The Moral Gatekeeper? Moderation and Deletion of User-Generated Content in a Leading News Forum

Frischlich, Lena; Quandt, Thorsten; Schatto-Eckrodt, Tim; Boberg, Svenja

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

COGITATIO

Media and Communication (ISSN: 2183–2439)
2018, Volume 6, Issue 4, Pages 58–69
DOI: 10.17645/mac.v6i4.1493

Article

The Moral Gatekeeper? Moderation and Deletion of User-Generated Content in a Leading News Forum

Svenja Boberg *, Tim Schatto-Eckrodt, Lena Frischlich and Thorsten Quandt

Department of Communication, University of Muenster, 48143 Muenster, Germany;
E-Mails: svenja.boberg@uni-muenster.de (S.B.), tim.schatto-eckrodt@uni-muenster.de (T.S.-E.),
lena.frischlich@uni-muenster.de (L.F.), thorsten.quandt@uni-muenster.de (T.Q.)

* Corresponding author

Submitted: 24 March 2018 | Accepted: 27 June 2018 | Published: 8 November 2018

Abstract
Participatory formats in online journalism offer increased options for user comments to reach a mass audience, also enabling the spreading of incivility. As a result, journalists feel the need to moderate offensive user comments in order to prevent the derailment of discussion threads. However, little is known about the principles on which forum moderation is based. The current study aims to fill this void by examining 673,361 user comments (including all incoming and rejected comments) of the largest newspaper forum in Germany (Spiegel Online) in terms of the moderation decision, the topic addressed, and the use of insulting language using automated content analysis. The analyses revealed that the deletion of user comments is a frequently used moderation strategy. Overall, more than one-third of comments studied were rejected. Further, users mostly engaged with political topics. The usage of swear words was not a reason to block a comment, except when offenses were used in connection with politically sensitive topics. We discuss the results in light of the necessity for journalists to establish consistent and transparent moderation strategies.

Keywords
community management; computational methods; forum moderation; gatekeeping; journalism; participatory media; Spiegel Online; topic modeling; user comments; user participation

Issue
This article is part of the issue "News and Participation through and beyond Proprietary Platforms in an Age of Social Media", edited by Oscar Westlund (Oslo Metropolitan University, Norway) and Mats Ekström (University of Gothenburg, Sweden).

1. Introduction

Which information makes it into the news, thus gaining the possibility of attracting the attention of a mass audience? From the very beginning, this fundamental gatekeeping decision has accompanied the work of journalists. Nowadays, the figurative "gate" journalists are keeping has changed. Since the emergence of participatory formats, journalists are no longer the only communicators publishing content on their news outlets; user comments have become a widely established supplement to journalistic output (Walther & Jang, 2012), though the value of comment sections has been questioned by news organizations resulting in the transfer of participatory spaces to non-proprietary platforms such as Twitter and Facebook (Karlsson, Bergström, Clerwall, & Fast, 2015). One reason for these developments might be that the maintenance of comment sections is costly and challenging. In contrast to earlier hopes of encouraging constructive discussions (Papacharissi, 2004) and actively integrating users in news production processes (Bruns, 2008), user comments have been found to also open the floor for "dark participation" (see Quandt, 2018), ranging from misinformation and hate campaigns to individual trolling and cyberbullying. Researchers have focused on these uncivil forms of communication, such as the spreading of vulgar language, disrespect, and aggression, highlighting their possible negative impacts (Coe, Kenski, & Rains,

2014). Yet other readers tend to respond adamantly to such uncivilities, increasing the risk of derailing debate (Ziegele, Breiner, & Quiring, 2014).

To prevent the abuse of participatory comment sections, journalists no longer only guard their open gates (Singer et al., 2011) by pre-selecting valuable user comments, but they also keep a vigilant eye on the comment section, ready to throw out anyone who transgresses the rules (Ksiazek, 2015). Although professional journalists feel morally obliged to create a discussion-friendly environment (Meltzer, 2015), they often lack a shared understanding of not only *what* they find uncivil and threatening (Frischlich, Boberg, & Quandt, 2017) but also *how* to deal with it (Muddiman & Stroud, 2017). Studies on the underlying factors of these inconsistent moderation decisions are rare and rely mostly on journalists' subjective perceptions (Diakopoulos & Naaman, 2011).

Which factors affect the actual gatekeeping decision? Do journalists mainly respond to widely acknowledged taboos, like offensive language and swearing, or are there problematic topics that are moderated with greater care? The current study aims to fill this void. We conducted an automated content analysis of a unique large dataset set from a six-month period of the entire comment section (pre- and post-moderation) of the leading German news outlet Spiegel Online (SPON). Our analysis unveils hidden gatekeeping decisions and allows for the observation of real gatekeeping processes.

2. Gatekeeping in Times of Participatory Journalism

Gatekeeping represents one of the most studied areas of communication research, dealing with the question of how editorial space is filled and how topics, events, and interpretative patterns are prioritized (Heinderyckx, 2015). Since White's (1950) depiction of journalists as rational news selectors who mainly depend on their individual freedom of choice, the concept has expanded in regard to several other factors, such as institutional structures on organizational and societal levels (Vos, 2015).

The emergence of participatory journalism not only changed journalistic decision-making processes but also the position of journalistic organizations in the information network. As citizens have gained opportunities to add information to the news, multiple "gates" have opened (Williams & Delli Carpini, 2000), resulting in a myriad of information sources and actors involved in communication processes. Traditional media gates have not lost their importance, but as communication hierarchies have flattened, researchers rather refer to "curated flows" (Thorson & Wells, 2015, p. 27) or "gatewatching" (Bruns, 2005, p. 1) when they describe the selection and editing of news content online. This development has also changed journalism on the output level with participatory formats, particularly user comments (Walther & Jang, 2012), becoming an established feature on newspaper websites.

