Article

Social Science Computer Review
2023, Vol. 41(3) 728–747

DOI: 10.1177/08944393211052042
journals.sagepub.com/home/ssc


Gender-Related Differences in Online Comment Sections: Findings From a Large-Scale Content Analysis of Commenting Behavior

Constanze Küchler[1], Anke Stoll[2], Marc Ziegele[2], and Teresa K. Naab[1]

Abstract

Comment sections below news articles are public fora in which potentially everyone can engage in equal and fair discussions on political and social issues. Yet, empirical studies have reported that many comment sections are spaces of selective participation, discrimination, and verbal abuse. The current study complements these findings by analyzing gender-related differences in participation and incivility. It uses a sample of 303,342 user comments from 14 German news media Facebook pages. We compare participation rates of female and male users as well as associations between the users' gender, the incivility of their comments, and the incivility of the adjacent replies. To determine the incivility of the comments, we developed a Supervised Machine Learning Model (classifier) using pre-trained word embeddings and word// frequency features. The findings show that, overall, women participate less than men. Comments written by female authors are more civil than comments written by male authors. Women's comments do not receive more uncivil replies than men's comments and women are not punished disproportionately for communicating uncivilly. These findings contribute to the discourse on gender-related differences in online comment sections and provide insights into the dynamics of online discussions.

Keywords

online discussions, user comments, incivility, gender, discrimination, machine learning, word embeddings, automated content analysis, multilevel modeling

[1]Department for Media, Knowledge and Communication, University of Augsburg, Augsburg, Germany
[2]Department of Social Sciences, Junior Research Group 'Deliberative Discussions in the Social Web' (DEDIS), Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Corresponding Author:
Constanze Küchler, Department for Media, Knowledge and Communication, University of Augsburg, Universitätsstraße 10, Augsburg 86159, Germany.
Email: constanze.kuechler@uni-a.de

Many citizens access news on Facebook and comment on news articles (Newman et al., 2017; Pew Research Center, 2017). While some scholars hoped that social media could foster inclusive and civil public online discussions (Dahlberg, 2001; Ruiz et al., 2011), research has taught us differently: comment sections are spaces of selective participation and prone to incivility (Chen, 2017; Coe et al., 2014; Muddiman & Stroud, 2017; Rowe, 2015).

The current study adds to these findings by focusing on gender-related differences in online discussions in comment sections. It considers two aspects: (1) the participation rates of female and male users and (2) incivility by and against male and female users, that is, their communication of incivility and their likelihood of receiving uncivil feedback. Thereby, the study addresses two gaps of existing research. First, previous studies on public (political) participation of women in online comment sections have often relied on self-report data (e.g., Bergström & Wadbring, 2015; Diakopoulos & Naaman, 2011; Stroud et al., 2016; van Duyn et al., 2021; for exceptions, see(Baek et al., 2021) Baek et al., 2021; Lee & Ryu, 2019; Vochocová et al., 2016). Self-reported behavior, however, often differs from actual behavior (e.g., Vochocová et al., 2016). The present study, therefore, uses data from a large-scale content analysis of actual discussions to examine whether a gender-related gap in the participation rates of female and male users exists in online comment sections. Second, only few studies have investigated gender-related differences regarding the communication of incivility and the likelihood of becoming a target of uncivil communication (e.g., Rheault et al., 2019). In fact, most of these studies have examined persons of public interest, such as politicians, journalists, or popular YouTubers (e.g., (Chen et al., 2020); Döring & Mohseni, 2020; Rheault et al., 2019; for an exception see Nadim & Fladmoe, 2021). The present study adds insights into how female laypersons in online comment sections are affected by incivility and whether they use incivility themselves.

The current study relies on data from a large-scale content analysis of 303,342 user comments on 14 German news media Facebook pages. We investigate whether male and female commenters differ regarding their participation rates in comment sections, whether they differ in their use of uncivil communication, and whether they face the same degree of uncivil reactions from other users. Data collection and analysis were conducted with the help of computational methods. To determine the incivility of user comments, we trained an *Incivility Classifier*, which achieved an acceptable accuracy of .68. The gender of comment authors was assigned automatically by matching usernames with dictionaries of male and female first names. To account for the hierarchical structure of comments and replies, we used multilevel modeling. Using this innovative combination of approaches, we contribute to the discourse on gender-specific differences in online comment sections and complement existing research on the dynamics of online discussions.

Gender-Related Differences in Online Discussions

Gender and Participation Rates

Gender-related differences in comment sections can be approached with a view on the participation rates of men and women. Discussions in comment sections are public and often about political topics (e.g., Stroud et al., 2016; van Duyn et al., 2021). Therefore, research on political deliberation (e.g., Polletta & Chen, 2013; Vochocová et al., 2016) can be used to derive hypotheses about the associations between gender and participation rates in comment sections. This research suggests that women are less likely than men to participate in political deliberation because they were socialized to avoid these discussions or because they were, for a long time, restricted from political spaces in general (van Duyn et al., 2021). Additionally, historically, women were blamed to lack the resources to successfully conduct political deliberation (Polletta & Chen, 2013). Some research has also linked women's lower engagement in political deliberation to the psychological

trait of conflict avoidance (Ulbig & Funk, 1999) or to their compliance to gender stereotypes that still prevail (see next sections). Ultimately, recent research has found that female politicians who are highly visible on social media platforms are more likely to be attacked uncivilly than their male colleagues (Rheault et al., 2019). From a social learning perspective (e.g., Bandura, 1969), female users could learn from these observations that engaging in public political discussions is harmful and, therefore, they refrain from participating.

In line with these arguments, most previous studies examining participation in online comment sections have found that more male users than female users write comments. These findings are largely consistent across different platforms (e.g., websites of news media outlets, Stroud et al., 2016; comment sections on Facebook sites of political parties, Vochocová et al., 2016; comment sections in general, van Duyn et al., 2021) and countries (Czech Republic: Vochocová et al., 2016; Korea: Baek et al., 2021; Lee & Ryu, 2019; USA: Stroud et al., 2016; van Duyn et al., 2021). The findings also align with research on offline political discussions, which has shown that women contribute less than men in terms of speech participation (Karpowitz et al., 2012).

One study, however, has found that more women than men reported to comment on news on social networking sites (Kalogeropoulos et al., 2017). In contrast, men were more likely to write comments on news websites. Yet, the authors only measured the frequency but not the type of commenting (e.g., private vs. public). It might well be that women are more active in certain online behaviors (e.g., commenting on photos or status updates shared by their network; Junco, 2013), but are less active in public commenting on political content (Lee & Ryu, 2019; Vochocovà et al., 2016).

In sum, the theoretical arguments and most empirical studies on participation behavior in public comment sections suggest that men are more active than women in terms of participation rates. Therefore, we can derive the following hypotheses:

H1: (1) More men than women participate in comment sections on Facebook pages of news outlets and (2) men contribute more comments than women.

Gender and Incivility

Gender-related differences in online discussions can also be approached as differences in the communication behavior of female and male users and as different reactions of other users to their behavior. In this paper, we focus on the communication of incivility. Definitions of incivility range from incivility as rhetorical and stylistic elements such as insulting vocabulary, ad hominem attacks, or verbal intimidation (Coe et al., 2014) to incivility as a set of behaviors that "threaten democracy, deny people their personal freedoms, and stereotype social groups" (Papacharissi, 2004, p. 267). Uncivil behaviors then include racism, sexism, attacking people for belonging to social or ethnic groups, or threatening other individuals' rights (Kalch & Naab, 2017; Papacharissi, 2004).

