Preprint version of

Marc Ziegele, Johannes Daxenberger, Oliver Quiring and Iryna Gurevych (2018): Developing Automated Measures to Predict Incivility in Public Online Discussions on the Facebook Sites of Established News Media.

Paper presented at the 68th Annual Conference of the International Communication Association (ICA).

Developing Automated Measures to Predict Incivility in Public Online Discussions on the Facebook Sites of Established News Media

Abstract

Incivility in public discourse on the Internet, e.g., in user comment sections on social network sites, is discussed as a significant problem for democratic societies. Due to the massive amount of conversation data, however, it is a challenge to code these discussions manually and make valid claims about the prevalence and effects of uncivil comments. We therefore trained an automated classifier on more than 10,000 hand-coded user comments and used it to predict incivility in user comments in an extensive dataset of more than one million user comments collected on the Facebook sites of nine German news media outlets over a three-months-period in 2015. Additionally, we use the classified data to analyze a) differences in user engagement with uncivil and civil comments, b) prevalence of incivility in reactive and interactive comments as well as differences regarding the share of uncivil comments across c) different news media and d) different topics.

Developing Automated Measures to Predict Incivility in Public Online Discussions on the Facebook Sites of Established News Media

Introduction

Public user comments on the websites and Facebook pages of established news media are a popular element of digital online communication (Stroud et al., 2015). Many news consumers use the comment sections to learn about the issue-related attitudes of other users or to voice their own opinion towards news topics or other user comments (Rowe, 2015; Springer et al., 2015). For German online users, reading comments is now almost as widespread as reading printed newspapers (Ziegele, Köhler, & Weber, 2017): 41 percent of German onliners read user comments at least once a week, and 50 percent read printed newspapers on a regular basis. 30 percent contribute own comments at least once a month. Other research shows similar data for countries such as the U.S. (Newman et al., 2016).

Reading and writing comments hence is widespread in today's online media landscape. But what do the comments that users read and write look like? The quality of user comments is often assessed through the lens of deliberation theory (e.g., Manosevitch & Walker, 2009; Rowe, 2015; Ruiz et al., 2011). The deliberation framework sketches a public sphere that can be accessed by everyone, and in which citizens discuss social and political issues in a rational, reciprocal, and respectful manner (Gastil, 2008). Ideally, these discussions result in an "agreement based on the best argument" (Ruiz et al., 2011, p. 478) which provides the most promising solution for the issue at hand (Dryzek, 2004). From this theoretical perspective, comment sections could be a promising forum for an open, non-discriminatory and constructive exchange of opinions between citizens on socially significant topics. Like other forms of online deliberation, they could support users' political knowledge, promote political tolerance, and contribute to their political and civil engagement (Friess & Eilders, 2015).

However, researchers, journalists, and politicians have raised concerns regarding the quality of user comments. Many comments are not constructive, respectful, and result-oriented. Rather, they include a high degree of so-called *incivility* (Coe et al., 2014). Incivility is defined as the "expression of disagreement by denying and disrespecting the justice of the opposing views" (Hwang et al., 2016, p. 5). In user comments, incivility encompasses elements such as verbal intimidation, ad hominem attacks, overgeneralizations, pejorative speech, as well as racism and hateful language (Coe et al., 2014; Ziegele & Jost, 2016).

Uncivil comments are problematic because they can undermine democratic values and lead to attitude polarization (Anderson et al., 2014). Moreover, they increase aggressive cognitions and stereotypical attitudes among their readers, and have a negative impact on the perceived news quality of established news media (Hsueh et al., 2015, Prochazka et al., 2016). In the long run, overly uncivil discussions prevent users from writing comments and can make news organizations shut down their comment sections (Stroud et al., 2016, Ziegele, 2016).

Previous research has found that around 30 percent of user comments on various U.S. news sites include some degree of incivility (Coe et al., 2014; Rowe, 2015). However, the generalizability of these findings is heavily limited as the studies often rely on comment samples from single news sites and/or are based on small samples. To overcome these shortcomings, we are interested in developing an automated "incivility predictor" that can be used to reveal the prevalence of uncivil comments across various Facebook news sites and amongst entire discussion threads. Hence, our first research question is as follows:

RQ1: Can an automated classifier be trained to predict incivility in public user discussions on the Facebook sites of established news media?