The emergence of these new communication channels was accompanied by euphoric hopes that the world was witnessing a new form of the deliberative public sphere (Bruns, 2008). Early studies show that journalists embrace the idea of being in touch with their readers but are also reluctant to offer them full access to their platforms (Domingo et al., 2008). According to Lewis and Westlund's (2015) systematization of cross-media news work, audience perceptions vary according to journalistic roles and activities. Considering tasks of community management, like observation and selection, journalists see users as active participants, yet the way journalists communicate with their readers has not changed fundamentally. Indeed, editors in online newsrooms are still in charge of the production process and are only willing to allow small "walled gardens" for actual user participation (Hanitzsch & Quandt, 2012). Since media organizations also implement participative offerings as additional channels of distribution (especially in case of non-proprietary platforms like Facebook), audiences are perceived as commodities or statistically aggregated target groups (Lewis & Westlund, 2015) resulting in the challenge of balancing editorial and economic goals. Nevertheless, possibilities of user engagement are usually limited to polls, comment sections, and social media sites as another outlet of news.

3. Guarding the Gates against Uncivil Intruders: Why Journalists Perceive Moderation to Be Necessary

Participatory formats offer journalists a great way to get directly in touch with their readership (Vos, 2015) and for users to articulate their views and evaluate the journalistic output. Besides constructive discussions, participatory formats allow irrelevant or even uncivil content to reach the public's eye. Coe et al. (2014) define incivility as an "unnecessarily disrespectful tone toward the discussion forum, its participants or its topics" (p. 660), which manifests as offensive attacks against other persons (Gagliardone et al., 2016), social groups (Engelin & De Silva, 2016), or disruption of the discussion for one's own amusement (i.e., trolling; Binns, 2012). Often, incivility is accompanied by swearing in terms of using highly arousing and offensive language (Kwon & Cho, 2017). In contrast to incivility as a whole, which can be quite hard to detect (Ross et al., 2016), swearing is less difficult to recognize for forum moderators. The presence of obscene language can be easily detected by both community managers and keyword-based algorithms. Also, journalists' ethical guidelines clearly condemn the use of offensive language (Muddiman & Stroud, 2017).

Since user comments are often discussed as indicators of the opinion climate, journalists fear how offensive and hateful comments could affect public discussions. Community managers not only deal with single users but also with orchestrated attacks exploiting the trustworthy environment of traditional news outlets (Tandoc, Lim, & Ling, 2017). Also, recent political controversies, like the election of USA President Donald Trump, have turned the interaction with user-generated content itself into a con-

tentiously discussed topic (Hofseth, 2017). The reason why news media still enable user comments is rooted in the journalistic role of the "press advocating for the public [and] serving as its voice in a mass-mediated society" (Braun & Gillespie, 2011, p. 385). In that regard, comments are seen as an additional tool to create a deliberative public sphere.

Following this line of thought, Lewis, Holton and Coddington (2014) introduced the concept of "reciprocal journalism" (p. 230), which describes the relationship between journalism and participatory formats as an interaction both sides benefit from. Journalists function as community builders who encourage an active discourse. In order to sustain the bond between media outlets and users, community managers have to establish an environment that "operates on and continues to foster trust" (Lewis et al., 2014, p. 235). This also implies that their responsibility is to protect the public from cyberhate and content that could harm vulnerable groups (Pöyhtäri, Con, In, Bassi, & Bretagna, 2014). The prevalence of hate and disrespectful communication might damage this trusting relationship by putting off users who want to engage in a constructive discussion as well as making journalists question the overall benefit of comment sections. The latter recently caused several media outlets to disable their comments (Moosa, 2014). Also, research shows that the engagement of users is rather low, which also diminishes the commercial value of comment sections from the media organizations' perspective (Karlsson et al., 2015).

As a consequence, community managers operate in a field of tension between their perceived moral obligation to keep undesirable content out of the comment section and their efforts to engage and reach people via participatory formats. They have to balance the risk of letting undesirable content slip through and scaring off users who would prefer a focused, theme-oriented discussion or rejecting too much, thereby restricting their forum and possibly being accused of censorship. Furthermore, they are often challenged by a vast amount of content that has to be handled in tandem with other daily tasks. As a result, journalists need to develop strategies to help integrate the moderation of user comments into their daily newsroom routines.

4. How to Deal with Undesirable Comments: Strategies of Community Management

Facing the challenges of participatory journalism, gatekeepers have been forced to differentiate their journalistic roles in order to handle problematic user comments. In that regard, comment moderation can be more or less restrictive. Community managers mostly rely on non-interactive strategies (Frischlich et al., 2017), which basically involve the decision of whether to block a comment or not. Non-interactive strategies include the laissez-faire approach of trusting the community's self-regulatory efficacy, more restrictive means like deactivating the comment sections below articles dealing with potentially sensitive topics (*closing the gates*; Nielsen, 2012; Reich, 2011), enabling single comments after inspection (*guarding the gates*), or scanning the comments for unwanted content and deleting it retroactively (*patrolling behind the gates*; Ksiazek, 2015).

Analogous to the "hierarchy of influences" conception of the journalistic working process (Reese & Shoemaker, 2016), the individual moderation decision is affected by newsroom routines, media organizations for which the journalists work, and the societal institutions and social system in which they operate. As described above, journalists feel obligated to provide a public forum for increasing awareness of relevant societal issues (Braun & Gillespie, 2011). On the level of newsroom routines, moderation is influenced by the political leaning, editorial policy, and quality standards of the media brand (Pöyhtäri et al., 2014), which include editorial guidelines such as netiquette.

