although the term is established in communication studies, it is less visible in public discourse. However, we argue that the concept of incivility is often alluded to in this context, for example, when reference is made to the harsh tone in online forums. Generally speaking, incivility is norm-transcending communication that can lead to an increase of polarization and aggression, on the one hand, and a decrease of trust in news reporting and media brands (Chen et al. 2019, 2), on the other. There are various definitions of incivility in the literature: Papacharissi (2004), for example, relates the concept to democratic norms and defines public-level incivility as "the set of behaviors that threaten democracy, deny people their personal freedoms, and stereotype social groups" (267). Other definitions are less specific and refer to norms of politeness and personal-level civility. With regard to user comments, Coe, Kenski, and Rains (2014) define incivility as "features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topics" (660).

At the *social institutions level*, national legislation has been identified as an important explanatory factor for comment moderation (Pöyhtäri 2014). German law, however, does not cite racism or hate speech as separate criminal offenses. Instead, it refers to a number of other criminal offenses such as defamation or insult, incitement to hatred and violence, Holocaust denial, and the display of the swastika (Brugger 2003b). The criminalization of these offenses is justified primarily on the basis of the fundamental rights guaranteed in the German constitution, including the protection of human dignity, or the prohibition of discrimination (Brugger 2003a). In the same manner, the German Press Code, the prevailing standard for journalism ethics and self-regulation in Germany, refers to both human dignity and discrimination (German Press Council 2017; Sections 1 and 12). One of the most recent additions to the Code indicates the relevance of these guidelines for comment moderation on news websites. Since 2015, the guideline has included a paragraph that commits editors to "ensure compliance with journalistic principles if they detect violations through user-generated content or if such violations are pointed out to them by third parties" (Section 2.7).

In debates about restricting and sanctioning hate speech, however, the constitutional rights mentioned above are often contrasted in Germany with "freedom of opinion" as another fundamental right in the constitution. In the case of hate speech on social media, German lawmakers specifically addressed these conflicting rights in the fall of 2017 with a heretofore internationally unprecedented law: the Network Enforcement Act ("Netzwerkdurchsetzungsgesetz" or "NetzDG"). This law obliges big social media platforms to delete "obviously illegal content" within 24 h, imposes fines in the case of systematic failure, and requires the platform providers to publish transparency reports (Tworek and Leerssen 2019). Although primarily aimed at big social media platforms, the NetzDG may also affect comment moderation in newsrooms. Newsrooms not only use social media to distribute their content and engage users, but the law may also influence users' expectations of comment moderators (Loosen and Schmidt 2012, 876–878), particularly with regard to the deletion of user comments they consider problematic.

As discussed in Section 2, moderation software and the use of (semi-) automated detection tools are becoming more prevelant in comment moderation. Among other things, the concept underlying the implementation and training of a detection tool determines its prediction results and biases. In the emerging field of algorithmic classification, this often involves concepts such as "toxicity," which are not derived from a theoretical position, but rather stem from the practical challenges of content moderation. Generally referring to the effects of specific comments on other users, toxicity is used inconsistently and serves as a label for various forms of harmful content (van Aken et al. 2018). The Perspective API, for example, developed by Google's subsidiary *Jigsaw*, defines a toxic comment as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion" (Jigsaw 2021). Another relevant concept in this area is "offensive language," which refers to more commonplace forms of norm-breaking language, for example, the use of profanity (Xiang et al. 2012).

What scholars, policy makers, industrial developers, and developers of automatic detection tools have in common is that they all have difficulty drawing a clear line between hate speech, incivility, defamation, or toxicity, on the one hand, and ordinary disagreement, dislike, or civil speech, on the other (Stryker, Conway, and Danielson 2016, 541). Any understanding of these phenomena is shaped by norms, context, and the

interpretative framework (Chen et al. 2019; Saleem et al. 2017), which is indicated by low reliability scores (Ross et al. 2016) and the susceptibility of automated tools to manipulation and bypassing (Hosseini et al. 2017). Keipi et al. (2017) stress that the definition and measurement of hate is relatively straightforward only in cases of extreme hate (54). It is therefore reasonable to assume that comment moderators have difficulty clearly delineating acceptable from unacceptable comments. Our first research question is this the following:

RQ1: Which types of user comments do comment moderators consider hate comments?

To add to our understanding of the various moderation practices discussed in Section 2, while also accounting for subjective differences between comment moderators themselves, we further asked:

RQ2: How do comment moderators respond to different types of hate comments?

Finally, since there remains little actual evidence on the factors influencing moderation practices and the existing studies focus on selected factors (see Section 2), we sought for an open exploration of our interview material to expand the base of empirical knowledge of moderation factors. Toward this end, and to demonstrate the capacity of the model to reconcile otherwise inconsistent findings, we asked:

RQ3: What explains the possible differences in the moderators' understandings and moderation of hate comments?

Method

To answer these research questions, we conducted 20 interviews with a total of 23 comment moderators working for German-language news sites and social media channels.[1] We selected interviewees from a variety of organizational contexts to cover a broad range of experience with moderating user comments. To this end, we identified three relevant factors in the literature that are likely to cause disparities in this area: media type (Domingo 2011), readership (Ziegele, Weber et al. 2018, 1420–1421), and discourse architecture (i.e., the technological design of a discussion environment; Ziegele 2016, 161–163). Since public broadcasting in Germany has an opinion-forming and integrating function as part of its public mandate (Schulz et al. 2008), we further sought to

Table 1. Selection criteria.