Empirical research particularly investigating incivility against women follows the aforementioned definitions to a greater (e.g., Southern & Harmer, 2021) or lesser extent (e.g., Nadim & Fladmoe, 2021; Rheault et al., 2019). For the current study, we follow the definition of Coe et al. (2014) who understand incivility as "*features of discussion that convey an unnecessarily disrespectful tone toward the discussion forum, its participants, or its topics"* (Coe et al., 2014, p. 660, italics in original), including name-calling, aspersion, lying, vulgarity, and pejorative for speech. We chose to adhere to this definition because (a) it is a comprehensive definition that offers clear indicators for uncivil expressions, and (b) it is an established definition of incivility and will, therefore, enable us to compare our results with existing research.

We use this definition to investigate incivility as a gender-related difference in comment sections along two aspects, namely, the communicators (i.e., do women communicate less uncivilly than men?) and the objects of incivility (i.e., are women targeted more often by incivility than men?). Additionally, we analyze a combination of these two aspects, namely, whether women are punished with uncivil responses disproportionally stronger than men when they communicate uncivilly.

Uncivil Communicators. Studies on computer-mediated communication have shown that communication styles vary between male and female online users (Herring & Stoerger, 2014; Kapidzic & Herring, 2011; Park et al., 2016; Thelwall et al., 2010). These studies indicate that female and male users can be identified by their way of communicating even in largely anonymous online environments because they discuss and express themselves differently (Herring & Stoerger, 2014; Kapidzic & Herring, 2011). Accordingly, women tend to communicate more positive emotions compared to men (Thelwall et al., 2010). Similarly, Park et al. (2016) found that female Facebook users communicate politely and more warmly (e.g. positive emotion). In contrast, male Facebook users used more impersonal, cold, assertive language (e.g., swearing, criticism). Additionally, studies have found that men are more likely than women to engage in trolling behavior (Buckels et al., 2014; Craker & March, 2016). Trolling includes deliberate behavior to provoke others, insults as well as bullying (March & Marrington, 2019)). Based on these definitions, trolling can be considered a form of online incivility (Coe et al., 2014).

Studies often do not provide a theoretical explanation for these differences (e.g., Kapidzic & Herring, 2011; Montgomery et al., 2004), but some research has referred to Social Role Theory (e.g., Wilhelm & Joeckel, 2019). One implication of Social Role Theory is that people expect others to behave in line with their gender stereotypes (Eagly, 2013; Eagly & Wood, 2012). Gender stereotypes attribute domestic, subordinate, and communal behaviors to women, while men are considered dominant and agentic (Eagly & Wood, 2012). Translated to communication behavior, Social Role Theory would predict that due to still-existing gender stereotypes, women communicate more warmly and less aggressively than men.

In sum, these findings and arguments suggest that women and men show different online communication behavior. Based on our understanding of incivility (Coe et al., 2014), incivility manifests in a disrespectful and often aggressive tone that aligns more with the (stereotypical) communication behavior of male than female commenters. We therefore hypothesize:

H2: More uncivil comments will be written by male users than by female users.

Uncivil Reactions. The question of whether the same messages are treated differently when they are communicated by men or women is an important subject of research on gender-related differences in online harassment (e.g., (Chen et al., 2020); Gardiner et al., 2016; Rheault et al., 2019; Rossini et al., 2021; Ward & McLoughlin, 2020; Wotanis & McMillan, 2014). Chen et al. (2020), for example, have argued that gendered harassment is "a particular aspect of online incivility" (p. 879). Based on in-depth interviews, the authors reported that female journalists regularly feel harassed and actively restricted to do their job properly (Chen et al., 2020). A quantitative content analysis of user comments posted on the website of the British *The Guardian* has shown that more uncivil comments are posted below the articles written by female journalists than below the articles written by their male colleagues (Gardiner et al., 2016). Similarly, a study by Southern and Harmer (2021) found that although male Members of Parliament (MPs) in the UK overall receive a higher number of uncivil tweets than female MPs (see also Ward & McLoughlin, 2020), female MPs are more likely than their male colleagues to receive at least one uncivil tweet. Ultimately, analyses of YouTube comments found that female YouTubers receive

more hostile feedback than male YouTubers (Döring & Mohseni, 2020; Wotanis & McMillan, 2014).

Social Role Theory and the so-called backlash effect can help to explain these findings (Wilhelm & Joeckel, 2019). According to this research, violating gender stereotypes can produce dissonance in communication partners and lead to a backlash effect (Rheault et al., 2019). The backlash effect has originally been researched in professional environments. Findings are that women are misjudged on the job or in hiring situations when showing counter-stereotypic behavior, such as agentic or aggressive communication (Eagly & Karau, 2002; Rudman & Glick, 1999, 2001). Not only women face stereotyping or discrimination due to gender—men also do (Davison & Burke, 2000). Yet, negative consequences due to the ascription of gender stereotypes are more likely to occur for women than for men ((Chen et al., 2020); Wotanis & McMillan, 2014). Similarly, a backlash effect can occur in the field of online journalism and online discussions. It can manifest in, among others, negative evaluations (i.e., flagging of online comments), social isolation, and even harassment ((Chen et al., 2020); Wilhelm & Joeckel, 2019).

In the current paper, we focus on a potential backlash effect that manifests in uncivil reactions to messages that are posted by male or female users in online discussions in comment sections. Social Role Theory and the backlash effect suggest two assumptions: First, given that discussions in comment sections are often political in their nature (see previous sections) and given that politics are still often considered a predominantly male arena (e.g., Rheault et al., 2019; Schneider & Bos, 2019), it can be assumed that women who publicly voice their opinion on politics will face more uncivil feedback than men. Second, this effect could be particularly strong when female users themselves engage in uncivil communication. In this case, women would violate two gender stereotypes, namely, the expectation not to (overly) engage in the field of politics *and* the expectation of communicating in a domestic, subordinate, and communal way (Eagly & Wood, 2012). Wilhelm and Joeckel (2019) have argued that such behavior could be considered "an act of double deviance" (p. 384). In line with the backlash effect, (Chen et al., 2020) found that female journalists receive uncivil reactions especially when reporting about "male topics." Moreover, in an online experiment (which, however, focused on flagging behavior as an operationalization of backlash), Wilhelm and Joeckel (2019) demonstrated that online users perceive it as disproportionately negative when women compared to men write counter-speech or hate-speech comments, that is, when they violate existing gender norms. The participants were more likely to sanction female users compared to male users for such comments (Wilhelm & Joeckel, 2019).

Still, studies have predominantly reported these effects for persons who are in the spotlight, such as politicians (Rheault et al., 2019; Southern & Harmer, 2021; Rossini et al., 2021; Ward & McLoughlin, 2020), journalists (Chen et al., 2020; Gardiner et al., 2016), or popular YouTubers (Döring & Mohseni, 2020; Wotanis & McMillan, 2014). Some evidence exists that the less well-known a person is, the smaller inequalities between men and women could be: Southern and Harmer (2021, p. 263, highlights in original), for example, investigated "'ordinary', or less high profile" MPs[1] in the UK and found only little differences between male and female MPs. Similarly, Rheault et al. (2019) found that gender-related differences in uncivil reactions to the tweets of politicians existed only for highly visible politicians. Finally, a survey focusing on laypeople suggested that "online harassment does not appear as specifically a 'woman problem'" (Nadim & Fladmoe, 2021, p. 255) because men also reported to face incivility. Explanations for a less pronounced backlash effect for laypeople could be that others perceive the behavior of laypeople as less norm-violating and as less threatening to the "status quo" than the behavior of well-known persons whose public actions are often considered exemplary for a society.