Looking at the content of comments, research shows that the topic of the discussion influences the amount of incivility journalists discover (Ksiazek, 2018) as well as the perceived necessity for a moderator to intervene (Loosen et al., 2017). Comment threads on sports or hobbies are perceived as less problematic, whereas political issues are often accompanied by uncivil content, which not only applies to general topics but also to the framing of issues (i.e., portraying refugees as potential criminals). Beyond the respective topic of the comment thread, user comments also address the development of the comment thread *itself* as a subject of discussion. These examples of meta-discussion often manifest themselves as discontent with journalistic news production or forum moderation (i.e., allegations of journalists being partial or even lying; Prochazka & Schweiger, 2016) or critical remarks towards other commenters (Loosen et al., 2017) and thus raise the awareness of community managers. Further, comments that include swearing are blocked rather consistently (Muddiman & Stroud, 2017). Most plausibly, this is because the prevalence of swearing is an obvious and easy-to-detect feature in the comment and also a clear violation of discussion norms to which journalists adhere (Pöyhtäri et al., 2014). Since swearing in relation to political topics attracts readers' attention, Kwon and Cho (2017) conclude that the norms around the acceptable degree of swearing vary across topical areas, so it can be assumed that the prevalence of swearing and the topic of the respective comment serve as the most obvious characteristics to be considered in the moderation decision.

Even though editorial guidelines serve as a point of reference, the decision of which comments to reject is often based on personal experiences (i.e., frequent exposure to hateful content) or even gut feelings (Frischlich et al., 2017). Therefore, differences not only between distinct media outlets but also within the same newsroom can be expected.

Little is known about the effectiveness of moderation. Requiring user registration and pre- and post-

moderation of discussion threads clearly promotes a more civil platform (Ksiazek, 2015). To date, studies that are able to compare the actual incoming comments with community managers' moderation decisions are scarce. As a notable exception, the study by Muddiman and Stroud (2017) on moderation of comments in the New York Times' online forum showed that community managers partially tolerated forms of incivility other than swearing because readers engaged heavily with swearing, and swearing merely poisons the climate of discussion.

5. The Case of the SPON Forum

The research object of this study, SPON, is one of the most important German news websites. Launched in 1994, it carries on a long tradition in the online market. The website has 20.64 million unique users per month (Statista, 2018) and is the third most frequently visited news website in Germany. SPON has the largest online forum in the country with comments visible to everyone, although users have to register in order to write a comment.

The user comments in the SPON forum are handled by 11 trained social media editors who have long-standing experience in the moderation of content. Along with maintenance of the forum, they are also responsible for other social media channels, such as Facebook and Twitter. Comments are checked for violations of the netiquette individually in the context of the discussion thread. Thereby, the forum aims to encourage an "open, friendly and respectful climate of discussion" and further seeks a "fair and factual tone of argumentation" (SPON, 2018). Comments that include swearing, vulgar language, or other elements of disrespectful and aggressive communication are banned. In the SPON comment section, mostly post-moderation is used. Additionally, for about 30% of the articles, a form of pre-moderation takes place, namely closing the discussion threads on sensitive topics like Middle East conflicts or the refugee crisis (Kriesel, 2017).

Analogous to the existing literature on forum moderation and as outlined in the SPON netiquette, the prevalence of swearing seems to be an important cue to be considered in the moderation decisions of SPON community management, since swear words are easy to detect by only scanning a comment or with the technical support of keyword filters. Also, the fact that SPON disables the comment sections under certain topics shows their sensitivity to problems of incivility that might arise with regard to issues that have been perceived as problematic in the past. But do the community managers of SPON also use the prevalence of swearing as an obvious reason to block a comment in order to preserve a friendly tone in the discussion? Are they more alerted to political topics in which swearing is less likely to be tolerated? To explore these questions, we formulated the following research questions.

- RQ1: Which topics are brought up in the user comments of the SPON forum (before moderation)?
- RQ2: To what extent do the comments include swearing (before moderation)?
- RQ3: Are comments that include swearing more likely to be banned by the forum moderators (moderation decision)?
- RQ4: Are comments that include swearing more likely to be banned when they occur in political contexts compared to non-political contexts (moderation decision)?

6. Method

To explore these questions, we used a six-month dataset of the complete SPON forum, which gives meaningful insight into how community managers handle user comments. This unique data resembles the whole input in the form of pre- and post-moderation comments, allowing the analysis of comments that were not publicly accessible.

6.1. Data

During the examined period (November 30, 2016–May 16, 2017), a total of 673,361 comments were posted referring to 9,548 articles. More than one-third of the comments (35%) were rejected by community managers after publication.

Before the analysis, a number of common preprocessing steps were applied (for an overview, see Günther & Quandt, 2016), including removing HTML markup, URLs, and stop words. Still, the data contained a lot of meaningless tokens, which were removed by excluding words that occurred less than 20 times ($n = 643,298$). To manage the ambiguous use of names (i.e., "Mrs. Merkel", "Angela"), the named entities of the comments were extracted with the Python software library spaCy (Honnibal & Johnson, 2015) and standardized manually.

6.2. Analysis

To explore what people in the SPON forum were talking about, we identified comment topics using latent Dirichlet allocation (LDA; Blei, Ng, & Jordan, 2003). LDA is an unsupervised learning algorithm that discovers latent topics inductively based on patterns of words that co-occur in the same document. It provides information on (i) to what extent each word of the corpus characterizes each topic (β) and (ii) to what extent each topic is present in each document (γ). Each comment can be a mixture of several topics (Günther & Domahidi, 2017). There is no clear-cut definition of characteristics of topics in theoretical terms; the meaning of the LDA-detected topics is assessed empirically by the interpretation of characteristic features of the respective topics (Maier et al., 2018). Since the topics are derived from co-occurring words, they do not necessarily resemble general topics of me-

dia coverage, like politics or sports or certain events like elections, but capture prevalent patterns of the way in which certain issues are addressed or framed (Jacobi, Van Atteveldt, & Welbers, 2016). For example, the LDA might identify two different topics that deal with the same issue but differ regarding the valence of the co-occurring top terms.