Using the classifier, we then want to investigate various further research questions. First of all, previous research has shown that uncivil comments – despite their detrimental effects – can stimulate user engagement: For example, in a case study of the *New York Times*, Muddiman

and Stroud (2017) have shown that uncivil comments receive more "recommendations" from other users than civil comments. We aim at replicating this finding in a Facebook context and therefore ask:

RQ2: Will uncivil user comments receive more "Likes" than civil comments?

Additionally, users can post different types of comments on Facebook; they can post *reactive comments* in response to a news article or they can post *interactive comments* that respond to one or more comments from other users. Heated and uncivil comments might particularly occur in discussions where users respond to each other. Hence we will investigate whether the share of incivility differs between reactive and interactive comments:

RQ3: Does the share of incivility differ between reactive comments and interactive comments?

Finally, previous research has suggested that the prevalence of incivility varies across news topics and across news media outlets: Particularly political topics seem likely to attract uncivil discourse (Coe et al., 2014) and some media outlets "cultivate" communities where incivility appears to be an accepted norm (Ruiz et al., 2011). Hence we ask:

RQ4: Does the prevalence of incivility vary between different news media outlets?

RQ5: Does the prevalence of incivility vary between different topics?

Method

Sampling

Based on two criteria, nine Facebook sites from established German news media outlets were included in the sample: First, the sample was meant to include a broad variety of media genres. Hence, the Facebook sites of public broadcasters, private broadcasters, newspapers, news magazines, and the popular press were considered. From each media genre, we chose up to two national outlets that had the highest number of Facebook followers (second criterion). Following these criteria, the following Facebook news sites were included

in the sample: *Tagesschau*, *ZDF Heute*, *N24*, *RTL Aktuell*, *Die Welt*, *Sueddeutsche Zeitung*, *BILD*, *Spiegel Online*, and *ZEIT Online.*

For the sampling of news articles and user comments, we constructed three artificial weeks between May and August 2015. That is, for each month, we randomly defined seven different weekdays as access days. At 9AM in the morning of the day after an access day, the nine Facebook sites were accessed in random order and crawled for the preceding day's news articles and user comments with the help of the tool *netvizz*. The tool is free to use for research purposes and it is regularly updated to account for changes in Facebook's API. This procedure yielded a total of 27,728 news articles and 1,056,002 comments (*reactive* comments and *interactive* comments). For the hand-coding procedure, a sample of this data was drawn according to the following criteria: On each access day, three or four news articles were randomly selected from each news site in the sample (resulting in a total of 27 or 36 news articles per access day). We did not code the specific topics of the news articles, however, the links of these articles refer to specific topic resorts on the news websites. After deleting articles that did not cover any news, we arrived at a total of 619 articles with 118,058 comments. From the 118,058 comments, we again drew a sample of a maximum of 20 *reactive* and *interactive comments* per article. For each article, the oldest five comments, the most recent five comments, five random comments from the 'middle' of the discussion, and the five most popular comments were selected. This procedure resulted in a total of 10,170 comments.

The comments were randomly assigned to nine coders who were trained extensively regarding the use of the incivility measure. This measure relied on previous measures of incivility and included 'name-calling', 'aspersion', 'lying', 'vulgarity', and 'pejorative for speech' (Coe et al., 2014). For each comment, coders indicated whether the comment contained none of these forms of incivility ('0'), if it contained scattered incivility ('1') or if it was predominantly/exclusively uncivil ('2'). At the end of the training, a random sample of

100 user comments was given to all coders for reliability testing. Reliability of the incivility measure was assessed using Krippendorff's α (Krippendorff, 2004). The overall reliability of the incivility measure was satisfying (α = .81).