Reviewing the previous section, it is fair to say that the empirical evidence regarding online incivility against women in general is inconclusive. This applies even more to the research on

incivility against women engaging in comment sections. Social Role Theory and the backlash effect suggest that female users of these predominantly political spaces will face uncivil feedback when they participate in general, and in particular when they use uncivil communication themselves. However, previous empirical research has also suggested that the differences regarding uncivil reactions may diminish among laypeople. As the inconclusive state of research prevents deriving clear hypotheses, we ask the following research questions:

RQ1: Will comments written by female users receive more uncivil replies than comments written by male users?

RQ2: Will the backlash effect investigated in RQ1 be particularly strong when female users post uncivil comments?

Method

Sample

To answer the research questions and test the hypotheses, this study used a dataset of 303,342 user comments below news articles on 14 German Facebook news pages.[2] We collected posts and comments via Facebook Graph API. First, we collected all posts published between July and August 2018 on the 14 Facebook pages. The articles covered a broad range of topics (e.g., sports, politics, etc.). During the collection period, no specific events took place that would bias the news coverage. From this corpus of posts, we drew a sample along the following criteria: we only considered posts with at least 60 comments, and we only included original posts on the Facebook pages of the news outlets, which linked to a respective article on the news outlets' websites (i.e., we excluded shared posts from other Facebook pages that appeared in the feed of the selected news page). This procedure resulted in a total of 792 news posts that were included in our sample. Due to the selection criteria for news pages (i.e., national news pages) and posts (i.e., number of comments) it is very likely that most of the posts included in our analyses report about political topics although we did not specifically concentrate on topics as a selection criterion.

In a second step, we collected all online discussions that were linked to the selected posts and included them in the analyses. The comment sections on Facebook are organized hierarchically in *comments* and *replies*. Replies are displayed (in chronological order) below each comment. In total, the dataset contains 139,830 comments and 163,512 corresponding replies.

Identifying Gender from Usernames

Since Facebook urges its users to register with their real names, their gender can be inferred from the author's username in most cases. For our analyses, we automatically determined gender from the first name of each user. We used the python-package *gender-guesser*,[3] which works with data bases of international first names. Names can be categorized as either (mostly) male, (mostly) female, or ambiguous (androgynous). To validate the automated measurement, we manually evaluated a random subsample of 500 comments. Coding (male, female, androgynous, unknown) was done by two coders. Subsequently, we conducted a qualitative error analysis to investigate if and when human and algorithmic gender coding differed significantly. We found that the gender assignment was accurate in 99% of the cases in which users stated their names in the order "first name" followed by "surname". Users who intentionally misspelled their names on purpose (e.g., "Tho Mas") or used an alias (e.g., "Cutie Pie") were categorized as "unknown" by the gender guesser (7% of the sample of 500 comments), whereas human coders often could identify

misspelled names (e.g., "Tho Mas" for Thomas) and consequently the gender of a user. Further, we found that gender was assigned correctly by the algorithm for both German and non-German first names, whereas human coders needed further research to assign the gender correctly to non-German names. Interestingly, misclassification was therefore mainly caused by the lack of knowledge of the human coders, not the gender guesser. Based on the sample, we found no significant evidence for a bias in the gender measurement in favor of male or female users. Also, aliases and misspellings occurred both among female and male users.

In our analyses, we included all users whose names were automatically determined as either female, male, mostly female, or mostly male. In sum, 93,171 comments (including replies) were assigned to female authors (31%), and 179,422 comments were assigned to male authors (59%). For 30,749 comments, the author's gender could not be assigned (10%). Those comments were excluded from the analyses, resulting in a total of $n = 272{,}593$ comments that were considered for further analyses.

Classifying Incivility in User Comments

To measure incivility in a user comment, we applied a Supervised Machine Learning model (classifier) that classifies user comments automatically into *uncivil* and *civil*. A classifier is a statistical model that predicts a certain output (e.g., *incivility*) given a certain input (e.g., comment text). It is necessary to train the classifier on a dataset that includes information on both the input and the output. Accordingly, user comments in the training dataset must be labeled (manually) as uncivil or civil. After successful training, the classifier is then applied on a large dataset of unlabeled text to test the hypotheses (in our case: $n = 272{,}593$). To apply a pre-trained classifier to a new dataset, both datasets should be as comparable as possible. In the following, we describe the training dataset and the approach of the incivility classifier in more detail.

*Training Data***.** The incivility classifier (see next subchapter) was trained on a sample of 10,114 hand-coded German Facebook comments that had been collected as part of an earlier content analysis (Stoll et al., 2020). This subsample was drawn from a corpus of more than 1,000,000 comments from Facebook pages of nine German news media that were collected in 2016 via Facebook Graph API. The hand-coded subsample included comments and replies from the nine news media outlets as well as comments on various topics and from various stages of the discussions. It therefore offers a solid basis for identifying online incivility in a broad variety of contexts.

For every comment, coders rated the level of incivility on a three-point scale (0 = civil, 1 = slightly uncivil, 2 = predominantly uncivil). The scale was adapted from the incivility measure by Coe et al. (2014). That is, a comment was coded as uncivil when it included name-calling, aspersion, lying, vulgarity, or pejorative for speech. Contrary to the procedure by Coe et al. (2014), we did not code every manifestation of incivility separately, but rather coded a comment as uncivil when it included at least one of the incivility-related characteristics. Inter-coder reliability was tested on a sample of 100 comments and reached a satisfactory level of Krippendorff's $\alpha = .83$.

A difficulty is that uncivil user comments occur less frequently than civil comments in the training data. When coding along the described three-point scale, there are very few comments per incivility-level (i.e., slightly or predominantly uncivil). In consequence, the classification results were unstable and rather unsuitable to distinguish between all three levels of incivility accurately, especially between slightly and predominantly. Therefore, we aggregated slightly and predominantly uncivil to uncivil. By doing so, we work with a dichotomous incivility measure (0 = *civil*, 1 = *uncivil*) in the current analysis. Within the training data, the category *civil* was assigned 6678 times (67%) and the category *uncivil* was assigned 3294 times (33%).

Model Building with Word Embeddings and Bag-of-Words. To predict the outcome variable incivility, we tested several classification approaches (including features constellations and Deep Learning architectures) that vary regarding complexity, performance, and costs of both calculation and requirements. Finally, we chose a combination of different text-based features (independent variables) including a) word frequency distributions (*Bag of Words, BoW*) and b) *Word Embeddings.* In the BoW-approach, the occurrence of a term (e.g., a word, phrase, or an emoji) is used to predict the text category (e.g., incivility). However, the BoW representation leads to high variance and in most cases, the training data are not sufficient to learn all conceivable terms for a text category (e.g., incivility). The word embedding approach addresses this problem. A word embedding is a vector of a word that can be described as a relative "location" of a word in a *n* dimensional vector space. To "arrange" the words in the vector space, words of an extensive corpus of documents are fed into a *Neural Network* and are mapped ("embedded") into lower dimensional vector representations (word embeddings). In this representation, words that are used in similar contexts (i.e., occur with the same words) have similar vectors. For example, words that are used synonymously, such as "super" and "awesome," often end up with similar word embedding vectors. In contrast, words that are used in different contexts (i.e., are surrounded by different words), such as "super" and "hello," have different word embedding vectors (they are more distant from each other in vector space). In contrast to the BoW approach, the word embedding approach does not consider the frequency of words to assign documents to a category but uses distance in vector space. This kind of text representation has proven to be more accurate than word frequencies for many text classification tasks (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013; Pennington et al., 2014; Young et al., 2018; Wiegand et al., 2018).