Before estimating the topic model, two parameters have to be predefined: (i) the number of topics (k) and (ii) the number of topics allowed per document (α). To find the ideal numbers of k and α, a series of 200 topic models were computed based on training and test samples using the LDA function of the R topicmodels package (Grün & Hornik, 2011). We found that the model of $k = 25$ and $\alpha = 5$ had a large increase in predictive power. These parameters were run several times on various random samples of the data, providing reproducible results with only minor deviations. The 25 topics were characterized by looking at the top terms and documents most representative for each topic. The interpretation was validated by two additional coders who were able to find the same labels for each topic. As a second validation step, an analysis of intercoder-reliability was performed to assess to what degree human coders and the topic modeling concur. The resulting kappa (Cohen, 1960) indicated substantial agreement, $\kappa = 0.76$ (Landis & Koch, 1977). It is noteworthy that the human coders agreed as much as the comparison of algorithm and human coders.

Swear words were detected following a deductive rule-based approach (Günther & Quandt, 2016). A swearing dictionary was implemented based on an actual keyword list used by journalists to prefilter insulting comments (Frischlich et al., 2017), which was extended by an online search of further swear words, resulting in 1,829 terms (i.e., "asshole", "idiot", or racial or misogynist slurs). The dictionary was matched with the text of the comments, extracting the respective swear word and the variable "contains swearing" (yes/no). With regard to RQ3, a subsample was created, including all comments that contain swear words.

Due to the large number of cases, it is challenging to infer meaningful relationships. For instance, using standard null-hypothesis significance testing on the given sample size would most likely result in finding a significant difference between the published and rejected corpus, even though the difference might be close to non-existent (Weber & Popova, 2012). To bypass this problem, the logic of the independent sample t-test is reversed; instead of testing for *difference* and rejecting the null hypothesis (*no difference*), the data is tested for *equivalence*, which means rejecting the rephrased H_0 (*true effect*) and supporting the alternative hypothesis (*absence of an effect that is worth examining*; Lakens, 2017). Naturally, a null-effect cannot be supported; thus, a maximum-no-effect (Δ) has to be predefined as a threshold. In the current study, the equivalence tests were calculated following Weber and Popova (2012), applying the mathematical formula to an R function and testing common effect sizes (small: $\Delta = 0.1$; medium: $\Delta = 0.3$; large $\Delta = 0.5$).

7. Results

To some degree, the extracted topics resemble the typical repertoire of news media coverage, including politics, sports, culture, and education (see Table 1). In line with prior studies, the data shows that users engaged heavily with political topics. Not only were almost half (10/25) of the identified topics about general political issues, such as democracy, or specific events, such as elections or the refugee crisis, political comments were also rather frequent in the corpus, especially the German federal election ($n = 28,018$), the civil war in Syria ($n = 25,849$), and diplomatic relations to Turkey ($n = 22,842$).

Apart from generic topics and current events, the LDA also revealed several forms of meta-discussions that were brought up by SPON forum users, namely constructive discourse, uncivil discourse, "fake news" accusations, and trolling. The constructive and uncivil discourse topics both addressed netiquette as an issue but through different frames. On the one hand, they were contrasted in a call for a civil debate and, on the other, used to discredit other users or community management. The topic addressing "fake news" did not cover the ongoing public debate on this phenomenon (see Quandt, Frischlich, Boberg, & Schatto-Eckrodt, in press) but used the term "fake news" as a complaint against SPON. Often, this complaint was associated with accusations of censorship against community management, thus representing a disclaiming remark towards legacy media in general and SPON in particular rather than referring to the general issue of media coverage. Finally, the trolling topic was characterized by rather pointless disruptive or uncivil language. These comments did not address media critique in a direct manner but, nonetheless, qualified it as stance against the general discussion thread by disrespecting discussion norms, such as relevance to the issue and civility.

In general, there is no topic which appears exclusively in the published or blocked comments. Plausibly, the comments that hint at a disrespectful way of communication, such as accusations of mainstream media being liars or "Fake News", trolling, and uncivil discourse, are rejected more often. Also, comments on controversial political issues are often subject to moderation. The distribution of topic-means among the published and blocked comments does not seem to indicate that community managers are more alert to political hot topics. Naturally, these topics evoke more engagement and maybe even more uncivil behavior. Nevertheless, the differences are barely noteworthy.

As community managers widely rely on keyword-based classification of presumable uncivil content that requires further inspection, swearing can be considered one of the key identifying features of rejected comments. With a total of 58,176 (8.6%), the number of comments

Table 1. Description of user comment topics.

Description	Most representative terms	Prevalence of topic (*n* of comments where $\gamma > 0.3$)
Politics		
German federal election	SPD, Schulz, Merkel, CDU, Green party, party, AfD, voter, politics	28,018
war in Syria	Russia, USA, war, Syria, Putin, Assad, NATO, Western World, weapon, Ukraine	25,849
tensions in turkey	Turkey, Erdogan, Germany, Turkish people, nation, Merkel, government, Europe, politician	22,842
USA election & trump administration	Trump, USA, Obama, president, Putin, Clinton, world, American people, Democrats, Russia	20,076
refugees & threat of crime and terror	Germany, nation, refugee, police, live, Islam, religion, immigrant, victim, Berlin	19,090
Eu & Brexit	Europe, Germany, UK, Brexit, nation, France, Poland, Italy, Switzerland, Brussels	17,755
right-wing populism	AfD, right, left, party, the Left Party, opinion, Höcke, democracy, Germany, Nazi	15,914
democracy	election, democracy, majority, elected, the people, voter, citizen, politics, parties, politicians	14,389
Society		
Families & education	children, woman, parent, man, school, live, learn, teacher, student, family	17,690
Societal norms	people, live, society, politics, freedom, democracy, nation, capitalism, future, population	14,661
Elite critique	Politician, Mr., Mrs., responsibility, military, Merkel, Germany, official, boss, DDR	13,269
Law	law, case, state, rule, apply, judge, court, citizen, judgement, question	13,161
Science	question, earth, number, statement, actual, study, comparison, statistics, science, result	13,109
Economy		
Employment, taxes & pension	money, pay, tax, work, Euro, Germany, state, cost, income, pension	22,694
European financial crisis	money, Euro, Germany, billion, Greece, bank, depts, millions, pay, cost	16,695
Global economy	USA, Germany, China, company, product, world, market, economy, land, Trump	14,221
Consumer Service		
Automobile & energy	car, drive, VW, diesel, electricity, PS, vehicle, Tesla, kilowatt hour	26,157
Infrastructure	railway, internet, Berlin, data, customer, Hamburg, city, airport, fast, smartphone	18,170
Health	eat, people, living, water, doctor, meat, alcohol, patient, couple, beer	15,965
Leisure		
Sports	FC Bayern, BVB, game, player, soccer, fan, club, team, rank, last	21,930
(Pop)Culture	woman, watch, movie, music, picture, sad, Tatort (German TV show), nice, art, show	17,364