Results

RQ1: Automatically predicting incivility in user-generated comments

As mentioned above, we had 10,170 hand-coded comments which were used to train an "incivility predictor". Of those, 27% had been coded as containing some sort of incivility (18% "predominantly/exclusively incivil"). Our predictor had as input the comments themselves, the number of likes each comment received, and a three-class variable for the incivility of the comment (none, scattered, predominantly/exclusively uncivil). We processed the comment messages with OpenNLP Segmenter and Part-of-speech (POS) Tagger[1], and extracted the following information from each comment: the unigrams (words)[2] contained in the comments, two- to four-grams of POS tags, sequences of repeated punctuation (such as "!!!"), the total number of words, the number of likes the comment received, and the number of words corresponding to concepts extracted from the Linguistic Inquiry and Word Count (LIWC) such as "affect" or "anger" (Tausczik & Pennebaker, 2010). Altogether, we extracted more than 4,500 "features" for each comment.[3] As classifier, we used L2-regularized logistic regression, which is a solid predictor for many text classification tasks (e.g., Ferreira & Vlachos, 2016).

We evaluated the accuracy of this classifier using 5-fold cross-validation. Although the overall accuracy of this classifier was 72.4%, our preliminary results show that incivility prediction is a difficult task (macro-F1 score over three classes 0.44). This is mainly because the classifier does a bad job in recognizing uncivil comments: of the comments it labeled as

[1] https://opennlp.apache.org
[2] Only for the 1,500 most frequent words in the overall set of comments.
[3] An analysis of the information gain of the overall feature set reveals that among the top-5 ranked features we find LIWC-words related to "affect" and "swear", which intuitively seems correct.

uncivil, less than 40% are actually uncivil; furthermore, it thinks that most of those comments which actually contain incivility do not. It does, however, well in predicting comments which do not contain incivility (F1 score 0.84). Being fully aware of the deficits of this classifier, we trained it on the full set of 10,170 manually coded comments and carried out a preliminary experiment for automatically coding 1,045,832 comments which we had crawled but not manually coded. The resulting dataset contains 1,020,254 comments predicted as not containing incivility, 17,319 (2%) predicted as containing scattered incivility, and 8.740 (1%) as predominantly or exclusively uncivil.

RQ2: Received "Likes" of civil and uncivil comments

We used the automatically-coded incivility category to investigate the further research questions. Regarding the "popularity" of uncivil and civil comments, a one-way ANOVA revealed significant differences between the number of received "Likes" of uncivil and civil comments ($F(2, 1049025) = 105{,}940$, $p < .001$). Civil comments received an average of $M = 2.87$ Likes ($SD = 23.03$), while comments that included scattered incivility received an average of $M = 4.96$ Likes ($SD = 23.00$). Entirely uncivil comments received an average of $M = 5.00$ Likes ($SD = 34.17$). Games-Howell post hoc tests revealed that partly and entirely uncivil comments received more Likes than civil comments ($p < .001$) but there was no significant difference between the average number of received Likes of partly and entirely uncivil comments ($p = .99$).[4]

RQ3: Incivility in reactive comments and interactive comments

We cross-tabbed the distribution of incivility in user comments with the distribution of the comment type (reactive, interactive). 96.9% of reactive comments did not contain incivility compared to a slightly smaller share of 96.6% of interactive comments. Accordingly, 2.4% of interactive comments included scattered incivility compared to 1.9% of reactive comments.

[4] Results of an analysis with a log-transformation of this count data revealed similar patterns.

1% of reactive and interactive comments were predominantly uncivil. Although the difference between partly uncivil reactive comments and partly uncivil interactive comments is significant (Chi^2 (2) = 180.239, $p < .001$, $Phi = .02$) and supports our line of thought that incivility might particularly occur in heated debates between the users, the difference also appears somewhat negligible small.

RQ4: Incivility across different news outlets

A Chi² test (Chi^2 (16) = 5877.00, $p < .001$, $Phi = .08$) revealed significant differences of the prevalence of uncivil comments across the different news outlets. The share of partly or entirely uncivil comments was particularly high among the Facebook sites of public broadcasters (4.6%) and – interestingly – particularly low on the Facebook sites of the Tabloid press (1.3%). In a further analysis, we will investigate whether these findings are a result of different topic foci of the respective media outlets (e.g., tabloid press might publish fewer political topics on Facebook, and political topics are likely to attract uncivil commentary).