Word embeddings must be trained separately on enormous amounts of text documents. Therefore, researchers often use pre-trained word embeddings. Word embeddings should be trained on a dataset that is comparable to the dataset that will be classified. For classification tasks on non-English and non-formal language texts, such as online discussions, few pre-trained models are available. For our sample of German user comments, we used 300-dimensional *fasttext* word embeddings (Bojanowski et al., 2017) that have been trained on a corpus of 50 million German tweets (Cieliebak et al., 2017; Deriu et al., 2017). It turned out to be the most appropriate word embedding model for our classification task. To use these word embeddings as independent variables (features) for incivility classification, they must be transformed from a representation on word-level into a representation on document-level. Since every word in a comment has its own 300-dimensional word vector, we averaged the word embedding vectors for all words in a comment to one 300-dimensional vector per comment (Pérez & Luque, 2019). This way, words and characters that are not included in the pre-trained embeddings are ignored. To ensure that all relevant words are considered, we additionally used frequency distributions of single words and word combinations (BoW) to predict the incivility of a comment. For the BoW representation, the comment messages were transformed into weighted frequencies of unigrams, bigrams, and trigrams ((Stoll et al., 2020); Risch & Krestel, 2018).

Model Results and Evaluation. To train the model, we tested different classification functions, including *Support Vector Machines*, *Logistic Regression,* and *Decision Trees*. To get a more robust result, we applied 5-*fold cross validation*, so that all reported evaluation scores are averages of five model runs. Since we will use the predicted categories in the later analysis of the study, it is important that the model does not learn the distribution of incivility in the training data as predictor. Therefore, we applied the oversampling algorithm *SMOTE* (*Synthetic Minority Oversampling Technique*, Chawla et al., 2002) to the training data. This is a common and proven technique in Machine Learning to balance the distribution of the categories. This way, oversampling can prevent the classification model from choosing the more frequent category in case of missing information (for further detail see (Stoll, 2020), Haixiang et al., 2017). The best results

were achieved by a linear Support Vector Machine with accuracy = 0.68 and macro-F1 = 0.61 ($F1_{civil} = 0.76$, $F1_{uncivil} = 0.46$, $Recall_{civil} = 0.80$, $Recall_{uncivil} = 0.42$). This indicates that in 68% of the cases, the model classifies a comment correctly as uncivil or not uncivil. To test our hypotheses and answer the research question, we trained the classifier again on the whole training dataset and applied it to the new set of user comments that was not labeled manually. On this new dataset, the classifier predicted the category *civil* for 164,200 comments (60%) and the category uncivil for 108,393 comments (40%). Based on the performance values for the two classes, we can assume that uncivil comments are identified reliably, but civil comments tend to be misclassified as uncivil. Therefore, absolute (or descriptive) values for civil and uncivil comments need to be interpreted with caution. Still, group comparisons between male and female users—which are the focus of our study—remain valid since the reliability of the incivility measurement does not vary by gender.

Multilevel Modeling with Logistic Regression

Comment sections on Facebook are organized hierarchically in comments and replies. RQ1 and RQ2 investigate whether the incivility of a reply depends on the gender of the comment's author and on the interplay between a comment author's gender and incivility. More precisely, we assumed that the incivility of a reply can be predicted by incivility of the comment and gender of its author. To consider these structural dependencies, we conducted multilevel analyses that model the incivility of a reply on level 1 as a function of level 2 (comment) gender (*RQ1*) and incivility by gender (*RQ2*). Additionally, since we analyze comments below 792 news posts, we introduced these news articles as a third level of analysis. Since we use both, binary dependent and independent variables, we applied multilevel logistic regression models using the R-Package *lme4* (Bates et al., 2012).

Results

A total of 94,923 identifiable individual users were involved in the discussions. Thereof, 59,781 users were identified as male (63%) and 35,142 users were identified as female (37%). Consequently, of all users whose gender could be identified, more than 1.7 times as many men than women were active in the online discussions. Furthermore, an average male user wrote more comments than an average female user (male: $M = 3.00$, $SD = 7.64$; female: $M = 2.65$, $SD = 7.92$, $t = 6.66$, $df = 71{,}489$, $p < .001$). These findings support *H1,* which assumes that online discussions in comment sections are dominated by male authors indicated by the facts that (1) more men than women participate and (2) an individual man, on average, writes more comments than an individual woman.

H2 assumed that the comments written by female and male users differ regarding their incivility. To test this hypothesis, we computed a simple cross-table on the full dataset (we excluded comments by authors whose gender was not identifiable), including all 122,822 comments and all 149,771 replies. Forty-two percent of the comments written by male users contained incivility. In contrast, only 35% of the comments written by female users were uncivil. The relationship between gender and incivility thus points in the expected direction, although it is quite weak ($\chi^2(1) = 1239.08$, $p < .001$; $\varphi = 0.07$). These findings support *H2*.

To investigate *RQ1* and *RQ2*, we conducted multilevel logistic regression models that describe the incivility of a reply as a function of the related comment's incivility, gender of its author, and the interaction term of incivility and gender. Thus, for this analysis, we only considered comments that received at least one reply. This reduced the dataset to $n = 27{,}922$ comments and $n = 128{,}557$ replies. Table A1 shows the results of the multilevel logistic regressions. The intraclass correlation coefficient (ICC) shows that incivility in a reply is explained to 12% by the groups it is nested in

(i.e., the related comment and news article). In the empty model (null model), the odds ratio for $y = 1$ is $OR = 0.56$, $p < .001$, meaning the overall chance of a reply i over all groups j being uncivil ($y = 1$) is smaller than being civil ($y = 0$).

RQ1 asked whether comments written by female authors will be more likely to receive uncivil replies than comments written by male authors. To answer this question, we ran a model including the level-2 variable "gender of the comment author" (model 1). We also added the variable "incivility of the comment" as a control variable. Results show that the chance of a reply being uncivil *decreases* when the comment is written by a female author ($OR = 0.88$, $p < .001$). This means that contrary to the implications of Social Role Theory, female users are *less likely* than men to receive uncivil feedback.

RQ2 asked whether women, as compared to men, will be disproportionately sanctioned for commenting in an uncivil manner. To investigate this question, we added the interaction term of level-2 incivility and gender (model 2). Results show that the interaction term is not significant ($OR = 1.04$, $p = .21$). The interaction diagram (see Figure A1) shows that the incivility of the replies to uncivil comments does not differ depending on whether the comment is written by a male or female user. In contrast, independent from a comment author's gender, uncivil comments stimulated more uncivil replies ($OR = 1.36$, $p < .001$). In summary, these findings suggest that female comment authors are not more likely to receive uncivil reactions than male authors, neither to civil nor to uncivil comments.

Discussion

Since user comments on Facebook pages of news outlets have become a popular form of online participation, it is important to investigate gender-related differences in comment sections. This study has examined these differences from several perspectives. First, we asked for differences in women's and men's participation rates in user comment sections regarding both, the shares of women and men participating and the numbers of contributions by female and male individuals. Second, we analyzed whether comments authored by female and male users differ in their incivility. Third, we looked at differences regarding the incivility of the replies to the comments written by female and male users. More specifically, we asked whether women receive more uncivil replies than men. In addition, we investigated whether women are disproportionately sanctioned with uncivil reactions for writing uncivil comments themselves. In doing so, this study is one of the first to analyze gender-related differences on how laypersons communicate and receive incivility in comment sections. We tested our hypotheses on a large dataset of online discussions on German news media's Facebook pages. We automatically determined the incivility of the comments and the gender of the commenters. To analyze dependencies between gender and incivility in comments and replies, we conducted multilevel logistic regression models.