Selection criteria	Specification	No. of interviews
Media type	Print	13
	Television	3
	Radio	1
	"Internet only"	3
Funding	Public service	4
	Commercial	16
Readership	Local	4
	National	16
Discourse architecture[a]	Comment sections	14
	Discussion or Q&A forum	4
	Social networks	4

[a]Since two websites offer both comment sections and a discussion forum, they are listed twice in this table.

investigate whether the funding of media organizations affects their moderation practices. Accordingly, the group of interviewees included individuals working for newspapers (on both a national and regional level), for both public and commercial television, and a public radio program (see Table 1). With regard to discourse architectures, our respondents mainly moderate user comments in the comment sections of news sites ($n = 14$), in discussion or Q&A forums ($n = 4$), or on social media ($n = 4$), including YouTube, Twitter, and Facebook.

When approaching the selected newsrooms, we asked to interview a staff member with day-to-day experience in comment moderation. In this way, we left it up to the newsrooms to decide who we interviewed. We chose this open approach on the basis of informal conversations with journalists and commentators prior to the study, in which we learned that newsrooms integrate moderation quite differently due to their size, organizational structure and resources: In some newsrooms, comment moderation is done by journalists as one task amongst others, while others employ designated comment moderators, community managers, or audience developers.[2] Thus, while all the interviewees are employed by their news organizations, there were differences in their professional roles (see Table 2). Our respondents either worked hands-on in comment moderation (i.e., comment moderators, journalists) or were responsible for a team of moderators (e.g., community Managers, community development, editor in chief).[3] In three cases, we interviewed two staff members in a double interview, following the suggestion of the corresponding newsrooms.

All interviews were conducted in January and February 2018, most of them over the phone. Each interview was led by two trained interviewers on the basis of a semi-structured interview guide with 19 main questions and 10 follow-up questions (Strippel et al. 2021). All the questions were open-ended, covering four key areas: (1) working

Table 2. Sample description ($n = 20$ Interviews with $n = 23$ interviewees).

	Individual characteristics		Organizational characteristics		
No.	Professional role	Gender	Media type	Readership	Discourse architecture
1	Editor-in-chief	Female	Print	Local	Comments
2	Online journalist	Male	Print	Local	Comments
3	Community manager	Male	Print	National	Forum
4	Community manager Comment moderator	Male Male	Internet	National	Forum
5	Community manager	Male	Print	Local	Comments
6	Community manager	Female	Print	National	Comments
7	Editor-in-chief	Male	Internet	National	Comments, forum
8	Audience Development	Female	Radio (public)	National	Social network
9	Online journalist	Female	Print	National	Comments
10	Online journalist	Female	Print	National	Comments
11	Community manager	Male	Print	National	Comments
12	Community manager	Female	Print	Local	Comments
13	Community manager	Female	Print	National	Comments
14	Comment moderator	Female	Print	National	Comments, forum
15	Community manager	Male	Internet	National	Comments
16	(Deputy) Editor-in-chief	Female	TV (public)	National	Comments
17	Editor-in-chief Online journalist	Female Female	Internet (Public)	National	Social network
18	Community manager	Female	Print	National	Comments
19	Community manager Community manager	Male Female	TV (public)	National	Social network
20	Head of communication	Male	TV	National	Social network

conditions and professional self-image, (2) approaches to user comments in general, (3) approaches to hate comments in particular, and (4) moderation practices. In order to answer RQ1 and RQ2, we focused on areas 3 and 4. Thus, the interviewees were asked what they understand to be a hate comment and where they draw the line with respect to comments that are considered unproblematic. We also asked what types of hateful comments they encounter and what measures they take in moderating these comments. With regard to RQ3, questions from areas 1 and 2 were relevant. Consequently, we asked what goals the interviewees pursue in their work, how they view their user community, and what moderation style they pursue in their team. On average, the conversation lasted 50 min.

After the interviews, we carried out a qualitative content analysis of the interview transcripts using MAXQDA and combining techniques of deductive and inductive category building (Schreier 2014, 9). With regard to RQ1 and RQ2, we approached the material in an explorative way. For this reason, the initial coding scheme included only three pre-set categories: (a) working definition, (b) types of hate comments, and (c) moderation practices. Concerning RQ3, four main theoretical categories were derived from the hierarchy of influences model (Shoemaker and Reese 2014). These include the analytical levels of the (d) individual, (e) professional routine, (f) organization, and (g) social institutions. The quotes from the interviews used in the following were translated; the orginal German quotes are found in the supplementary document.

Findings

In this section, we answer our three research questions. The interview material is described in reference to the subcategories derived in the content analysis and by means of illustrating our findings using typical quotes. The findings concerning RQ1 and RQ2 are presented in 5.1 and 5.2, contextualized within the concepts and terms in Section 3. The findings on RQ3 are presented in 5.3, interpreted through the lens of the moderation factors model elucidated in Section 2.

