



# Moderating the Good, the Bad, and the Hateful: Moderators' Attitudes Towards ML-based Comment Moderation Support Systems

Holger Koelmann<sup>1</sup>(✉) , Kilian Müller<sup>1</sup> , Marco Niemann<sup>1</sup> ,  
and Dennis M. Riehle<sup>2</sup>

<sup>1</sup> University of Münster - ERCIS, Münster, Germany

{holger.koelmann,kilian.mueller,marco.niemann}@ercis.uni-muenster.de

<sup>2</sup> University of Koblenz-Landau, Koblenz, Germany

riehle@uni-koblenz.de

**Abstract.** Comment sections have established themselves as essential elements of the public discourse. However, they put considerable pressure on the hosting organizations to keep them clean of hateful and abusive comments. This is necessary to prevent violating legal regulations and to avoid appalling their readers. With commenting being a typically free feature and anonymity encouraging increasingly daunting comments, many newspapers struggle to operate economically viable comment sections. Hence, throughout the last decade, researchers set forth to develop machine learning (ML) models to automate this work. With increasingly sophisticated algorithms, research is starting on comment moderation support systems that integrate ML models to relieve moderators from parts of their workload. Our research sets forth to assess the attitudes of moderators towards such systems to provide guidance for future developments. This paper presents the findings from three conducted expert interviews, which also included tool usage observations.

**Keywords:** Community management · Machine learning · Content moderation · Comment moderation support system · Digital work

## 1 Introduction

You are a community manager for the online presence of a large newspaper. Your area of work should include tasks like user engagement, participating in discussions, and researching the validity of discussion points. However, you are often tasked with reviewing and deciding on publishing or withholding user-generated content. How did we get to this point?

Social media in general and user comments in particular have seen a remarkable rise over the last decades. While formerly, letters to the editors of newspapers were a comparatively rare phenomenon, nowadays, readers can engage with

the editorial team much easier on the website, using discussion fora and comment sections [19, 28, 31]. While bad behavior on the internet has a long tradition [4], the number of comments that readers, i.e., users visiting the newspaper's website, publish has increased massively (e.g., the New York Times received about 9,000 comments per day in 2015 [16]), leading to an increased reader engagement and yields positive effects on the economic sustainability of newspaper organizations [19].

With the rise of comments, the amount of problematic comments has risen as well. The digital feel-good society has—in parts—turned into a more unjoyful society, where hate, incitement, oppression, and discrimination exist. Hate speech [37, 41, 53] and other malicious content, like misinformation [30, 60] or cyberbullying [29, 61], are disrupting the online discourse and keep increasing in volume [58]. Studies estimate abusive user-generated content somewhere between 2% and 80% [3, 10, 19, 25, 35, 42], which is an inaccurate estimate but shows that the problem of critical user comments exists and is relevant. As hateful or insulting comments can cause legal consequences [44], newspapers tend to and are, to a certain extent, forced to keep these toxic comments off their platforms by moderating incoming comments [2].

A moderation process, in whatever form it might take place, requires resources, most often human resources; as a consequence, newspapers locked their comment sections either for highly debated topics or, in some cases, even for all topics [3, 27, 44, 56]. Therefore, community managers, who normally would foster a healthy discussion culture and interact with the readers, are forced to fill the roles of content cleaners [3]. The only other option would be to close the discussion sections completely, an action some newspapers have already taken [9, 32, 38]. Thus, community managers are in need of support.

One idea to ease the burden of community managers, allowing them to perform their indented tasks, is the use of machine learning (ML) embedded in (semi-)automated comment moderation support systems (CMSS) [21, 23, 27]. These systems can either act as a decision support system (DSS) flagging comments for the moderators to review or as an automated system that accepts or rejects comments. The (automated) moderation of user-generated comments is a complex issue, as aside from legal restrictions of publishable comments, most free, democratic states also guarantee “freedom of speech”; therefore, these systems have to be handled with care; both by the developers as well as by the community managers [23]. As the community managers are the ones already walking this narrow ridge, we want to find out how to best support them in their current work. Are CMSS already a viable option for moderators, and if not, what factors could lead to a possible adoption of CMSS to support moderators in their daily business?