Table 1. (Cont.) Description of user comment topics.

Description	Most representative terms	Prevalence of topic (*n* of comments where $\gamma > 0.3$)
Meta-Discussion		
"Fake News"	media, SPON, fake, news, fact, press, article, Spiegel magazine, truth	18,496
Uncivil discourse	contribution, thanks, read, SPON, write, comment, topic, question, forum, opinion	18,328
Trolling	people, say, bla, nix, money, whatever, believe, stupid, blame, real	13,756
Constructive discourse	question, problem, situation, opinion, politics, effective, condition, manner, behavior, topic	11,887

Notes: LDA (method = Gibbs, $k = 25$, $\alpha = 5$, $n = 67,336$).

that included swear words or racial slurs was surprisingly low (*RQ2*). In fact, the significant equivalence test ($t = -36.68$, $\Delta < 0.1$, $p < 0.001$) shows that the presence of swearing does not discriminate between the published and rejected comments, or at least, the effect size is minimal ($r < 0.1$). Thus, comments that include swearing are not more likely to be banned (*RQ3*). It is worth mentioning that the individual swearing terms of the dictionary were also distributed equally in the published and rejected comments except for some terms of xenophobic slurs, like "goatfucker", or political insults, such as "nazi-slut", which were blocked in over 90% of the cases.

So, if the occurrence of swearing as an agreed-upon violation of the netiquette is alone not enough to attract the attention of community managers, which comment characteristics are? Relatedly, in which topics is swearing tolerated or handled more restrictively? With regard to *RQ4*, all comments that contained swearing were tested for equivalence among the moderation decisions for each topic. Again, the equivalence among the published and blocked corpus was tested with a threshold of a presumed maximum-no-effect of $\Delta = 0.1$, $\Delta = 0.3$, and $\Delta = 0.5$.

For the vast majority of the topics, the assumption of equivalence can be supported, meaning there is no appreciable difference between topic and moderation decision in comments with swearing (see Figure 1). However, for the topics "automobile" ($p = 0.061$), "right-wing populism" ($p = 0.99$), "fake news" ($p = 0.28$) and "threat (terror/refugees)" ($p = 0.26$), the assumption of a maximum-no-effect of $\Delta = 0.1$ is not supported. The data shows that community managers were more likely to tolerate swearing in the context of automobiles, for instances regarding the diesel emissions scandal. Swearing was less tolerated in the context of right-wing populism, fake news allegations, and associating refugees with threats to national security. Yet the differences between the published and the rejected corpus are rather small; when applying a medium maximum-no-effect of $\Delta = 0.3$, the equivalence tests for all topics are highly significant.

In sum, the results show that the users of the SPON forum engage heavily in political discussions as well as meta-discourses on the netiquette of the forum. We found no topic-related differences between the published and rejected comments. Also, the use of swear words was not a key indicator in the rejection of comments, whereas racial slurs were blocked rather consistently. Even though the forum moderators were slightly more restrictive on the co-occurrence of swearing and topics dealing with refugee politics, fake news allegations, or right-wing populism, systematic moderation or even exclusion of certain topics can be denied. Thus, to understand moderation decisions, further context factors must be considered.

8. Discussion

Community managers and digital editors are expected to guard the open gates of online newspapers (Singer et al., 2011) against dark participation—with the obvious challenge of finding an adequate level of intervention. The current study aimed at providing empirical insights into the gatekeeping processes of community managers.

The results show that there is neither consistency nor a systematic way of blocking certain topics or styles of communication. Not even swearing as a generally agreed-upon violation of both journalistic professional norms and netiquette was eliminated consistently from published comments. There is a slight indication, though, that racial and misogynistic slurs are more strictly, yet not completely, blocked demonstrating journalists' efforts to protect vulnerable social groups. We also found that the use of swear words is not handled more or less restrictively in conjunction with specific topics. However, we found small differences in moderation behavior of swearing in conjunction with comments on the refugee crisis, fake news, and right-wing populism. This finding hints at the community managers' endeavors to keep offensive language out of already sensitive topics that refer to nationally prevalent political controversies in order to fulfill a mediating role in the discourse. Notably, SPON does not enable all articles to be commented on, so topic-related moderation decisions that took place beforehand are not reflected in our results.