How Comment Moderators See Hate Comments

Consistent with the findings of other studies mentioned at the outset, hateful user comments are also a challenge for the editorial teams in our sample. Most interviewees remarked that they or their teams have felt overwhelmed by the volume and tone of user comments since some years. At the same time, several interviewees had trouble describing what they understand to be a hate comment. For one interviewee, hate speech was "a completely vague term" (#13). Others noted similarly:

> For me, it becomes clearer when I break it down into racism, sexism and death threats, for example. (#8)

> I think the term is imprecise and flawed. There are racist comments, there are anti-Semitic comments, there are comments that incite hate […]. All this is subsumed under the term "hate comments." But I think that this term is too ambiguous to be of any real use. (#7)

Although overarching concepts such as hate speech are increasingly used in the field of research and automation, some interviewees thus prefer to work with more specific

concepts. Their criticism of terms such as "hate comments" or "hate speech" supports Brown's (2017) observation that they become diluted and ultimately useless if society adopts them as blanket terms for myriad purposes (427).

On the other hand, the various forms of group-focused enmity mentioned by our interviewees exemplify what they all agreed were clear cases of hate speech. These include threats of violence or murder, calls for violence, Holocaust denial, and insults and comments that discriminate against a certain group. Some interviewees described them as "cases relevant under criminal law." This is striking as the designation implies that the legal relevance of such comments is readily apparent to them. This reminds of the phrase "obviously illegal content" in the NetzDG, which also suggests that the legality of content can be evaluated and classified more or less immediately. The correspondence here between real-world practice and the law seems to indicate that the public discussion about the NetzDG has had a significant impact on the practice of comment moderation in German newsrooms, despite the legislation's focus on big social media platforms.

Nevertheless, several interviewees emphasized that such clear-cut cases are rare. They spent most of their time dealing with less obvious cases, as illustrated by the following quote:

> The comments in the gray area are actually the rule. And that is the problem with hate speech. It is not the norm that someone posts an explicit call to murder or radical right-wing slogans. (#12)

This "gray area" includes ironic comments, word play, rhetorical questions, or disparaging modifications of people's names. The interviewees also mentioned non-linguistic references such as when a gun is posted in a comment thread on refugees, when cheering emoticons appear under news on drowned refugees in the Mediterranean Sea, or when a news article about a rape or murder is shared to threaten someone. The practice of comment moderation thus confirms the context sensitivity of the phenomenon highlighted in the literature (Saleem et al. 2017):

> We often see comments in which experienced users refrain from using racist language, but nevertheless make racist remarks. […] Such statements would perhaps be acceptable in another context, but if you take a closer look and put them into context, they are racist or insult a religious community or homosexuals, for example. (#16)

In summary, while our respondents were largely dismissive of the term "hate speech," their sense of clear cases corresponded to the term's conceptual underpinning. Their general understanding of hate comments, however, was broader than most hate speech definitions. It even included insults against individuals that rather fall under the definitions of incivility or offensive language (see Section 3). One interviewee explained this discrepancy by emphasizing that it is crucial to ask how people should address each other in the public sphere and that this question cannot be adequately answered solely by legal assessments:

> If you take the criminal code as a basis, you make things relatively easy for yourself. The law allows for many things that we do not want to see in social interactions and above all in social networks. (#4.2)

As discussed in Section 3, hate speech definitions tend to focus on three different elements: intentions, harms, and content. All of these aspects were in fact addressed in

our interviewees' answers. When it comes to deciding whether a comment is a hate comment, most interviewees referred to its content. By the same token, some also assessed the intention behind the comment or its anticipated consequences:

> What do we consider a hate comment? I would put it like this: First of all, the intention behind it is to cause harm. Be it to insult or degrade another person, to defame or denigrate an entire group, or simply to stir up political turmoil. (#4.2)

> Opinions are also formed in this way when there are stupid comments under each of our articles and we are called into question, our readers are called into question, and politics is called into question. You have to be careful that this doesn't have a negative effect on the passive readers. (#14)

Comments, then, are problematic if they can potentially harm others or damage the reputation of the news site or the political system in general. Other interviewees reported that even otherwise acceptable comments can have a negative influence on the further course of a discussion and trigger a veritable downward spiral:

> It starts with a first comment that insinuates something but is not yet worth deleting, which then gives many subsequent readers the justification to go one step further and write hate comments. (#3)

In response to RQ1, our interviewees generally agree that group-focused enmity and insults are clear cases of hate comments. Their understanding is thus consistent with the hate speech concept in the literature, but additionally includes certain forms of incivility and offensive speech. When evaluating user comments, respondents cited the aspects of content, intent, and harm as relevant touchstones. Simultaneously, they emphasized the contextual sensitivity that makes it difficult to draw a clear line between acceptable and unacceptable comments. This finding is in line with the observation of Keipi et al. (2017) that identifying online hate is simpler in the case of extreme hate compared to those comments at "the other end of the hate classification spectrum" (54).

How Comment Moderators Respond to Hate Comments

It is precisely this distinction, however, that is significant for the question of how our interviewees moderate hate comments. A consensus emerged among our interviews regarding extreme or clear cases. Specifically, comments of this type are usually deleted or hidden ("shadowbanned"). Sometimes, they are even reported to the police. Other authoritative practices the interviewees named in the context of clear hate comments include blocking users or disabling user comments altogether.