Therefore, with this study, we aim to provide insights into the general possibility of (semi-)automated content moderation, how community managers have already adopted it, and which requirements still need to be fulfilled to reach a reasonable amount of productivity. Therefore, we aim to answer our main research question **RQ**: *What are the attitudes of community managers towards (semi-)automated comment moderation support systems?* Subsequently, we address the

following sub-research questions: **SRQ-1**: *What is the current state of digital comment moderation, and which systems are already utilized by community managers?*, **SRQ-2**: *How well do the requirements for CMSS derived from literature reflect the needs of community managers?*, and **SRQ-3**: *Which factors may lead to the adoption or non-adoption of (semi-)automated comment moderation support systems by community managers of news outlets?*

To tackle these research questions, we interviewed community managers from different newspapers. Therefore, we conducted a series of interviews with representatives from the community management field of some of the major newspapers in Germany. Within these interviews, we are evaluating community managers' experiences with CMSS, how such systems might already be used in practice, and which improvements have to be made in the future. In order to construct an interview guideline that encompasses the relevant aspects, we reviewed the already existing literature. An excerpt of our findings from the literature is presented in Sect. 2. Section 3 details our research approach while Sect. 4 presents the results from our interviews. Finally, Sect. 5 concludes the paper.

## 2 Theoretical Background

Regarding the comment moderation process, two scenarios are possible: A *post-moderation* refers to a process where each received comment is published. If the newspaper receives complaints about a comment or if the comment is reported, the comment will be read by a moderator, evaluated, and, if necessary, blocked or deleted. In contrast, a *pre-moderation* refers to a process where each comment is moderated before it is published online. While both approaches have pros and cons, the most notable downside of a post-moderation is that newspapers might run into legal issues with criminal comments published, while for pre-moderation, a large amount of human resources is required [44, 46].

A solution that is suitable both from an economic (less human resources required), as well as legal (do not publish hate comments) perspective can be the inclusion of ML as part of the moderation process [47]. Here, comment moderation can be seen as a two-stepped process. On the first level, an ML algorithm scans incoming comments and puts them into different categories like critical and uncritical. On a second level, only a subset of all comments is moderated by humans. For instance, only comments in the category "critical" might be manually moderated, while comments in the category "uncritical" could be published immediately without human interaction.

Abusive speech and the challenges of moderation endeavors are a long-standing problem—one that in Europe, and especially in Germany, reached a culmination point in the wake of the refugee crisis in 2015/2016 [27, 48]. Up to 50% of the newspapers in Germany decided to give up their comment sections back then [54]—a trend that could also be observed elsewhere [14, 32, 45]. However, the last years indicate that comment sections are not dead and still desirable for newspapers [15]. Evidence for this assumption can be found in recent studies indicating that a substantial number of newspapers are still operating

comment sections—especially the small and medium-sized ones [40]. Furthermore, solutions such as the Coral Talk project are still actively maintained and developed, indicating a need for specialized commenting software [11, 26]. Beyond these approaches, newspapers added further protective measures such as upfront registrations and rule sets [40].

So far, all these adjustments are still based primarily on human efforts. In addition, academia and practice started to work on automation more than a decade ago [41, 62]. Till today no conclusive approach to automatically detect abusive language has been found [18, 24, 34, 51, 63]. However, the same research indicates that considerable progress is made on this classification task. With the increased efficacy of ML algorithms in detecting abusive language, initial research is coming up that seeks to design systems to integrate such algorithms in a moderation interface [7, 33]. While the approaches differ in the specific goals and pathways, they all aim to leverage the ML capabilities to reduce the community managers' and moderators' workload—not to replace them. Some follow a more punitive strategy [7], whereas others aim at rewarding creators of quality comments [33, 43]. One consistent notion is the goal to provide decision support instead of decisions—sometimes phrased implicitly [33] and sometimes explicitly [7]. This does not necessarily preclude partial automation [7]; however, it typically then connects back to the decision support idea by giving moderators the power for corrective actions, which can be *used as learning input for future algorithm generations* [33].

Given the complexity of many modern ML approaches as well as the *typical black box nature of their decision processes* [50], and with most community managers being no ML experts, transparency is becoming an important aspect [7]. This aligns with legal regulations that recommend—partially even mandate—being able to explain automated (or semi-automated) decisions [17, 22, 57]. Despite the increased focus on ML tools for comment moderation, support systems are often conceptualized to include “non-ML” capabilities that should provide additional guidance to users. This can range from the provision of statistical information (*e.g., comments written by a user; rejections vs. accepts per user; ...*) to the provision of context information or simple static analysis (*e.g., comment length, ...*) [33]. Another line of thought that has been brought forth recently is the notion of community management as a collaborative effort that requires the provision of tooling to support the corresponding workflows (*e.g., assigning problematic comments to other moderators; reviewing prior decisions; ...*) [40]. However, to date, most of this is conceptual work enhanced only by mockups and wireframes [7, 33], with only a few prototypical systems being under development (*e.g., [47]*).