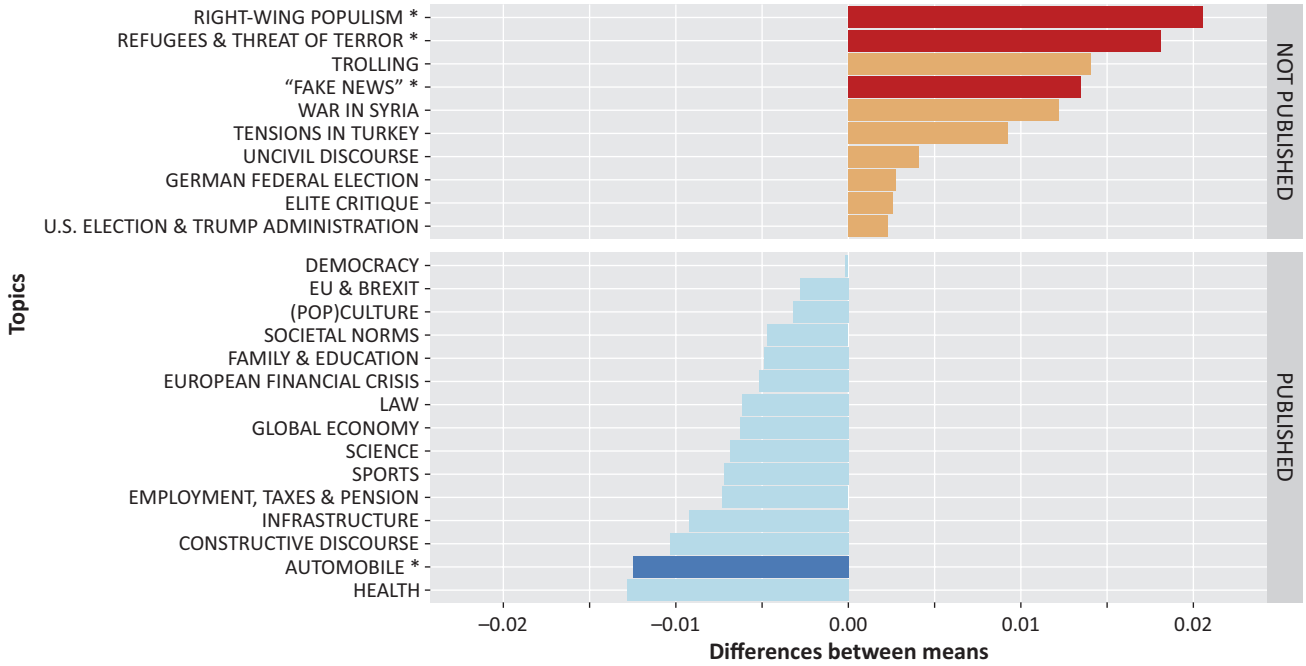

Figure 1. Equivalency among topic and gatekeeping decision on the comments that include swearing. $N = 58,176$, $\Delta = 0.1$; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Overall, the data shows that there was no systematic cleaning by topic. Simple heuristics, such as abusive language, are unrelated to whether a comment was rejected. This allows for two possible conclusions: either there are context factors other than the topic or independent from the content of the comment, or there is no systematic strategy, and moderation fully depends on the individual instincts of community managers. The non-consistent blocking of swearing could also hint at certain contexts in which swearing is not perceived as problematic; insults could, for example, also be used to admonish other users (i.e., "don't act like an [swear word]"). Further research should take the context of swearing into account—especially to whom the words refer—in order to evaluate if they are used to disrespect other users in a hostile way. Context factors that are independent from the content of a single comment could include personal or organizational constraints, such as work-load, the lack of supportive resources, the number of comments streaming in at the same time, or the recognition and blocking of known troublemakers independent from the content of the post.

There are some limitations to this study. The analysis was limited to one forum, and therefore, influencing factors on the organizational level, such as the political leaning of the newspaper in regard to the restrictiveness of moderation policies, could not be explored. Also, to further explore the inter-individual effects of gatekeeping decisions, it would be necessary to know who exactly moderated each respective comment.

Although we discovered no general patterns of moral red lines, the mere fact that more than one-third of the comments were rejected demonstrates that journalists feel morally obligated to protect their comment sections from harmful content or, at least, content that is perceived as such. Yet we do not know which standards they apply or which concept of an ideal moderation they pursue. Further research should investigate which aspirations individual community managers associate with a functioning forum moderation and why they think certain levels of restrictiveness are vital to online discussions.

Finally, the current study finds strong evidence against media-critical conspiracy theorists who believe that the mainstream media systematically conceals issues that are opposed to the political mainstream and blocks comments that offer alternative views. Nevertheless, non-transparent moderation practices make it difficult for users to understand why their posts have not been published and stir up feelings of mistreatment. With regard to the concept of reciprocal journalism, media outlets should define for themselves which benefits they derive from enabling comment sections and, further, what kind of forum they want to offer to their readers. Following this, moderation guidelines should be developed that are not only in line with this strategic decision but are also application-oriented and provide more detailed instructions than the general framework of the netiquette. Most importantly, the selection and rejection of user comments should be transparent to the users of the forum. Even if this might not silence every "fake news" accusation, it could help to regain trust from the readers who feel misunderstood from time to time but are generally willing to engage in a deliberative discussion.

9. Conclusions

The current study demonstrates that forum moderators face the continuous challenge of creating spaces for user participation that are beneficial for both the media organizations and their readership while having to protect these spaces from dark forms of participation, like hateful content, disruptive or nonsense comments, or even threatening accusations. The mere fact that a substantial amount of user comments is perceived to be not suitable to reach the public eye raises the question of why media organizations even bother to encourage user participation—or differently phrased, what do media organizations envision as the ideal forum for user participation? In this context, it is worth investigating for future research how much proprietary and non-proprietary platforms of user participation vary in terms of audience perception and journalistic intervention. Proprietary platforms have the potential to target the media outlet's core audience while leaving the journalists in charge. As platform providers, community management could think of new measures to guarantee the kind of online discussion for which they aim, for example, making sure that commenters have to read the article before participating or constantly identifying and blocking users who violate the rules of the forum. However, we observe that these measures of control are not fully taken advantage of, yet. When it comes to non-proprietary platforms, media outlets let go of these means of control even more; in return, they potentially reach a broader audience. These circumstances make it all the more necessary for media organizations to develop a consistent and transparent roadmap for handling user comments.