The results show that fewer women than men participated in online discussions. Additionally, women, on average, wrote fewer comments than men. This suggests that discussions in comment sections are indeed dominated by male users. The results contradict the findings from a survey that reported that women and men participate to the same extent in comment sections (e.g., Kalogeropoulos et al., 2017). It is possible that female users report a general willingness to write comments that equals the willingness of male users. However, the results of our study suggest that when it comes to actual commenting behavior, women are less inclined to engage to the same extent as men in discussions on (political) news in online comment sections. These results support the findings of earlier studies on political participation (Karpowitz et al., 2012; Polletta & Chen, 2013) and the findings of the few available content analyses of gender-related participation in comment sections (Baek et al., 2021; Lee & Ryu, 2019; Vochocová et al., 2016). The findings also

raise the question for future research on what could be done to create public spaces that are equally inviting for people of all genders, to actively reduce traditional hierarchies (van Duyn et al., 2021; Polletta & Chen, 2013), and to make women more comfortable to discuss potentially controversial topics (Ulbig & Funk, 1999).

Additionally, we found weak gender-related differences in the use of incivility within comments. Male commenters, on average, wrote slightly more uncivil comments than women. These findings are in line with previous research on different communication styles of men and women (e.g., Thelwall et al., 2010). Still, 35% of the comments written by female users were uncivil, which suggests that many women participating in comment sections do not adapt to the stereotypic expectation of behaving warm and subordinate. Future research could investigate whether this is a result of a gender-specific online disinhibition effect (Suler, 2004) or of fading gender stereotypes in general.

Regarding gender-related differences in the reactions to comments written by female or male authors, women's comments did not receive more uncivil replies than comments written by men. This finding contradicts the expectations of Social Role Theory and the backlash effect and supports research suggesting that gender-related backlashes could be less pronounced for female laypeople than for females who are in the spotlight, such as journalists, MPs, and YouTubers (e.g., (Rheault et al., 2019); Southern & Harmer, 2021). For the domain of comment sections, the current study supports the assumption that incivility against lay females might be less of a severe problem compared to well-known or high profile people. A reason may be that users responding to well-known authors have more background knowledge on the authors, are more aware of their gender, and actively attribute social roles. These processes could be responsible for a stronger backlash when well-known authors seemingly behave in contradiction with users' expectations. When interacting with other laypersons, users may, however, not necessarily pay equal attention to the authors' profile names and thus their gender.

Our findings also show that uncivil commenting leads to more uncivil replies. This supports previous studies that found a "vicious circle of incivility," which means that uncivil comments trigger further incivility in the subsequent discussions. Interestingly, our results show that this effect is independent of the commenters' gender. Here, our findings contradict a previous study that found small but significant "double backlash effects" against women in online discussions (Wilhelm & Joeckel, 2019). While the different results might be explained by the fact that this previous study used a different operationalization of backlash (i.e., flagging behavior instead of written comments) and relied on experimental settings instead of content analyses, our study still leaves room for more detailed investigations (see further below).

Overall, our findings provide some evidence for gender-related differences in participation rates, but limited evidence for differences in communication behavior and a gender-specific backlash. In terms of participation rates, our study has only investigated whether participation rates are equal. However, our data does not give insights into whether or how incivility affects the future commenting behavior of users (e.g., throughout a comment section or even across several comment sections), especially for female commenters. Future studies should address this in more detail. Although comments written by female authors are underrepresented in our dataset, men do neither overly dominate the discussions in terms of uncivil communication behavior nor are women's comments more likely to be attacked uncivilly. These findings should certainly not suggest that gendered harassment against individual women of interest or groups of women in the spotlight is not a relevant problem in online discussions (the opposite is shown in other studies, e.g., (Chen et al., 2020)Rheault et al., 2019; Southern & Harmer, 2021). Yet, our findings do not suggest that such attacks are generally more frequently directed at women than at men. One explanation, again, could be that we investigated incivility toward "average" commenters who do not possess certain expertise, status, or power. Such characteristics might make authors,

particularly females, prone to uncivil attacks (Rheault et al., 2019). Future studies should address this limitation and investigate the differences between laypersons and persons of interest to gain a deeper understanding of gendered incivility.

A second explanation could be that we did not distinguish types of gendered incivility. Studies have already shown that female and male MPs experience different types of harassment. Female MPs suffer from gender-based stereotyping, whereas male MPs face incivility due to professional aspects, such as party affiliation or political stances (Southern & Harmer, 2021; Ward & McLoughlin, 2020). Further, compared to men, women are subject to "more sexist, racist, or sexually aggressive hate comments" (Döring & Mohseni, 2020, p. 73). They are more objectified due to gender and physical appearance, while receiving less supportive feedback on their content (Döring & Mohseni, 2020; Wotanis & McMillan, 2014). Accordingly, it is the task of future studies to disentangle different types of gendered incivility and reinvestigate the relationship between gender and types of incivility in more detail.

In sum, our findings provide important implications for research on gender-related differences in online discussions regarding participation and communication behavior in online comment sections. More specifically, our results suggest that users primarily reply to "what" is said in comments instead of replying to "who" has said it. This would be an important prerequisite for inclusive online discussions in which women can equally contribute their perspectives. This interpretation is supported by research showing that users evaluate persuasive messages online (such as user-generated product reviews) mainly based on the arguments and rhetorical and stylistic devices (Willemsen et al., 2011). However, various studies have also shown that users consider the identity-related disclosures of online communicators when judging their messages (Forman et al., 2008). Thus, future research needs to untwine the conditions under which users do or do not consider identity-related cues of messages when evaluating their content and responding to it.

Limitations

Along with new insights that add to existing research on gender-related differences in online communication, this study has several limitations: First, we did not code for news topics discussed in the posts below which users commented. However, our sampling procedure (e.g., including only posts that received at least 60 comments) increased the probability that we predominantly included political, and, therefore, "male" topics, which often attract high numbers of comments (e.g., Coe et al., 2014). This might explain the gender-related gap between the participation rates of female and male users; in fact, van Duyn and colleagues (2021) showed that women and men comment on different types of topics, that is, women are more likely to comment on local news and men are more likely to comment on national or international news. Despite the high probability that our sample included many political topics, we did not find that women are disproportionately sanctioned with uncivil feedback for voicing their opinion on these topics. Having said that, future studies should investigate in more detail the relations between participation rates and communication behavior of female and male users by controlling for topics during analyses.

A second limitation is that our dataset has not been collected in real-time. This means that exceptionally uncivil comments may already have been filtered out before we collected the data. Consequently, extreme forms of incivility, which might have affected the results, could not be analyzed. This is a problem of many content analyses and although we do not assume that this limitation overly biased our results, future research should consider collaborating with news outlets or platform providers to gain insights into deleted or filtered comments as well.

Third, it is unclear in our dataset who is addressed in an uncivil reply. Replies can possibly contain a high level of incivility against the author of a comment, but also against a third person (e.g., a politician) or against an issue. Additionally, in Facebook comment sections, all replies are subordinate to the related comment, yet there are no further sub-levels of responses to replies. However, replies may well address a preceding reply within the thread instead of the related comment. Future studies need to take this into consideration and measure the object of incivility more comprehensively. Further, by studying incivility in replies to comments, our research employed a very specific operationalization of the gender backlash. Future studies may consider investigating additional manifestations of backlash, such as flagging behavior (Wilhelm & Joeckel, 2019) or being ignored. In content analyses, the latter form of backlash could possibly be investigated by analyzing the number of replies that female and male comment authors receive.