Regarding the much larger proportion of implicit and less clear cases of hate comments, the moderators seem to rely more on their subjective judgment. After all, these comments are often highly context-sensitive (as discussed in 5.1) and therefore allow for less restrictive and more interactive responses. Such moderation practices include various ways of showing presence in discussions such as reminding people of the netiquette, asking for civility, raising counterarguments, or asking questions:

> It always helps a lot to ask what is actually meant [or] why the person made the statement […] It feels like they blurt out something pretty quickly and if you then follow up with questions, they have a bit more time to think it over. Most of the time this works quite well to bring things to a more factual basis. (#18)

In reflecting on the tone of moderation responses, our interviewees confirm the findings of prior studies (Ziegele and Jost 2020). They stated that factual responses were helpful in deescalating user discussions; however, they also cautioned against using humor, irony, or sarcasm:

> We had the best experience when we can respond to users with a factual comment. I think humor is a very difficult thing, because with humor there's always the question of who is supposed to get the joke. The moment you make fun of the person who commented, it can be very tricky. In my experience, that can really backfire. (#12)

> The problem with [responding to] hate comments is that sarcasm, irony, or jokes are often not recognized. (#5)

However, our interviewees agreed that deciding whether a specific user's comments are better off being deleted or directly addressed is a challenge that requires both experience and instinct. In this context, several interviewees reported that they double-checked with colleagues, deferred to team decision-making, or relied on regular trainings using real-world examples.

Most respondents described appropriate moderation responses as being constrained by limited resources that often do not allow for interactive moderation. We learned that this is one of the reasons why some of them have begun to focus their resources on constructive rather than aggressive user comments. It turns out that this has improved both their own sense of well-being and the civility of the discussion environments:

> In the meantime, we have noticed that it is not helpful to pay so much attention to these negative posts, but rather to emphasize the more constructive ones. Now not only are we better off, but it also seems that the forum and the community are as well. It feels like the discussions are more constructive and of higher quality. (#18)

> I think this whole discourse around hate speech is so focused on the idea that the discussions are all negative, and that's just not true. I think that if you somehow start to change your perspective and say, "Ah, someone is saying something positive, I'd rather focus on that," then the mood also changes. And I think that we also are able to notice this. (#8)

To summarize in reference to RQ2, it can be stated that clear cases of hate comments are usually responded to with authoritative, restrictive measures (e.g., deletion), while less clear cases receive a larger range of responses (incl. interactive, discursive moderation practices). This "toolbox" is similar to the moderation practices described in other studies in the literature (Wintterlin et al. 2020; Ziegele, Jost et al. 2018). Moreover, our analysis additionally points to the strategy of shifting the focus to constructive comments as a way of discouraging hateful or otherwise hostile user comments.

Explaining Differences in the Perception and Moderation of Hate Comments

The findings presented in 5.1 and 5.2 reveal differences in how hate comments are understood and handled. However, these differences seem to relate mainly to less extreme forms of hate comments, while there was a strong consensus among our interviewees regarding extreme cases. In our view, this may be explained by Germany's legal requirements. This macro-level factor has come to bear on similar routines for handling extreme cases across different newsrooms and their community management teams (e.g., netiquettes, terms of services). In addition, the influence of German law also became

evident when many of our interviewees used the formulation "cases relevant under criminal law" to label extreme cases. Nevertheless, the interviews also revealed differences in the respondents understanding of what hate comments are, especially with regard to the question of where the line between acceptable and unacceptable comments should be drawn and how comments in the gray area should be handled. We discuss these differences in terms of the levels of the model of moderating factors described in Section 2.

Individual Level

Many interviewees emphasized that the evaluation of a specific user comment can be subjective. Statements in this vein generally referred to individual differences in the assessment of user comments. When asked about how they would distinguish between acceptable and inacceptable comments, one interviewee answered:

> The boundaries are blurry. For example, I'm relatively desensitized; I just often don't notice [hate comments], think it's okay and say, "Yes, you can say that," while some of my colleagues then say, "No, you can't say that." (#8)

At the same time, some respondents specifically referred to socioeconomic differences and attitudes as the underlying explanatory factors for these individual differences:

> There will always be differences depending on your background. You could be a young person, an old person, a man or a woman. Everyone perceives things differently. (#13)

Routines Level

The practices of double-checking with colleagues and working out guidelines (see 5.2) indicate that individual approaches in comment moderation are often viewed critically. There appears to be an effort toward establishing shared working routines, both formal and informal:

> We're all human, and of course you try to refrain from expressing your own opinion a little bit. That's not always easy and doesn't always work. But if we don't like something and actually want to delete it, we at least try to follow up. We at least have this double check again so that our own opinion doesn't come through too strongly. (#18)

Our interviewees in fact expressed that working routines regarding the content of user comments have been established. Several interviewees explained that their moderation depends on the topic of a discussion or the group that is targeted in a comment and that they tend to moderate more strictly if vulnerable groups are targeted:

> In the case of hate towards us, of course, we can withstand a little more of it, as long as no one specific person is affected by it. But when it comes to insults among [the community], we simply have to intervene with community management. (#19)

> When difficult topics arise, we are vigilant and give them the necessary attention. […] We definitely moderate such topics more strictly. (#6)

The interviewees, moreover, referred to the different perceptions they have of their communities. This factor can also be assigned to the level of routines. Some of them stressed their intention to provide a safe space for all users and therefore to moderate more

strictly; others emphasized the importance of an independent opinion-making process and a less strict moderation practice:

> We make sure that everyone feels welcome here, regardless of age, nationality, gender, or origin. That's why everything that is discriminatory or offensive is filtered out. (#5)