Therefore, little research exists about the acceptance of such technologies by moderators or community managers. Articles by Brunk et al. [6] and Bunde [7] analyze trust and transparency using a mocked CMSS, with limited capabilities. Nevertheless, trust and technology acceptance are core concerns regarding the adoption of such information systems in practice [36], with trust in, *e.g., the system provider* [55] needed to overcome the perceived risk by its user [12, 36].

Empirical quantitative research often refers to the technology acceptance model (TAM) [13] and the unified theory of acceptance and use of technology (UTAUT) [59] to measure the acceptance of a given technology, with some approaches including the users' trust in the technology as an essential factor for a users' acceptance of given technology [20]. To gain some deeper insights into the critical aspect of the acceptance of CMSS by our expert interview partners, we are also addressing their thought processes and attitudes towards CMSS within this study as well.

### 3 Research Approach

Building upon this background, we conducted semi-structured expert interviews to generate insights for our research questions. These allowed us to find in-depth answers from practitioners in the field about their opinions, work behavior, and attitudes to the discussed developments and systems. For these reasons, and the limited number of experts available, semi-structured interviews seem to be the most suitable method for data collection [5, 49]. Also, it allows us to have the interviewee test a working prototype of a CMSS designed to address the issues and requirements brought up in Sect. 2. This approach further allows us to ask follow-up questions when needed [5, 49]. Since transparency in the research process is essential for qualitative research [1, 8], we will go deeper into the structure of the interview guideline, the sampling process for the study in progress, and our interview setting:

In search of relevant questions for the interviews, we followed the suggestions by Rowley [49]. We generated potential questions inductively based on our theoretical background, reworked them to fit the practice case, and reduced them to the most fitting ones for the interview guideline. We structured the interview guideline according to the following three main question sets, which we derived from the existing theoretical background described in Sect. 2. After a short self-introduction of the interviewer and interviewee, the first question set covers general information about the current moderation process and used system. In addition, the fundamental attitudes toward objectivity and transparency in the moderation process and the interviewee's attitudes towards ML-based automation are covered. After the first set of questions, a functioning prototype of an ML-supported CMSS<sup>1</sup> is introduced, which the interviewee is asked to use to moderate a predefined set of comments. Based on this experience, the requirements for such systems [39] are discussed in the second set of questions, covering the five different aspects of team moderation, interpretability and transparency, control and correction of ML-based decisions, decision support beyond ML, and the openness of the system. The third set then includes questions regarding the acceptance of such systems, such as the potential intention to use such a system, its trustworthiness, the risks of using such a system, as well as the potential influence of system use on the organization. Finally, the last questions following the three main question sets are designed for an open discussion about issues

<sup>1</sup> <https://www.moderat.nrw>.

left untouched, which are important to the interviewee, and some demographic questions for further context to the statements made during the interview. An overview of the discussed topics and an exemplary question, representative of each topic, can be found in Table 1 to give an impression of the discussed content.

To find experts in the field, we have invited 24 representatives of different German news outlets to participate in our study. We received six responses from interested experts, to which we explained the planned interview procedure. We were then able to schedule three virtual interviews. The interviewed experts all operated in the field of comment moderation and were employed by one medium-sized and two of the largest national news outlets in Germany. We asked questions from the interview guideline, gave the participants an introduction to an existing prototype with the opportunity to test it, and ended with an open discussion about topics important to the respective interviewee. Each interview took between 44 and 103 min, depending on the intensity the interviews used the open discussion elements for further in-depth statements. The interviews were conducted between September and October of 2021, resulting in 214 min of recordings. The recorded interviews were transcribed by a professional transcription service<sup>2</sup>. The transcripts were then fine-tuned by the researchers where necessary and analyzed afterwards. The main findings of these interviews will be reported in the next section.