Finally, to conduct our analyses on a large sample, we decided to determine the incivility of a comment automatically, applying a supervised machine learning approach (classifier). The classifier was trained on a comparable dataset of 10,000 user comments from an earlier content analysis. The model achieved 68% accuracy in classifying user comments as either civil or uncivil. This means that about 32% of the comments have been misclassified. Incivility often is a matter of personal perception (Chen, 2017), and some forms of incivility can only be deduced from context. It remains challenging for machines and even for humans to determine incivility based on text patterns. To provide the best measurement of incivility, we tested several classification approaches that differ regarding their complexity, expense, and performance, including different feature constellations and Deep Learning approaches (i.e., LSTM model architecture). Results showed that complex Deep Learning architectures clearly overfit the training data. This led to higher model performance but, at the same time, resulted in unreliable measurement on new data. We therefore chose the model described since it ensures a reliable identification (recall) of uncivil user comments. Model performance scores showed that the classifier identified most uncivil comments ($Recall_{uncivil} = 0.80$) but tended to misclassify many civil comments as uncivil, too. This must be kept in mind when interpreting the findings of the study. Still, the group comparisons between male and female users should remain valid since the reliability of the incivility measurement does not differ by gender. Nevertheless, future studies that are mainly interested in interpreting descriptive or absolute values may consider using a larger sample of manually labeled training data that assures lower heterogeneity in most cases and, therefore, leads to better model results.

It is also important to emphasize that we only used a binary gender classification. We did not consider, for example, androgynous names, such as "Dominique" or "Charly." Additionally, based on the applied automated measurement of gender, users' first names were categorized as (mostly) female or (mostly) male. According to our manual evaluation of the gender guesser algorithm on 500 comments, this procedure was not biased by any systematic misclassification. Yet, some users register with aliases or intentionally misspell their names (e.g., "Tho Mas" or "Cutie Pie"). The gender guesser classified these names as unknown, but they might have been correctly assigned by human coders in some cases. Future research should consider using refined methods to be able to investigate these cases, too.

Lastly, we do not know whether the users who were classified as female or male also identify as women or men. Even more so, authors might intentionally choose an ambiguous name or one of another gender category to avoid gendered harassment. There is a long history of research on identity deception in computer-mediated communication. It manifests as identity concealment, attractiveness deception, or category deception (Utz, 2005), with the last one referring to a switch in gender. Motivations behind identity deception online can be privacy concerns, status elevation, idealized self-presentation, or identity play, among others (Caspi & Gorsky, 2006;

Utz, 2005). Having said that, we are fully aware of the gender variety in our society. Nonetheless, not only most existing research, but also international first name lists are mostly based on a binary understanding of gender. It is up to future studies to offer innovative solutions for this limitation.

Conclusion

Commenting on news articles on Facebook pages of news media is currently considered one of the most popular forms of public user participation. Previous research has argued that the discussions in comment sections could foster equally accessible and respectful exchange. However, various studies reported that a high share of incivility threatens the discussion atmosphere in comment sections and drives readers away from reading and writing comments. Yet, only few studies have used content analyses to investigate whether users' gender is related to their commenting behavior and whether the comments written by female and male users are differently likely to trigger uncivil responses. The present research used a large dataset of comments from various news outlets to investigate these questions. The results suggest that gender-related differences in comment sections are mainly related to unequal participation rates and commenting frequencies of male and female users. We did not find striking differences regarding gender-based (un-)civil communication behavior or regarding gendered abuse through uncivil replies to women. These findings extend previous research on gender inequalities that often used self-reports or experimental designs with a limited number of comments or focused on well-known authors of user-generated content. We hope that our research will stimulate further studies that disentangle when and why women and men are treated (un)equally in comment sections.

Software Information

Python-Package gender-guesser: https://github.com/lead-ratings/gender-guesser

300-dimensional fasttext word embeddings: The pre-trained Tweet embeddings are provided under the Creative Commons License CC BY 4.0 by Spinningbytes: https://www.spinningbytes.com/resources/wordembeddings/(accessed June 19, 2020).

Weighted frequencies of unigrams, bigrams, and trigrams: We used Tf-idf weighting (Term frequency-inverse document frequency weighting), which is the standard weighting metric in natural language processing.

Removal of stopwords: We used the NLTK Stopword Corpus for German. Full documentation: https://www.nltk.org/book/ch02.html

Removal of long sequences of letters or punctuation to a maximum of three consecutive characters: We applied the TweetTokenizer from the nltk. tokenize package. Full documentation: https://www.nltk.org/api/nltk.tokenize.html

Support Vector Machine: We applied the Support Vector Classifier SVC from the scikit-learn python package (Pedregosa et al., 2011). Full documentation: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

Statistical analysis: R-package lme4, For full documentation: https://www.rdocumentation.org/packages/lme4/versions/1.1-23.

Appendix

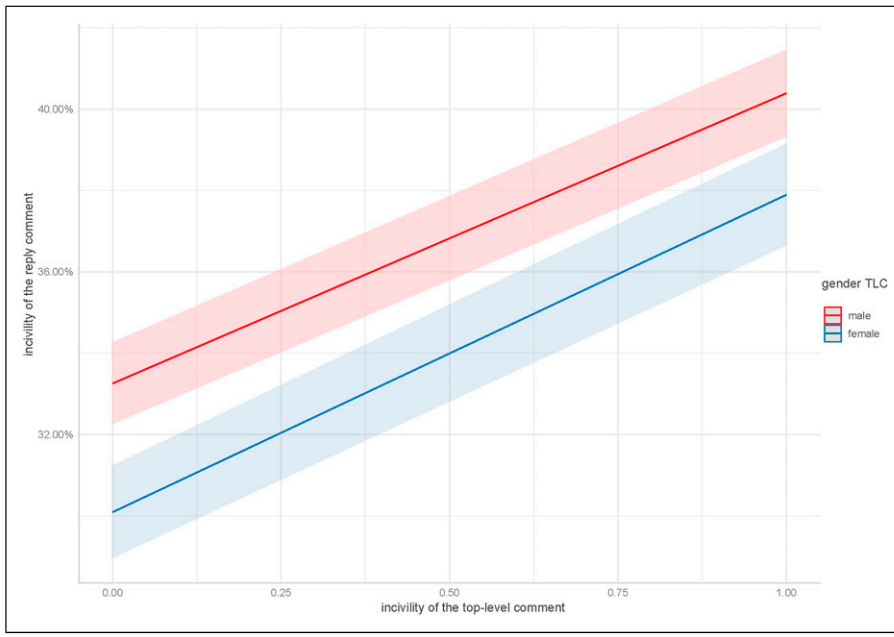

Figure A1. Incivility of a reply (Level one; 0 = *civil*, 1 = *uncivil*) predicted by comment Gender*Incivility (level two; 1 = *male*, 2 = *female*); *N*reply = 133,941; *N*comment = 29,352.

Table A1. Model Overview – Associations Between Comment Gender/Incivility and Incivility of Replies.

	Null Model	Model 1	Model 2
Predictors (comment, level two)	OR	OR	OR
Intercept	0.56***	0.49***	0.50***
Gender (1 = *female*)		0.88***	0.86***
Incivility (1 = *uncivil*)		1.38***	1.36***
Gender * incivility			1.04
Random effects			
σ^2	3.29	3.29	3.29
$\tau 00_{Article}$	.30	.25	.24
$\tau 00_{comment}$	.16	.16	.11
ICC	.12	.11	.11
Approx. R^2	.11	.12	.12
AIC	166562.674	166109.563	166110.008
Model improvement (Chi^2)		457.11***	1.56
n (Articles)	792	792	792
n (comment)	27,922	27,922	27,922
n (Reply)	128,557	128,557	128,557

Notes. Dependent Variable: Incivility of reply (level one; 0 = *civil*, 1 = *uncivil*); OR = odds ratio; *N*reply = 133,941; *K*comment = 29,352; α = 0.05, ***p < .001, **p < .01, *p < 0.5; Full ML.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Ministry of Culture and Science of the German State of North Rhine-Westphalia and Deutsche Forschungsgemeinschaft, Grant NA 1281/1-1, project number 358324049.