> We let people decide who they want to agree with. They can usually see the initial post, they can see our response, and every reader gets to make up their own mind and decide how they're going to respond. […] I think most readers are intelligent enough to make the decision themselves. (#13)

With respect to the community and other users, one interviewee described a rule of thumb for estimating the intent behind a comment and assessing whether the user is genuinely interested in a discussion or prefers to disrupt it:

> Sometimes I feel like a user just posts a short message to get all the other people involved and create a stir. 50 comments then follow, but the user who originally posted that message is no longer part of the discussion. It then looks like he wasn't really interested in the discussion in the first place, but just in nudging people in a certain direction. […] But I'm still going to be critical and look at what happened there and how I evaluate it. (#3)

Finally, the interviews echo the growing importance of software for the practice of comment moderation (see Section 2). Our interviews showed that almost all newsrooms use some sort of software. Here, tools from external providers are more common than in-house solutions. On the basis of word lists ("blacklists"), these tools typically filter incoming user comments (either before or after their publication) into three categories: publish, delete (or hide), or review by human moderator. Several interviewees reported that such word lists are prone to errors due to the lack of context and regularly need to be manually updated. Moreover, external solutions initially needed to be adapted to internal moderation rules. At the time of the interviews, only few newsrooms used algorithmic systems that are continuously trained by the decisions of the human moderators. Thus, the use of moderation software has affected moderation routines in the newsrooms, for example, by imposing the post-moderation mode or the extra step of reviewing the tool's classification.

Organizational Level

We further identified several organizational factors that shape decision making in comment moderation. First, an interesting distinction emerged when comparing the moderation approaches of interviewees working for commercial and public service media. Interviewees working for public service media emphasized their public mandate and, in this context, their efforts to reflect a broad range of positions in every public discussion on a topic. Accordingly, users and comments would only be blocked in extreme cases. One interviewee explained this attitude as follows:

> As a public broadcaster, it is really difficult to block users. People pay broadcasting fees, so we can't really do that much. (#8)

In contrast, moderators of commercial media referred to their right to decide about what happens on their website:

> Basically, I support the rights of the publisher, and I think we can publish what we want to. […] I consider it our right to intervene. This is our site and that also determines how we are perceived as a news site. (#6)

Second, our interviews indicated that the editorial line is relevant for comment moderation. The following quote shows that the political leaning of a medium is also mirrored in its discussion environment:

> There is no right-wing discourse here, because we don't allow it. It's different at other newspapers. I would say that the way in which the discourse is managed or the extent to which right-wing discourse is allowed reflects the attitude of a newspaper to a certain extent. (#9)

Third, the targeted audience seems to be a relevant factor. For example, an interviewee working for a news site with lots of opinion pieces said that this orientation is reflected in both the comment culture and the moderation style. We further observed that interviewees working for media with younger audiences noted that they use images, GIFs and humor in their moderation practice.

With respect to RQ3, our findings showed that respondents explicitly and implicitly referred to different explanatory factors when asked about their understanding of a hate comment and their moderation regarding these comments. Legislation appeared to be a very influential factor at the social institutions level. However, our findings indicate that the law interacts with other factors at the individual moderator, professional routines, and organization level. As a result, hate comments are handled within this legal framework in German newsrooms in different ways. Our study thus shows that these moderation factors can reinforce or mitigate, but also contradict each other, and can thus help explain the apparently inconsistent findings on moderation practices within media organizations (Boberg et al. 2018).

Discussion

The interview study presented in this paper was guided by three questions: Which types of user comments do comment moderators consider hate comments (RQ1)? How do they respond to these different types of hate comments (RQ2)? And how can we explain possible differences (RQ3)? Based on the theoretical background of moderation factors, drawing on the hierarchy of influences model (Shoemaker and Reese 2014), the findings of this study suggest that several factors at different levels can be expected to influence the comment moderators' perception of and response to hate comments. While there is broad agreement on extreme cases of hate speech such as death threats or discriminatory insults (which are in fact rare), there are disparities in comment moderation when it comes to less clear-cut cases such as irony or word plays. Accordingly, the biggest differences in comment moderation are found in this gray area.

Possible reasons for these differences can be attributed to the various levels of the moderation factors model introduced in Section 2. First of all, there seem to be individual differences in sensitivity to problematic content. We also found various working routines for moderating user comments, significantly shaped by the technological support provided by moderation software. Further, the political orientation, business model, and target group of newsrooms or media organizations seem to have an impact on moderation decisions. Finally, legislation, professional ethics, and the cultural-historical

context (e.g., in cases of Holocaust denial) seem to play an important role for comment moderators. Even though we exclusively interviewed staff from newsrooms in German-speaking countries, we found strong indicators for a decisive influence of these factors on the macro levels of social institutions and social systems.

In our view, these findings have important implications for future research:

With regard to our empirical findings, it is necessary to assume that comment moderators have different viewpoints on hate comments or hate speech. In some cases, they are even critical of these terms or reject them. Hate speech in user comments is a multifaceted and at times subtle phenomenon, which is why its identification and moderation demands high context sensitivity. These terms, however, are also used inconsistently in public debate and the literature. Accordingly, we would do well to account for possible differences in this regard, for example, when interviewing content moderators or when analyzing the outcome of their moderation decisions.