The results show that even journalists of a single outlet do not share common rules when it comes to the selection of user comments, except a very small effect was noted in the blocking of severe racial slurs in connection with topics related to refugees, right-wing populism, or fake news accusations. Instead, gatekeeping decisions depend to a substantial degree on inter-individual differences. From the users' perspective, participative formats offer the chance to discuss a broad variety of different issues. Even though possibilities of actively participating in news production processes are limited by the restrictions of media outlets, the results clearly show that single voices or views are not systematically silenced by forum moderation.

So is user participation an enrichment or a daily struggle? Community managers are eager to ban dark forms of participation but also want to leave their users room for discussion at the same time. In this context, the traditional questions of gatekeeping research are still interesting: Which comments are considered to be worth publishing and, therefore, selected by forum moderators? The current study contributes to this field of research by integrating methods of computational social science and, therefore, offering insights into the *actual gatekeeping decision*. Although these journalists were partly able to keep their gates against aggressive and disrespectful language, their decisions were not fully based on a set of obvious standards like the consequent filtering out of swearing but, rather, shaded by the moderator's personal moral compass. Still, participative formats offer unique possibilities for media outlets to get in touch with their audience. However, if media organizations want to fully tap into this potential, they must figure out how to deal with these challenges. The fact that moderation decision-making processes are often not fully comprehensible might unintendedly fuel censorship-critique among readers, thus damaging the image of participatory journalistic media in the long run.

Acknowledgments

The research was funded by the Federal Ministry of Education and Research. We acknowledge support by the open access publication fund at the University of Muenster.

Conflict of Interests

The authors declare no conflicts of interest.

References

Binns, A. (2012). Don't feed the trolls! Managing troublemakers in magazines' online communities. *Journalism Practice*, *6*(4), 547–562.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4/5), 993–1022.

Braun, J., & Gillespie, T. (2011). Hosting the public discourse, hosting the public. *Journalism Practice*, *5*(4), 383–398.

Bruns, A. (2005). *Gatewatching. Collaborative online news production*. New York, NY: Peter Lang.

Bruns, A. (2008). Life beyond the public sphere: Towards a networked model for political deliberation. *Information Polity*, *13*(1), 65–79.

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

Cohen, J. (1960). A coefficient of agreement for dominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Diakopoulos, N., & Naaman, M. (2011). Towards quality discourse in online news comments. In *Proceedings of the ACM 2011 conference on computer supported cooperative work* (pp. 133–142). Hangzhou: Alibaba Group.

Domingo, D., Quandt, T., Heinonen, A., Paulussen, S., Singer, J. B., & Vujnovic, M. (2008). Participatory journalism practices in the media and beyond: An international comparative study of initiatives in online newspapers. *Journalism Practice*, *2*(3), 326–342.

Engelin, M., & De Silva, F. (2016). *Troll detection: A comparative study in detecting troll farms on Twitter using cluster analysis* (Unpublished Dissertation). KTH, School of Computer Science and Communication, Sweden. Retrieved from urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-186406

Frischlich, L., Boberg, S., & Quandt, T. (2017). *Online newspapers as target of manipulative user-generated content: Dealing with hate speech, fake news, and covert propaganda*. Paper presented at the Annual Conference of the International Journal of Press and Politics, Oxford.

Gagliardone, I., Pohjonen, M., Beyene, Z., Zerai, A., Aynekulu, G., Bekalu, M., Teferra, Z. (2016). *Mechachal: Online debates and elections in Ethiopia: From hate speech to engagement in social media*. Retrieved from http://dx.doi.org/10.2139/ssrn.2831369

Grün, B., & Hornik, K. (2011). Topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, *40*(3), 1–30.

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

Günther, E., & Quandt, T. (2016). Word counts and topic models. *Digital Journalism*, *4*(1), 75–88.

Hanitzsch, T., & Quandt, T. (2012). Online journalism in Germany. In E. Siapera & E. Veglis (Eds.), *The handbook of global online journalism* (pp. 429–444). Oxford: Wiley-Blackwell.

Heinderyckx, F. (2015). Gatekeeping theory redux. In T. P. Vos & F. Heinderyckx (Eds.), *Gatekeeping in transition* (pp. 253–268). New York, NY: Routledge.

Hofseth, A. (2017). *Fake news, propaganda, and influence operations: A guide to journalism in a new and more chaotic media environment*. Oxford: Reuters Institute. Retrieved from reutersinstitute.politics.ox.ac.uk/news/fake-news-propaganda-and-influence-operations-—guide-journalism-new-and-more-chaotic-media

Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on empirical methods in natural language processing* (pp. 1373–1378). Lisbon: EMNLP.

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*(1), 89–106.

Karlsson, M., Bergström, A., Clerwall, C., & Fast, K. (2015). Participatory journalism: The (r)evolution that wasn't. Content and user behavior in Sweden 2007–2013. *Journal of Computer-Mediated Communication*, *20*(3), 295–311. Retrieved from 10.0.4.87/jcc4.12115

Kriesel, D. (2017). Spiegel mining. *Dkriesel*. Retrieved from www.dkriesel.com/spiegelmining

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

Ksiazek, T. B. (2018). Commenting on the news: Explaining the degree and quality of user comments on news websites. *Journalism Studies*, *19*(5), 650–673.

Kwon, K. H., & Cho, D. (2017). Swearing effects on citizen-to-citizen commenting online: A large-scale exploration of political versus nonpolitical online news sites. *Social Science Computer Review*, *35*(1), 84–102.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lewis, S. C., Holton, A. E., & Coddington, M. (2014). Reciprocal journalism. *Journalism Practice*, *8*(2), 229–241.

Lewis, S. C., & Westlund, O. (2015). Actors, actants, audiences, and activities in cross-media news work: A matrix and a research agenda. *Digital Journalism*, *3*(1), 19–37.