## 4 Findings

To contextualize the findings, we want to give a brief overview of our interviewee demographics: Their age spanned from 29 to 56 years, and all worked directly related to community management in their respective news outlets. They all finished their school education (Abitur; A-level equivalent), and two of the interviewees had additional Master's degrees. All of them already have several years of work experience in the field and have stayed in their respective companies for multiple years already. Despite their elevated positions, they are all still in touch with hands-on community management; however, they also have a more managerial perspective allowing them to give strategic insights beyond the mere operative ones.

In the upcoming subsections, we present the most interesting and striking findings from these interviews along with each main question set of the interview guideline. The results of the analysis are then followed by a concluding discussion in Sect. 5.

### 4.1 Current Process, System, and Attitudes

We interviewed all participants regarding their current moderation process, used software systems and support tools, and their attitudes towards different aspects of the moderation process.

---

<sup>2</sup> <https://sonix.ai>.

**Table 1.** Structure of the used interview guideline, incl. topics and representative questions

No.	Topic	Representative question (translated from German)
1. Current process, system, and attitudes		
I.	Information about the current moderation process and system setup	“Which systems do you currently use to manage your content and especially your comments?”
II.	Community management	“Which type of moderation to you currently use, pre-, post-, or mixed-moderation?”
III.	Objectivity and transparency in moderation	“Do you provide feedback about the reason for blocking a comment to the affected commentator?”
IV.	Attitude towards ML-based automation	“Would you be willing to include ML in the moderation process in general?”
—Introduction to the prototype, incl. user test—		
2. Requirements of comment moderation support systems		
V.a	Team moderation	“To which degree does the presented system cover your needs in terms of team support?”
V.b	Interpretability and transparency	“Do you miss some information you would need to understand the ML-based decision?”
V.c	Control and correction of ML-based decisions	“Can you imagine to use the decisions of the presented system alone (maybe without the ability to correct the decision)?”
V.d	Decision support beyond ML	“How useful is the presented additional commenter’s information?”
V.e	Openness of the system	“Did you encounter any issues with the adaptability of the content moderation systems you have used so far in your career?”
3. Acceptance of ML-based comment moderation		
VI.a	Acceptance and intention to use	“Can you imagine that your organization would adopt such a system into its regular operations?”
VI.b	Trustworthiness of the system	“Do you consider the moderation of comments with such a system to be trustworthy?”
VI.c	Perceived risk of system Use	“Do you see risks for your work as a community manager through the use of such a system?”
VI.d	Influence on organization	“Do you see risks for your organization through the use of such a system?”
4. Closing remarks		
VII.	Open issues & discussion	“Would you like to discuss additional points with us, which have not yet been properly covered?”
VIII.	Demographics	“What is your highest level or education or academic degree?”

Noticeably, all three newspapers utilized shift operation to ensure the largest amount of moderation-coverage possible. Another common characteristic across the three outlets is the use of pre-moderation (No. II from Table 1). While one newspaper previously utilized post-moderation, they recognized a significant increase in critical comments and, thus, chose to switch their content moderation to pre-moderation. Handling rejected comments and informing commentators differed between organizations. In one case, the commentator does not receive any information, in the other cases, the commentator receives some basic information via mail (III). Additionally, while the largest newspaper was able to keep its comment section open during the night, the other two newspapers are either closing their respective comment sections during these hours or are not publishing comments until the morning, as they are not able to muster the necessary workforce to moderate comments during nighttime.

In terms of IT systems, every newspaper utilized a different system (I). While the largest newspaper has a custom solution, the other two use various applications, e.g., content management systems (CMS), with varying functions for different channels. These systems were not integrated, i.e. workers had to switch between different systems.

All newspapers agreed that user comments are critical to their online presence. Firstly, to increase customer loyalty and secondly, to increase the generated traffic (which again impacts the outlets' visibility on search engines). Basically, every interviewee could envision themselves utilizing ML-assisted content moderation to some extent (IV). However, the expected level of involvement and the degree of automation differed from newspaper to newspaper. While the largest newspaper could envision itself at some point utilizing fully automated content moderation, all newspapers agreed that the human should be kept in the loop (at least in the beginning).