The question therefore remains of how we as researchers should deal with the prevalent terms and concepts in the future, especially when working on issues that concern or even affect moderation practices. For example, if we work with terms in interviews or surveys that are misunderstood or rejected by the respondents, the outcome of our study will be skewed. Moreover, when we work on algorithms to help newsrooms to better identify hate speech in user comments, this also impacts the work of comment moderators. As a consequence, it is incumbent upon us to not only refine our terms, but also to think about their usefulness in different contexts. Just because a concept is well suited for content analyses does not mean that the same is true for the moderation of user comments, and vice versa.

Moreover, we must acknowledge that a moderation decision about a user comment is not only dependent on its content, but that there are other factors that fundamentally influence moderation decisions. These include individual, professional, organizational, cultural, and systemic factors that can overlap, reinforce, but also weaken or even contradict each other. Thus, by introducing the moderation factors model as a larger theoretical framework, this paper aims to contribute to research on comment moderation beyond the scope of this case study. In our view, the model can serve to better structure future research in three ways: First, it helps to systematically map and relate previous and future studies and findings in the field of comment moderation, and to identify research gaps. Second, it does justice to the complexity of moderation decisions and can thus theoretically explain inconsistencies observed in previous studies (Boberg et al. 2018; Muddiman and Stroud 2017). Third, the model can help us to develop research questions and study designs that consider factors at different levels, empirically relate and compare them, and thus expand our knowledge of interaction effects in explaining comment moderation.

Finally, we need to reflect on the limitations of this study and its findings. To be sure, one especially relevant limitation is that we almost exclusively interviewed comment moderators working in German newsrooms. This means that not only were other national contexts with their respective peculiarities not taken into account, but the same is true for the profession of comment moderators who work outside journalistic newsrooms (e.g., in those cases where comment moderation has been outsourced; see Ferrucci and Wolfgang 2021). We therefore cannot rule out the possibility that there may be more relevant influential factors than the ones we identified with regard to comment moderation in the field

of journalism. Thus, even though we contend the model of moderation factors introduced here can generally be applied to all forms of comment moderation, we acknowledge that it may still have gaps.

Furthermore, as we focused on comment moderators in German newsrooms, we did not examine moderating factors at the macro levels of social institutions and social systems in more detail. Such an analysis would ideally involve cross-national studies in which the factors at these two levels vary. For this study, we only collected data on the individual level; we then systematically searched the material for indications of influences on the other levels, using the moderation factors model as a theoretical lens. In future research, these shortcomings could be circumvented by investigating possible influences at each level more specifically or, if possible, by collecting data at multiple levels using method triangulation. In our view, our model of moderation factors and the findings of this study, despite its limitations, provide a helpful framework and constructive clues for such further research.

Notes

1. The sample consists of 19 newsrooms from Germany and one from Austria. We decided to additionally interview staff members of *Der Standard*, after two other interviewees recommended to do so. The news website claims to be the first one worldwide to offer comment sections under each published article (Hinterleitner 2020).
2. To our knowledge, there has not yet been any systematic research on this topic, which is why we were not able to systematically control this aspect in the selection process for our interviews.
3. Since our study touches on sensitive topics, we decided not to name the selected media or our interview partners in the following. However, as we describe our sample in terms of organizational and individual characteristics, we do not believe that this anonymization is associated with limited comprehensibility of our findings.

Acknowledgments

We thank Martin Emmer and Joachim Trebbe for their contribution to the design of this study, Laura Laugwitz for the interview transcripts, and all students who were involved in conducting the interviews.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This research is part of the project "NOHATE—Overcoming crises in public communication about refugees, migration, foreigners," funded by the German Federal Ministry of Education and Research [grant number: 01UG1735AX].

ORCID

Sünje Paasch-Colberg http://orcid.org/0000-0002-0771-9646
Christian Strippel http://orcid.org/0000-0002-7465-4918

References

Bahador, B., and D. Kerchner. 2019. *Monitoring Hate Speech in the US Media*. Working Paper. https://mediapeaceproject.smpa.gwu.edu/report/.

Bakker, P. 2014. "Mr. Gates Returns. Curation, Community Management and Other New Roles for Journalists." *Journalism Studies* 15 (5): 596–606. doi:10.1080/1461670X.2014.901783.

Binns, A. 2017. "Fair Game? Journalists' Experiences of Online Abuse." *Journal of Applied Journalism & Media Studies* 6 (2): 183–206. doi:10.1386/ajms.6.2.183_1.

Bliuc, A.-M., N. Faulkner, A. Jakubowicz, and C. McGarty. 2018. "Online Networks of Racial Hate: A Systematic Review of 10 Years of Research on Cyber-Racism." *Computers in Human Behavior* 87: 75–86. doi:10.1016/j.chb.2018.05.026.

Boberg, S., T. Schatto-Eckrodt, L. Frischlich, and T. Quandt. 2018. "The Moral Gatekeeper? Moderation and Deletion of User-Generated Content in a Leading News Forum." *Media and Communication* 6 (4): 58–69. doi:10.17645/mac.v6i4.1493.

Braun, J., and T. Gillespie. 2011. "Hosting the Public Discource, Hosting the Public. When Online News and Social Media Converge." *Journalism Practice* 5 (4): 383–398. doi:10.1080/17512786.2011.557560.

Brown, A. 2017. "What is Hate Speech? Part 1: The Myth of Hate." *Law and Philosophy* 36 (4): 419–468. doi:10.1007/s10982-017-9297-1.