ORCID iDs

Constanze Küchler https://orcid.org/0000-0002-9406-144X
Marc Ziegele https://orcid.org/0000-0002-2710-0955
Teresa K. Naab https://orcid.org/0000-0001-7345-2559

Notes

1. For Southern & Harmer (2021), high profile MPs are those with the most followers on Twitter, for example, former prime minister Theresa May. In this study they focus on the less high profile or ordinary MPs, that is, on those having less followers on Twitter and therefore excluded the 50 most followed MPs.
2. We selected the news pages according to ranking lists of the most used German news pages (Newman et al., 2017; Schröder, 2016). We selected from the top 16 news pages of both lists (since the list of Newman et al. only consisted of 16 news pages). News pages were considered for our study when they maintained a website and a respective FB page, allowed commenting on those sites, and had their own in-house news production.
3. Package gender-guesser version 0.4.0. Full documentation: https://pypi.org/project/gender-guesser/ (09.09.2020).

References

Baek, H., Lee, S., & Kim, S. (2021). Are female users equally active? An empirical study of the gender imbalance in Korean online news commenting. *Telematics and Informatics*, *62*, 101635. https://doi.org/10.1016/j.tele.2021.101635

Bandura, A. (1969). Social learning of moral judgments. *Journal of Personality and Social Psychology*, *11*(3), 275-279. https://doi.org/10.1037/h0026998

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., & Krivitsky, P. N. (2012). *Package 'lme4'*. Vienna, Austria: CRAN. R Foundation for Statistical Computing.

Bergström, A., & Wadbring, I. (2015). Beneficial yet crappy: Journalists and audiences on obstacles and opportunities in reader comments. *European Journal of Communication*, *30*(2), 137-151. https://doi.org/10.1177/0267323114559378

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146. https://doi.org/10.1162/tacl_a_00051

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, *67*, 97-102. https://doi.org/10.1016/j.paid.2014.01.016

Caspi, A., & Gorsky, P. (2006). Online deception: Prevalence, motivation, and emotion. *CyberPsychology & Behavior*, *9*(1), 54-59. https://doi.org/10.1089/cpb.2006.9.54

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357. DOI:https://doi.org/10.1613/jair.953.

Chen, G. M. (2017). *Online incivility and public debate: Nasty talk*. Cham, Switzerland: Palgrave Macmillan.

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

Cieliebak, M., Deriu, J. M., Egger, D., & Uzdilli, F. (2017). A twitter corpus and benchmark resources for german sentiment analysis. Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, (45-51). https://doi.org/10.18653/v1/W17-1106

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

Craker, N., & March, E. (2016). The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences*, *102*, 79-84. https://doi.org/10.1016/j.paid.2016.06.043

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

Davison, H. K., & Burke, M. J. (2000). Sex discrimination in simulated employment contexts: A meta-analytic investigation. *Journal of Vocational Behavior*, *56*(2), 225-248. https://doi.org/10.1006/jvbe.1999.1711

Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., Hofmann, T., & Jaggi, M. (2017). Leveraging large amounts of weakly supervised data for multi-language sentiment classification. Proceedings of the 26th international conference on world wide web, (1045-1052). https://doi.org//10.1145/3038912.3052611

Diakopoulos, N. A., & Naaman, M. (2011). Towards quality discourse in online news comments. CSCW '11 Proceedings of the ACM 2011 conference on Computer supported cooperative work, New York, 133-142. https://doi.org/10.1145/1958824.1958844

Döring, N., & Mohseni, M. R. (2020). Gendered hate speech in YouTube and YouNow comments: Results of two content analyses. *Studies in Communication and Media*, *9*(1), 62-88. https://doi.org/10.5771/2192-4007-2020-1-62

Eagly, A. H. (2013). *Sex differences in social behavior: A social-role interpretation*. New York: Psychology Press.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573-598. https://doi.org/10.1037/0033-295X.109.3.573

Eagly, A. H., & Wood, W. (2012). Social role theory. In P. A. M. Van Lange, A. W. Kruglanski, & E.T. Higgins (Eds.), *Handbook of Theories in Social Psychology,* (Vol. 2, pp. 458-476). Los Angeles: Sage Publications. https://doi.org/10.4135/9781446249222.n49

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, *19*(3), 291-313. https://doi.org/10.1287/isre.1080.0193

Friess, D., Ziegele, M., & Heinbach, D. (2021). Collective Civic Moderation for Deliberation? Exploring the Links between Citizens' Organized Engagement in Comment Sections and the Deliberative Quality of Online Discussions. *Political Communication*, 1-23. https://doi.org/10.1080/10584609.2020.1830322.

Gardiner, B., Mansfield, M., Anderson, I., Holder, J., Louter, D., & Ulmanu, M. (2016). *The dark side of Guardian comments*. The Guardian. https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-com

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220-239. https://doi.org/10.1016/j.eswa.2016.12.035

Herring, S. C., & Stoerger, S. (2014). Gender and (a)nonymity in computer-mediated communication. In S. Ehrlich, M. Meyerhoff, & J. Holmes (Eds.), *The handbook of language, gender, and sexuality*, (Vol. 2, pp. 567-586). Chichester, West Sussex: Wiley Blackwell. https://doi.org/10.1002/9781118584248.ch29

Junco, R. (2013). Inequalities in Facebook use. *Computers in Human Behavior*, *29*(6), 2328-2336. https://doi.org/10.1016/j.chb.2013.05.005

Kalch, A., & Naab, T. K. (2017). Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. *SCM Studies in Communication and Media*, *6*(4), 395-419. https://doi.org/10.5771/2192-4007-2017-4

Kalogeropoulos, A., Negredo, S., Picone, I., & Nielsen, R. K. (2017). Who shares and comments on news? A cross-national comparative analysis of online and social media participation. *Social Media+ Society*, *3*(4), 1-12. https://doi.org/10.1177/2056305117735754

Kapidzic, S., & Herring, S. C. (2011). Gender, communication, and self-presentation in teen chatrooms revisited: Have patterns changed? *Journal of Computer-Mediated Communication*, *17*(1), 39-59. https://doi.org/10.1111/j.1083-6101.2011.01561.x

Karpowitz, C. F., Mendelberg, T., & Shaker, L. (2012). Gender inequality in deliberative participation. *American Political Science Review*, *106*(3), 533-547. https://doi.org/10.1017/S0003055412000329

Lee, S. Y., & Ryu, M. H. (2019). Exploring characteristics of online news comments and commenters with machine learning approaches. *Telematics and Informatics*, *43*, 101249. https://doi.org/10.1016/j.tele.2019.101249

March, E., & Marrington, J. (2019). A Qualitative Analysis of Internet Trolling. *Cyberpsychology, Behavior, and Social Networking*, *22*(3), 192-197. DOI:https://doi.org/10.1089/cyber.2018.0210.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. Conference: Proceedings of the International Conference on Learning Representations (ICLR 2013), arXiv preprint arXiv:1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Montgomery, K., Kane, K., & Vance, C. M. (2004). Accounting for differences in norms of respect: A study of assessments of incivility through the lenses of race and gender. *Group & Organization Management*, *29*(2), 248-268. https://doi.org/10.1177/1059601103252105

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, *67*(4), 586-609. https://doi.org/10.1111/jcom.12312

Nadim, M., & Fladmoe, A. (2021). Silencing women? Gender and online harassment. *Social Science Computer Review*, *39*(2), 245-258. https://doi.org/10.1177/0894439319865518

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A., & Nielsen, R. K. (2017). *Reuters Institute digital news report 2017*. Oxford: Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/Digital%20News%20Report%202017%20web_0.pdf

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(6), 259-283. https://doi.org/10.1177/1461444804041444

Park, G., Yaden, D. B., Schwartz, H. A., Kern, M. L., Eichstaedt, J. C., Kosinski, M., Stillwell, D., Ungar, L. H., & Seligman, M. E. P. (2016). Women are warmer but no less assertive than men: Gender and language on facebook. *PLOS ONE*, *11*(5), e0155885. https://doi.org/10.1371/journal.pone.0155885

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, M., Weiss, R., Dubourg, V., Vanderplas, J., Passos, J., Cournapeau, D., Brucher, M., Perrot, M., &

Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*(85), 2825-2830.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing EMNLP, 1532-1543.