## 4.2 Requirements of Comment Moderation Support Systems

The findings regarding the requirements' fulfillment (respectively their adequacy) for CMSS come with the limitation of the prototypical nature of the system used for demonstration and experimentation in the interviews. This entails that certain elements are not yet implemented in a way that satisfies commercial requirements for UI/UX. Nevertheless, the interviews revealed several interesting findings: First, they indicate that none of the newspapers aims for full-fledged automation at the current point in time (V.c). All seek automated support to reduce workloads but only see any potential for automation after a more extended evaluation and experimentation period. Even though not much was discussed in the literature, the interviews confirmed the importance of collaborative features in a CMSS, with ML support present (V.a). Feedback ranged from the positive acknowledgment of useful features such as assigning comments for decision and review to other community managers to requests for additional functionality (e.g., real-time exchange of moderation decisions between software clients). Furthermore, all interviewees agreed upon the necessity and helpfulness of explanations for machine decisions (V.b). While highlighting individual



words with colors is uniformly considered appropriate, opportunities for further research appear to emerge for the exact configuration. One interviewee pointed out that having problematic (or anti-problematic) passages highlighted is helpful for all cases, the other two interviewees would prefer more restricted support. However, the latter were differing in their preferences of the exact configuration (e.g., only highlighting passages in selected comments vs. reducing the quantity of highlighted words) (V.d). Lastly, interviewees pointed toward existing software solutions into which they would like to integrate a CMSS (V.e). Thus, CMSS should be designed to both be able to attach to existing moderation software and/or to be able to integrate other solutions.

While this is not a complete assessment of the requirements—or the linked systems—, the results illustrate that such systems are needed and that the extant ideas are suitable guiding principles. However, the interviews also highlight the need for additional research and evaluation with practitioners, as such systems are complex and affect a critical aspect of journalism—the thin line between the right and necessity of free speech and the need for moderation to satisfy legal requirements.

### 4.3 Acceptance of ML-Based Comment Moderation

Regarding their acceptance factors of such systems (VI.a), it is noteworthy that all participants were interested in using such an ML-based CMSS with one organization already using a similar system in their routines. Though for one interviewee, it remained important that the system would only work as a DSS and not as a fully automated system, further stretching the importance of the human-in-the-loop approach.

The interview section about the system's trustworthiness (VI.b) also reflected the importance of the human-in-the-loop approach since the use of such a system was considered trustworthy due to its emotionless nature but not necessarily as fair without the emotional context.

This was also considered a potential risk for the community managers' work (VI.c): Their focus on the commenter in the comments section might get lost. Another identified concern is the sensitivity of the system. As with every technical test, it might miss hateful comments leading to their display on site—a condition that is perceived less likely using tight human moderation.


Another remarkable aspect of the findings is the expected impact on the trustworthiness and risks in terms of the news outlet's perception in its community (VI.d). Here, the interviewee from the mid-sized outlet stretched the potential loss of focus on the community through less engagement with the critical material as well. For them their claim for direct interaction and discussion with their smaller community is key. Machine-based moderation could harm that claim and the perceived trustworthiness of the outlet. The other two interviewees from larger outlets saw this rather differently. They currently cannot engage with their respective communities as personally as they would like to, also due to the number of critical comments that need moderation. For them, additional automation of the comment moderation process would free time they

could spend on engaging directly with the community, which could, in turn, positively affect their perception in the eyes of their community.

## 5 Concluding Discussion

We set forth to find out more about the current state, the requirements, and the potential acceptance for the use of CMSS. Our results show various similarities and dissimilarities between the current state of the art, requirements, and adoption criteria of different newspaper outlets. For practice, our work provides insights into the current state of CMSS and their use, such as the prevalence of pre-moderation in shifts. Furthermore, adoption criteria defined by community managers can be utilized by CMSS designers to either improve their current systems or guide the development phase from the start. It became clear that the human-in-the-loop approach is preferred for the time being and that CMSS should be developed as automated support for the human community managers, relieving them from some of the burdens of their work to concentrate on their core task: engaging with their community.

Besides these valuable contributions, the study also comes with limitations. The obvious one is the limited amount of available data since only three expert interviews have been conducted. However, we argue that these three interviewees are a good representation of the market, coming from three different newspapers, with both medium-sized and large organizations being represented. Further, the interviews are only representing the German market of news outlets and the findings should therefore be interpreted in light of this cultural and regulatory context. Future research should therefore look deeper into other contexts to see, if the same results appear internationally or if differences in culture and regulation play a larger role in the acceptance of ML-based CMSS. In addition, future quantitative research could dive further into the linkages of the involved factors for accepting and adopting these systems in the community moderator's work. On that note, further research into how the use of such CMSS is perceived by the affected commenters and how the systems need to give feedback to the commenters for a higher rate of acceptance needs to be done. Besides this practitioner group, it would also be possible to conduct interviews with other academic experts in the field of journalism and hate speech prevention as well, to get a less practice-based view of the topic. Lastly, interviewing potential commentators of the general public could be worthwhile as they are affected by decisions of these systems directly during their engagement in online discussions or indirectly by the resulting shift in online debating culture [3] and the potential distortion of their freedom of expression [52].