Brugger, W. 2003a. "The Treatment of Hate Speech in German Constitutional Law (Part I)." *German Law Journal* 4 (1): 1–22. doi:10.1017/S2071832200015716.

Brugger, W. 2003b. "The Treatment of Hate Speech in German Constitutional Law (Part II)." *German Law Journal* 4 (1): 23–44. doi:10.1017/S2071832200015728.

Chen, G. M., A. Muddiman, T. Wilner, E. Pariser, and N. J. Stroud. 2019. "We Should Not Get Rid of Incivility Online." *Social Media + Society* 5 (3): 1–5. doi:10.1177/2056305119862641.

Chen, G. M., and P. Pain. 2017. "Normalizing Online Comments." *Journalism Practice* 11 (7): 876–892. doi:10.1080/17512786.2016.1205954.

Chen, G. M., P. Pain, V. Y. Chen, M. Mekelburg, N. Springer, and F. Troger. 2020. "'You Really Have to Have a Thick Skin': A Cross-Cultural Perspective on How Online Harassment Influences Female Journalists." *Journalism* 21 (5): 877–895. doi:10.1177/1464884918768500.

Coe, K., K. Kenski, and S. A. Rains. 2014. "Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments." *Journal of Communication* 64 (4): 658–679. doi:10.1111/jcom.12104.

Domingo, D. 2011. "Managing Audience Participation. Practices, Workflows and Strategies." In *Participatory Journalism: Guarding Open Gates at Online Newspapers*, edited by J. B. Singer, D. Domingo, A. Heinonen, A. Hermida, S. Paulussen, T. Quandt, Z. Reich, and M. Vujnovic, 76–95. Oxford: Wiley-Blackwell.

Ferrucci, P., and J. Wolfgang. 2021. "Inside Or Out? Perceptions of How Differing Types of Comment Moderation Impact Practice." *Journalism Studies* 22 (8): 1010–1027. doi:10.1080/1461670X.2021.1913628.

Fortuna, P., and S. Nunes. 2018. "A Survey on Automatic Detection of Hate Speech in Text." *ACM Computing Surveys* 51 (4): 1–30. doi:10.1145/3232676.

Frischlich, L., S. Boberg, and T. Quandt. 2019. "Comment Sections as Tagrets of Dark Participation? Journalists' Evaluation and Moderation of Deviant User Comments." *Journalism Studies* 20 (14): 2014–2033. doi:10.1080/1461670X.2018.1556320.

German Press Council. 2017. "German Press Code. Guidelines for Journalistic Work as Recommended by the German Press Council – Complaints Procedure." https://www.presserat.de/en.html?file=files/presserat/dokumente/download/Press%20Code.pdf.

Harlow, S. 2015. "Story-Chatterers Stirring Up Hate: Racist Discourse in Reader Comments on U.S. Newspaper Websites." *Howard Journal of Communications* 26 (1): 21–42. doi:10.1080/10646175.2014.984795.

Hinterleitner, G. 2020. *Weit mehr als eine nerdige Idee: Wie DER STANDARD vor 25 Jahren ins Internet fand* [Much More Than Just a Nerdy Idea: How DER STANDARD Found Its Way Into the Internet 25