RQ5: Incivility across different news topics

This analysis is still in progress. We will either use automated topic analysis to identify topics in the news articles related to the user comments or analyze the links in the news articles to the specific topic resorts on the news websites. Then, we will analyze whether different topics will attract different shares of uncivil discourse.

Discussion

Incivility in public online discussions is considered a significant problem of democratic societies, and providing reliable evidence of the prevalence of uncivil discourse appears an important task for research. In this context, research collaborations between communication scientists and computer scientists, such as the one described here, are a critical endeavor. Our experiments provide a first step into this direction. Using a large dataset of hand-coded comments, we trained an automated “incivility predictor”, and, using the automatically

classified data, we could replicate some findings of previous studies. As a next step, we will improve the classification performance of our incivility predictor by applying more sophisticated deep learning algorithms (e.g., Kim, 2014). The results of these experiments will be presented at the conference in case of acceptance.

References

Anderson, A. A., Brossard, D., Scheufele, D. A., Xenos, M. A., & Ladwig, P. (2014). The "Nasty Effect:" Online Incivility and Risk Perceptions of Emerging Technologies. *Journal of Computer-Mediated Communication, 19*(3), 373–387.

Ferreira W. & Vlachos, A. (2016). Emergent: a novel dataset for stance classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1163–1168, San Diego, CA, USA.

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.

Dryzek, J. S. (2004). *Deliberative democracy and beyond: Liberals, critics, contestations*. *Oxford Scholarship Online*. New York: Oxford University Press.

Friess, D., & Eilders, C. (2015). A Systematic Review of Online Deliberation Research. *Policy & Internet, 7*(3), 319–339.

Gastil, J. (2008). *Political communication and deliberation*. Los Angeles: Sage.

Hwang, H., Kim, Y., & Kim, Y. (2016). Influence of Discussion Incivility on Deliberation: An Examination of the Mediating Role of Moral Indignation. *Communication Research.*

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave Your Comment Below": Can Biased Online Comments Influence Our Own Prejudicial Attitudes and Behaviors? *Human Communication Research, 41*(4), 557–576.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751.

Krippendorff, K. (2004). Reliability in Content Analysis. *Human Communication Research, 30*(3), 411–433.

Manosevitch, E. & Walker, D. M. (2009). *Reader Comments to Online Opinion Journalism: A Space of Public Deliberation: Paper presented at the 10th International Symposium on Online Journalism, Austin, TX, April 17-18*.

Muddiman, A., & Stroud, N. J. (2017). News Values, Cognitive Biases, and Partisan Incivility in Comment Sections. *Journal of Communication, 67*(4), 586–609.

Newman, N., Fletcher, R., Levy, D. A. L., & Nielsen, R. K. (2016). *Reuters Institute Digital News Report 2016*. Retrieved from https://reutersinstitute.politics.ox.ac.uk.

Prochazka, F., Weber, P., & Schweiger, W. (2016). Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism Studies,* 1–17.

Rowe, I. (2015). Deliberation 2.0: Comparing the Deliberative Quality of Online News User Comments Across Platforms. *Journal of Broadcasting & Electronic Media, 59*(4), 539–555.

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, 22,* 463–487.

Springer, N., Engelmann, I., & Pfaffinger, C. (2015). User comments: motives and inhibitors to write and read. *Information, Communication & Society, 18*(7), 798–815.

Stroud, N. J., Scacco, J. M., Muddiman, A., & Curry, A. L. (2015). Changing Deliberative Norms on News Organizations' Facebook Sites. *Journal of Computer-Mediated Communication, 20*(2), 188–203.

Stroud, N. J., van Duyn, E., & Peacock, C. (2016). *News Commenters and News Comment Readers*. Retrieved from http://engagingnewsproject.org.

Tausczik, Y.R. & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 29, 24-54.

Ziegele, M., Köhler, C., & Weber, M. (2017). *Socially destructive! Effects of hateful user comments on recipients' prosocial behavior.: Presentation at the 67th Annual Conference of the International Communication Association (ICA), May 25-29, San Diego, USA.*

Ziegele, M., & Jost, P. B. (2016). Not Funny?: The Effects of Factual Versus Sarcastic Journalistic Responses to Uncivil User Comments. *Communication Research,* 1–30.