Pérez, J. M., & Luque, F. M. (2019). Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification. Proceedings of the 13th International Workshop on Semantic Evaluation, 64–69.

Pew Research Center (2017). *More Americans are turning to multiple social media sites for news*. https://www.pewresearch.org/fact-tank/2017/11/02/more-americans-are-turning-to-multiple-social-media-sites-for-news/

Polletta, F., & Chen, P. C. B. (2013). Gender and public talk: Accounting for women's variable participation in the public sphere. *Sociological Theory*, *31*(4), 291-317. https://doi.org/10.1177/0735275113515172

Rheault, L., Rayment, E., & Musulan, A. (2019). Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics*, *6*(1), 1-7. https://doi.org/10.1177/2053168018816228

Risch, J., & Krestel, R. (2018). *Aggression identification using deep learning and data augmentation.* Paper presented at the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Santa Fe, NM. https://www.aclweb.org/anthology/volumes/W18-44/#abstract-W18-4418

Rossini, P., Stromer-Galley, J., & Zhang, F. (2021). Exploring the relationship between campaign discourse on facebook and the public's comments: A case study of incivility during the 2016 US presidential election. *Political Studies*, *69*(1), 89-107. https://doi.org/10.1177/0032321719890818

Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, Communication & Society*, *18*(2), 121-138. https://doi.org/10.1080/1369118X.2014.940365

Rudman, L. A., & Glick, P. (1999). Feminized management and backlash toward agentic women: the hidden costs to women of a kinder, gentler image of middle managers. *Journal of Personality and Social Psychology*, *77*(5), 1004-1010. https://doi.org/10.1037//0022-3514.77.5.1004

Rudman, L. A., & Glick, P. (2001). Prescriptive gender stereotypes and backlash toward agentic women. *Journal of Social Issues*, *57*(4), 743-762. https://doi.org/10.1111/0022-4537.00239

Ruiz, C., Domingo, D., Micó, J. L., Díaz-Noci, J., Meso, K., & Masip, P. (2011). Public sphere 2.0? The democratic qualities of citizen debates in online newspapers. *The International Journal of Press/politics*, *16*(4), 463-487. https://doi.org/10.1177/1940161211415849

Schneider, M. C., & Bos, A. L. (2019). The application of social role theory to the study of gender in politics. *Political Psychology*, *40*, 173-213.

Schröder, J. (2016). *IVW-News-Top-50: Januar bringt Rekorde für Focus, Welt, FAZ und viele andere*. https://meedia.de/2016/02/09/ivw-news-top-50-januar-bringt-rekorde-fuer-focus-welt-faz-und-viele-andere/

Southern, R., & Harmer, E. (2021). Twitter, incivility and "everyday" gendered othering: An analysis of tweets sent to UK members of parliament. *Social Science Computer Review*, *39*(2), 259-275. https://doi.org/10.1177/0894439319865519

Stoll, A. (2020). Supervised Machine Learning mit Nutzergenerierten Inhalten: Oversampling für nicht balancierte Trainingsdaten. *Publizistik*, *65*(2), 233-251. DOI:https://doi.org/10.1007/s11616-020-00573-9.

Stoll, A., Ziegele, M., & Quiring, O. (2020). Detecting impoliteness and incivility in online discussions: Classification approaches for German user comments. *Computational Communication Research*, *2*(1), 109-134. DOI:https://doi.org/10.5117/CCR2020.1.005.KATH.

Stroud, N. J., van Duyn, E., & Peacock, C. (2016). *News commenters and news comment readers*. https://mediaengagement.org/wp-content/uploads/2016/03/ENP-News-Commenters-and-Comment-Readers1.pdf

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, *7*(3), 321-326.

Thelwall, M., Wilkinson, D., & Uppal, S. (2010). Data mining emotion in social network communication: Gender differences in MySpace. *Journal of the American Society for Information Science and Technology*, *61*(1), 190-199.

Ulbig, S. G., & Funk, C. L. (1999). Conflict avoidance and political participation. Political Behavior. *Political Behavior*, *21*(3), 265-282. https://doi.org/10.1023/A:1022087617514

Utz, S. (2005). Types of deception and underlying motivation: What people think. *Social Science Computer Review*, *23*(1), 49-56. https://doi.org/10.1177/0894439304271534

Van Duyn, E., Peacock, C., & Stroud, N. J. (2021). The gender gap in online news comment sections. *Social Science Computer Review*, *39*(2), 181-196. https://doi.org/10.1177/0894439319864876

Vochocová, L., Štětka, V., & Mazák, J. (2016). Good girls don't comment on politics? Gendered character of online political participation in the Czech Republic. *Information, Communication & Society*, *19*(10), 1321-1339. https://doi.org/10.1080/1369118X.2015.1088881

Ward, S., & McLoughlin, L. (2020). Turds, traitors and tossers: The abuse of UK MPs via Twitter. *The Journal of Legislative Studies*, *26*(1), 47-73. https://doi.org/10.1080/13572334.2020.1730502

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 shared task on the identification of offensive language. In 14th Conference on Natural Language Processing KONVENS 2018, pp. 1-10.

Wilhelm, C., & Joeckel, S. (2019). Gendered morality and backlash effects in online discussions: An experimental study on how users respond to hate speech comments against women and sexual minorities. *Sex Roles*, *80*(7–8), 381-392. https://doi.org/10.1007/s11199-018-0941-5

Willemsen, L. M., Neijens, P. C., Bronner, F., & de Ridder, J. A. (2011). "Highly recommended!" the content characteristics and perceived usefulness of online consumer reviews. *Journal of Computer-Mediated Communication*, *17*(1), 19-38. https://doi.org/10.1111/j.1083-6101.2011.01551.x

Wotanis, L., & McMillan, L. (2014). Performing gender on YouTube: How Jenna Marbles negotiates a hostile online environment. *Feminist Media Studies*, *14*(6), 912-928. https://doi.org/10.1080/14680777.2014.882373

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe Magazine*, *13*(3), 55-75.

Ziegele, M., Weber, M., Quiring, O., & Breiner, T. (2018). The dynamics of online news discussions: Effects of news articles and reader comments on users' involvement, willingness to participate, and the civility of their contributions. *Information, Communication & Society*, *21*, 1419-1435. DOI:https://doi.org/10.1080/1369118X.2017.1324505.

Author Biographies

Constanze Küchler: Constanze Küchler is a research associate at the University of Augsburg, Germany. Her research interests include health communication, audience research and social science research methods.Anke Stoll:Anke Stoll is a research associate at the Institute for Social Sciences at the Heinrich Heine University in Düsseldorf, Germany. In the research projects DEDIS (Deliberative Discussions in the Social Web) and KOSMO (AI-Supported Collective-Social Moderation) she works on the automated identification of social media content and on the development of Artificial Intelligence software for moderation systems and participation platforms.

Prof. Dr. Marc Ziegele: Marc Ziegele is an Assistant Professor for Political Online Communication at the Department of Social Sciences at the University of Duesseldorf, Germany. In his research, he studies online incivility, deliberation, journalistic moderation, and media trust in the digital age.

PD Dr. Teresa K. Naab: PD Dr Teresa K Naab is senior researcher at the University of Augsburg, Germany. She received her Ph.D. from the University of Music, Drama and Media Hannover, Germany. Her research interests include digital communication, audience research, and social science methods.