Years Ago]. https://www.derstandard.de/story/2000114009456/weit-mehr-als-eine-nerdige-ideewie-der-standard-vor-25.
Hosseini, H., S. Kannan, B. Zhang, and R. Poovendran. 2017. "Deceiving Google's Perspective API built for detecting toxic comments." https://arxiv.org/abs/1702.08138.
Jigsaw. 2021. "Perspective: What If Technology Could Help Improve Conversations Online?" https://www.perspectiveapi.com.
Keipi, T., M. Näsi, A. Oksanen, and P. Räsänen. 2017. *Online Hate and Harmful Content. Cross-National Perspectives*. London: Routledge.
Loosen, W., and J.-H. Schmidt. 2012. "(Re-)Discovering the Audience: The Relationship Between Journalism and Audience in Networked Digital Media." *Information, Communication & Society* 15 (6): 867–887. doi:10.1080/1369118X.2012.665467.
Matsuda, M. J. 1989. "Public Response to Racist Speech: Considering the Victim's Story." *Michigan Law Review* 87 (8): 2320–2381. doi:10.2307/1289306.
Muddiman, A., and N. J. Stroud. 2017. "News Values, Cognitive Biases, and Partisan Incivility in Comment Sections." *Journal of Communication* 67 (4): 586–609. doi:10.1111/jcom.12312.
Paasch-Colberg, S., C. Strippel, L. Laugwitz, M. Emmer, and J. Trebbe. 2020. "Moderationsfaktoren: ein Ansatz zur Analyse von Selektionsentscheidungen im Community Management." [Moderation Factors: An Approach to the Analysis of Selection Decisions in Community Management]. In *Integration durch Kommunikation. Jahrbuch der Deutschen Gesellschaft für Publizistik- und Kommunikationswissenschaft 2019*, edited by V. Gehrau, A. Waldherr, and A. Scholl, 109–119. doi:10.21241/ssoar.67858.
Papacharissi, Z. 2004. "Democracy Online: Civility, Politeness, and the Democratic Potential of Online Political Discussion Groups." *New Media & Society* 6 (2): 259–283. doi:10.1177/1461444804041444.
Parekh, B. 2006. "Hate Speech: Is There a Case for Banning?" *Public Policy Research* 12 (4): 213–223. doi:10.1017/CBO9781139042871.006.
Pöyhtäri, R. 2014. "Limits of Hate Speech and Freedom of Speech on Moderated News Websites in Finland, Sweden, The Netherlands and the UK." *Annales. Series historia et sociologia Izhaja Štirikrat Letno* 24 (3): 513–524.
Quandt, T. 2018. "Dark Participation." *Media and Communication* 6 (4): 36–48. doi:10.17645/mac.v6i4.1519.
Reich, Z. 2011. "User Comments: The Transformation of Participatory Space." In *Participatory Journalism: Guarding Open Gates at Online Newspapers*, edited by J. B. Singer, D. Domingo, A. Heinonen, A. Hermida, S. Paulussen, T. Quandt, Z. Reich, and M. Vujnovic, 96–117. Oxford: Wiley-Blackwell.
Ross, B., M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. 2016. "Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis." *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication (Bochum)*, 6–9. doi:10.17185/duepublico/42132.
Saleem, H. M., K. P. Dillon, S. Benesch, and D. Ruths. 2017. "A Web of Hate: Tackling Hateful Speech in Online Social Spaces." https://arxiv.org/abs/1709.10159.
Schreier, M. 2014. "Qualitative Content Analysis." In *The SAGE Handbook of Qualitative Data Analysis*, edited by U. Flick, 1–19. Sage.
Schulz, W., T. Held, S. Dreyer, and T. Wind. 2008. "Regulation of Broadcasting and Internet Services in Germany: A Brief Overview." doi:10.21241/ssoar.71697.
Sellars, A. F. 2016. *Defining Hate Speech*. Research Publication No. 2016–20. Berkman Klein Center. doi:10.2139/ssrn.2882244.
Shoemaker, P., and S. Reese. 2014. *Mediating the Message in the 21st Century. A Media Sociology Perspective*. 3rd ed. Routledge.
Springer, N., I. Engelmann, and C. Pfaffinger. 2015. "User Comments: Motives and Inhibitors to Write and Read." *Information, Communication & Society* 18 (7): 798–815. doi:10.1080/1369118X.2014.997268.
Strippel, C., S. Paasch-Colberg, M. Emmer, and J. Trebbe. 2021. "Guidance: User Comments and Hate Speech. Expert Interviews With Comment Moderators of German-Language Online News Services." https://osf.io/6uwyh/.

Stryker, R., B. A. Conway, and J. T. Danielson. 2016. "What is Political Incivility?" *Communication Monographs* 83 (4): 535–556. doi:10.1080/03637751.2016.1201207.

Su, L. Y.-F., M. A. Xenos, K. M. Rose, C. Wirz, D. A. Scheufele, and D. Brossard. 2018. "Uncivil and Personal? Comparing Patterns of Incivility in Comments on the Facebook Pages of News Outlets." *New Media & Society* 20 (10): 3678–3699. doi:10.1177/1461444818757205.

Tworek, H., and P. Leerssen. 2019. *An Analysis of Germany's NetzDG Law*. Working Paper. https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/NetzDG_TWG_Tworek_April_2019.pdf.

United Nations. 2020. "United Nations Strategy and Plan of Action on Hate Speech." https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20PoA%20on%20Hate%20Speech_Guidance%20on%20Addressing%20in%20field.pdf.

van Aken, B., J. Risch, R. Krestel, and A. Löser. 2018. "Challenges for Toxic Comment Classification: An In-Depth Error Analysis." Paper Presented at the Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). https://arxiv.org/abs/1809.07572.

Wintterlin, F., T. Schatto-Eckrodt, L. Frischlich, S. Boberg, and T. Quandt. 2020. "How to Cope With Dark Participation: Moderation Practices in German Newsrooms." *Digital Journalism* 8: 904–924. doi:10.1080/21670811.2020.1797519.

Xiang, G., B. Fan, L. Wang, J. I. Hong, and C. P. Rose. 2012. "Detecting Offensive Tweets via Topical Feature Discovery Over a Large Scale Twitter Corpus." *CIKM '12 – Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 1980–1984. doi:10.1145/2396761.2398556.

Ziegele, M. 2016. *Nutzerkommentare als Anschlusskommunikation. Theorie und qualitative Analyse des Diskussionswerts von Online-Nachrichten*. Wiesbaden: Springer VS.

Ziegele, M., and P. B. Jost. 2020. "Not Funny? The Effects of Factual Versus Sarcastic Journalistic Responses to Uncivil User Comments." *Communication Research* 47 (6): 891–920. doi:10.1177/0093650216671854.

Ziegele, M., P. Jost, M. Bormann, and D. Heinbach. 2018. "Journalistic Counter-Voices in Comment Sections: Patterns, Determinants, and Potential Consequences of Interactive Moderation of Uncivil User Comments." *Studies in Communication and Media* 7 (4): 525–554. doi:10.5771/2192-4007-2018-4-525.

Ziegele, M., M. Weber, O. Quiring, and T. Breiner. 2018. "The Dynamics of Online News Discussions: Effects of News Articles and Reader Comments on Users' Involvement, Willingness to Participate, and the Civility of Their Contributions." *Information, Communication & Society* 21 (10): 1419–1435. doi:10.1080/1369118X.2017.1324505.