**Acknowledgments.** The research leading to these results received funding from the federal state of North Rhine-Westphalia and the European Regional Development Fund (EFRE.NRW 2014–2020), Project: **M** **DERAT!** (No. CM-2-2-036a).

## References

1. Aguinis, H., Solarino, A.M.: Transparency and replicability in qualitative research: the case of interviews with elite informants. *Strateg. Manag. J.* **40**(8), 1291–1315 (2019)
2. Bloch-Wehba, H.: Automation in moderation. *Cornell Int. Law J.* **53**(1), 41–96 (2020)
3. Boberg, S., Schatto-Eckrodt, T., Frischlich, L., Quandt, T.: The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media Commun.* **6**(4), 58–69 (2018)
4. Brail, S.: The price of admission: Harassment and free speech in the wild, wild west. *Wired.Women: Gender and new realities in cyberspace* (1996)
5. Brinkmann, S.: *Qualitative Interviewing*. Oxford University Press, United Kingdom (2013)
6. Brunk, J., Mattern, J., Riehle, D.M.: Effect of transparency and trust on acceptance of automatic online comment moderation systems. In: Becker, J., Novikov, D. (eds.) *21st IEEE Conference on Business Informatics*, pp. 429–435. Russia, Moscow (2019)
7. Bunde, E.: AI-assisted and explainable hate speech detection for social media moderators - a design science approach. In: *Proceedings of the 54th Hawaii International Conference on System Sciences*, pp. 1264–1273. HICSS 2021, ScholarSpace, Kauai, HI, USA (2021)
8. Burton-Jones, A., Boh, W.F., Oborn, E., Padmanabhan, B.: Editor's comments: advancing research transparency at MIS Quarterly: a pluralistic approach. *Manag. Inf. Syst. Q.* **45**(2), 3–8 (2021)
9. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, pp. 71–80. SOCIALCOM-PASSAT 2012, IEEE, Amsterdam, Netherlands (2012)
10. Cheng, J.: Report: 80 percent of blogs contain “offensive” content (2007). <https://arstechnica.com/information-technology/2007/04/report-80-percent-of-blogs-contain-offensive-content/>
11. Coral Project: Coral by Vox Media (2021). <https://coralproject.net/>
12. Das, T., Teng, B.S.: The risk-based view of trust: a conceptual framework. *J. Bus. Psychol.* **19**(1), 85–116 (2004)
13. Davis, F.: *A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results*. Ph.D. thesis, Massachusetts Institute of Technology, Massachusetts (1985)
14. Ellis, J.: What happened after 7 news sites got rid of reader comments (2015). <https://www.niemanlab.org/2015/09/what-happened-after-7-news-sites-got-rid-of-reader-comments/>
15. Engelke, K.M.: Enriching the conversation: audience perspectives on the deliberative nature and potential of user comments for news media. *Digit. J.* **8**(4), 447–466 (2020)
16. Etim, B.: The Most Popular Reader Comments on The Times (2015). <https://www.nytimes.com/2015/11/23/insider/the-most-popular-reader-comments-on-the-times.html>
17. Felzmann, H., Villaronga, E.F., Lutz, C., Tamò-Larrieux, A.: Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data Soc.* **6**(1), 1–14 (2019)

18. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. *ACM Comput. Surv.* **51**(4), 1–30 (2018)
19. Gardiner, B., Mansfield, M., Anderson, I., Holder, J., Louter, D., Ulmanu, M.: The dark side of Guardian comments (2016). <https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments>
20. Gefen, D., Karahanna, E., Straub, D.W.: Trust and tam in online shopping: an integrated model. *MIS Q.* **27**(1), 51–90 (2003)
21. Gillespie, T.: Content moderation, AI, and the question of scale. *Big Data Soc.* **7**(2), 1–5 (2020)
22. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision making and a “Right to Explanation”. *AI Mag.* **38**(3), 50–57 (2017)
23. Gorwa, R., Binns, R., Katzenbach, C.: Algorithmic content moderation: technical and political challenges in the automation of platform governance. *Big Data Soc.* **7**(1), 1–15 (2020)
24. Herodotou, H., Chatzakou, D., Kourtellis, N.: Catching them red-handed: Real-time aggression detection on social media. In: 2021 IEEE 37th International Conference on Data Engineering (ICDE), pp. 2123–2128 (2021)
25. Hine, G.E., et al.: Kek, cucks, and god emperor trump: a measurement study of 4chan’s politically incorrect forum and its effects on the web. In: Proceedings of the Eleventh International AAAI Conference on Web and Social Media, pp. 92–101. ICWSM-2017, Montreal, Canada (2017)
26. Kim, J.: Moderating the uncontrollable. *Intersect. Stanford J. Sci. Technol. Soc.* **10**(3), 1–9 (2017)
27. Köffer, S., Riehle, D.M., Höhenberger, S., Becker, J.: Discussing the value of automatic hate speech detection in online debates. In: Drews, P., Funk, B., Niemeyer, P., Xie, L. (eds.) *Tagungsband Multikonferenz Wirtschaftsinformatik 2018. MKWI 2018*, Leuphana Universität, Lüneburg, Germany (2018)
28. Kolhatkar, V., Taboada, M.: Constructive language in news comments. In: Proceedings of the First Workshop on Abusive Language Online. pp. 11–17. ALW1, Vancouver, Canada (2017)
29. Kowalski, R.M., Giumetti, G.W., Schroeder, A.N., Lattanner, M.R.: Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. *Psychol. Bull.* **140**(4), 1073–1137 (2014)
30. Lazer, D.M.J., et al.: The science of fake news. *Science* **359**(6380), 1094–1096 (2018)
31. Lewis, S.C., Holton, A.E., Coddington, M.: Reciprocal journalism: a concept of mutual exchange between journalists and audiences. *Journal. Pract.* **8**(2), 229–241 (2014)
32. Liu, J., McLeod, D.M.: Pathways to news commenting and the removal of the comment system on news websites. *Journalism* **22**(4), 867–881 (2021)
33. Loosen, W., et al.: Making sense of user comments: identifying journalists’ requirements for a comment analysis framework. *Stud. Commun. Media* **6**(4), 333–364 (2017)
34. MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: challenges and solutions. *PLoS ONE* **14**(8), 1–16 (2019)
35. Mansfield, M.: How we analysed 70m comments on the Guardian website (2016). <https://www.theguardian.com/technology/2016/apr/12/how-we-analysed-70m-comments-guardian-website>
36. McKnight, D.H., Carter, M., Thatcher, J.B., Clay, P.F.: Trust in a specific technology: an investigation of its components and measures. *ACM Trans. Manag. Inf. Syst.* **2**(2), 1–25 (2011)