Loosen, W., Häring, M., Kurtanović, Z., Merten, L., Reimer, J., van Roessel, L., & Maalej, W. (2017). Making sense of user comments. Identifying journalists' requirements for a comment analysis framework. *SCM Studies in Media and Communication*, *7*(4), 333–364.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., . . . Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2/3), 93–118.

Meltzer, K. (2015). Journalistic concern about uncivil political talk in digital news media: Responsibility, credibility, and academic influence. *The International Journal of Press/Politics*, *20*(1), 85–107.

Moosa, T. (2014, September 12). Comment sections are poison: Handle with care or remove them. *The Guardian*. Retrieved from www.theguardian.com/science/brain-flapping/2014/sep/12/comment-sections-toxic-moderation

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, *67*(4), 586–609.

Nielsen, C. (2012). Newspaper journalists support online comments. *Newspaper Research Journal*, *33*(1), 86–100.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(2), 259–283.

Pöyhtäri, R., Con, N., In, M., Bassi, P., & Bretagna, E. G. (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*, *5*(4), 513–524.

Prochazka, F., & Schweiger, W. (2016). Media criticism online. Allegations and criticism towards news media

in user comments. *SCM Studies in Media and Communication*, *5*(4), 454–469.

Quandt, T. (2018). Dark participation. *Media and Communication*, *6*(4), 36–48.

Quandt, T., Frischlich, L., Boberg, S., & Schatto-Eckrodt, T. (in press). Fake news. In T. P. Vos & F. Hanusch (Eds.), *The international encyclopedia of journalism studies*. Malden: Wiley-Blackwell.

Reese, S. D., & Shoemaker, P. J. (2016). A media sociology for the networked public sphere: The hierarchy of influences model. *Mass Communication and Society*, *19*(4), 389–410.

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

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2016). Measuring the reliability of hate speech annotations: The case of the European refugee crisis. In *NLP4CMC III: 3rd workshop on natural language processing for computer-mediated communication, September 2016* (pp. 6–9). Bochum: Sprachwissenschaftliches Institut.

Singer, J. B., Hermida, A., Domingo, D., Heinonen, A., Paulussen, S., Quandt, T., & Vujnovic, M. (2011). Introduction. In J. B. Singer, A. Hermida, D. Domingo, A. Heinonen, S. Paulussen, T. Quandt, & M. Vujnovic (Eds.), *Participatory journalism: Guarding open gates at online newspapers* (pp. 1–9). Sussex: Wiley Subscription Services, Inc.

SPON. (2018). Das SPON-Forum. So wollen wir debattieren [The SPON forum. This is how we want to debate]. *SPON*. Retrieved from www.spiegel.de/extra/spon-forum-so-wollen-wir-debattieren-a-1032920.html

Statista. (2018). Anzahl der Unique User von Spiegel Online von Februar 2017 bis Februar 2018 [Number of unique user of Spiegel Online between February 2017 and February 2018]. *Statista*. Retrieved from de.statista.com/statistik/daten/studie/324186/umfrage/besucher-von-spiegel-online

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news": A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153.

Thorson, K., & Wells, C. (2015). How gatekeeping still matters: Understanding media effects in an era of curated flows. In T. P. Vos & F. Heinderyckx (Eds.), *Gatekeeping in transition* (pp. 25–44). New York, NY: Routledge.

Vos, T. P. (2015). Revisiting gatekeeping theory during a time of transition. In T. P. Vos & F. Heinderyckx (Eds.), *Gatekeeping in transition* (pp. 3–24). New York, NY: Routledge.

Walther, J. B., & Jang, J. (2012). Communication processes in participatory websites. *Journal of Computer-Mediated Communication*, *18*(1), 2–15.

Weber, R., & Popova, L. (2012). Testing equivalence in communication research: Theory and application. *Communication Methods and Measures*, *6*(3), 190–213.

White, D. M. (1950). The "gate keeper": A case study in the selection of news. *Journalism Quarterly*, *27*(4), 383–390.

Williams, B. A., & Delli Carpini, M. X. (2000). Unchained reaction: The collapse of media gatekeeping and the Clinton-Lewinsky scandal. *Journalism*, *1*(1), 61–85.

Ziegele, M., Breiner, T., & Quiring, O. (2014). What creates interactivity in online news discussions? An exploratory analysis of discussion factors in user comments on news items. *Journal of Communication*, *64*(6), 1111–1138.

About the Authors

Svenja Boberg is a PhD student at the University of Muenster. Since 2016 she is a Research Associate in the BMBF-funded project "Identification, Detection and Countering of Propaganda Disseminations Via Online Media", mainly focusing on using computational methods to analyze user comments. Her dissertation deals with the articulation and spread of political anger via social network sites.

Tim Schatto-Eckrodt is a PhD student and communication scientist, who researches methods to identify attempts to influence public opinion through online-propaganda. Since October 2017 he is a member of the chair of online communication of the Department of Communication of the Westfälische Wilhelms-Universität Münster, Germany. His master's thesis dealt with diffusion and the acceptance of new technologies.

Lena Frischlich studied Psychology at the University of Cologne and completed her PhD in social and media psychology in 2016. She has joined the Institute for Communication Science (chair of Online Communication) at Münster University in 2016, working as a Post-Doc in the PropStop project on detection, identification and combating of covert propaganda-attacks via online media. In January 2018, she has started her own junior research group on democratic resilience in times of online-propaganda, fake news, fear- and hate-speech (DemoRESILdigital).

Thorsten Quandt is a Professor of Online Communication at the University of Münster, Germany. His research fields include online communication, digital games and (online) journalism. Quandt is particularly interested in the societal changes connected to the Internet and new media, and the question how human beings have evolved in sync with these changes. His earlier works on participatory journalism and online newsroom production have been widely cited in the field of (digital) journalism research.