37. Mondal, M., Silva, L.A., Benevenuto, F.: A measurement study of hate speech in social media. In: Dolong, P., Vojtas, P. (eds.) *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pp. 85–94. HT 2017, ACM, Prague, Czech Republic (2017)
38. Muddiman, A., Stroud, N.J.: News values, cognitive biases, and partisan incivility in comment sections. *J. Commun.* **67**(4), 586–609 (2017)
39. Niemann, M.: Elicitation of requirements for an AI-enhanced comment moderation support system for non-tech media companies. In: Stephanidis, C., Antona, M., Ntoa, S. (eds.) *HCII 2021. CCIS*, vol. 1419, pp. 573–581. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-78635-9\\_73](https://doi.org/10.1007/978-3-030-78635-9_73)
40. Niemann, M., Müller, K., Kelm, C., Assenmacher, D., Becker, J.: The German comment landscape. In: Bright, J., Giachanou, A., Spaiser, V., Spezzano, F., George, A., Pavliuc, A. (eds.) *MISDOOM 2021. LNCS*, vol. 12887, pp. 112–127. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-87031-7\\_8](https://doi.org/10.1007/978-3-030-87031-7_8)
41. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: *Proceedings of the 25th International Conference on World Wide Web*, pp. 145–153. WWW 2016, ACM Press, Montreal, Canada (2016)
42. Papacharissi, Z.: Democracy online: civility, politeness, and the democratic potential of online political discussion groups. *New Media Soc.* **6**(2), 259–283 (2004)
43. Park, D., Sachar, S., Diakopoulos, N., Elmqvist, N.: Supporting comment moderators in identifying high quality online news comments. In: Kaye, J., Druin, A., Lampe, C., Morris, D., Hourcade, J.P. (eds.) *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 1114–1125. CHI 2016, ACM, San Jose, CA, USA (2016)
44. Pöyhtäri, R.: Limits of hate speech and freedom of speech on moderated news websites in Finland, Sweden, the Netherlands and the UK. *Ann. Ser. Hist. Sociol.* **24**(3), 513–524 (2014)
45. Pritchard, S.: The readers’ editor on... closing comments below the line (2016). <https://www.theguardian.com/commentisfree/2016/mar/27/readers-editor-on-closing-comments-below-line>
46. Reich, Z.: User comments: the transformation of participatory space. In: Singer, J.B., (eds.) et al. *Participatory Journalism: Guarding Open Gates at Online Newspapers*, chap. 6, pp. 96–117. Wiley-Blackwell, Chichester, UK, 1 edn. (2011)
47. Riehle, D.M., Niemann, M., Brunk, J., Assenmacher, D., Trautmann, H., Becker, J.: Building an integrated comment moderation system – towards a semi-automatic moderation tool. In: *Proceedings of the HCI International 2020, Copenhagen, Denmark* (2020)
48. Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., Wojatzki, M.: Measuring the reliability of hate speech annotations: the case of the European refugee crisis. In: Beißwenger, M., Wojatzki, M., Zesch, T. (eds.) *Proceedings of the 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pp. 6–9. NLP4CMC III, Stefanie Dipper, Sprachwissenschaftliches Institut, Ruhr-Universität Bochum, Bochum, Germany (2016)
49. Rowley, J.: Conducting research interviews. *Manag. Res. Rev.* **35**(3/4), 260–271 (2012)
50. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**(5), 206–215 (2019)

51. Sadiq, S., Mehmood, A., Ullah, S., Ahmad, M., Choi, G.S., On, B.W.: Aggression detection through deep neural model on twitter. *Futur. Gener. Comput. Syst.* **114**, 120–129 (2021)
52. Sander, B.: Freedom of expression in the age of online platforms: the promise and pitfalls of a human rights-based approach to content moderation. *Fordham Int'l LJ* **43**, 939 (2019)
53. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Ku, L.W., Li, C.T. (eds.) *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pp. 1–10. SocialNLP 2017, Association for Computational Linguistics, Valencia, Spain (2017)
54. Siegert, S.: Nahezu jede zweite Zeitungsredaktion schränkt Online/Kommentare ein (2016). <http://www.journalist.de/aktuelles/meldungen/journalist-umfrage-nahezu-jede-2-zeitungsredaktion-schraenkt-onlinekommentare-ein.html>
55. Söllner, M., Hoffmann, A., Leimeister, J.M.: Why different trust relationships matter for information systems users. *Eur. J. Inf. Syst.* **25**(3), 274–287 (2016)
56. The Coral Project Community: Shutting down onsite comments: a comprehensive list of all news organisations (2016). <https://community.coralproject.net/t/shutting-down-onsite-comments-a-comprehensive-list-of-all-news-organisations/347>
57. The European Parliament: The Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Off. J. Eur. Union* **119**, 1–88 (2016)
58. Ullmann, S., Tomalin, M.: Quarantining online hate speech: technical and ethical perspectives. *Ethics Inf. Technol.* **22**(1), 69–80 (2019). <https://doi.org/10.1007/s10676-019-09516-z>
59. Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D.: User acceptance of information technology: toward a unified view. *MIS Q.* **27**(3), 425–478 (2003)
60. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. *Science* **359**(6380), 1146–1151 (2018)
61. Whittaker, E., Kowalski, R.M.: Cyberbullying via social media. *J. Sch. Violence* **14**(1), 11–29 (2015)
62. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on web 2.0. In: *Proceedings of the Content Analysis in the WEB*, pp. 1–7. CAW2.0, Madrid, Spain (2009)
63. Yin, W., Zubiaga, A.: Towards generalisable hate speech detection: a review on obstacles and solutions. *Peer J. Comput. Sci.* **7**, 1–38 (2021)