Automatically Detecting Incivility in Online Discussions of News Media

Johannes Daxenberger*, Marc Ziegele†, Iryna Gurevych*, Oliver Quiring‡
*Ubiquitous Knowledge Processing (UKP) Lab, Technische Universität Darmstadt, Germany
http://www.ukp.tu-darmstadt.de
†Institut für Sozialwissenschaften, Heinrich-Heine-Universität Düsseldorf, Germany
‡Institut für Publizistik, Johannes Gutenberg-Universität Mainz, Germany

***Abstract*—Detecting biased language in written discourse is a highly relevant area of research in political communication and other social sciences, given the large quantity of information exchanged in public online platforms. In this abstract, we discuss an approach based on the concept of "incivility"—assessing biased text on the Facebook pages of established news media. News outlets are forced to put increasing efforts into preventing heated debates from turning into disrespectful discussions on their social media platforms. By scaling the analysis from a few thousand manually coded samples to more than a million comments, we take a step towards supporting media outlets in (semi-)automatizing the detection of uncivil comments and enable a much broader analysis of the latter.**

I. Introduction

Based on a close collaboration between social and computer science research, this project seeks to advance the understanding of incivility, i.e. "expressions of disagreement by denying and disrespecting [...] opposing views" [1]. In particular, we are interested in user comments on the Facebook pages of nine German public and private media outlets. We used machine learning techniques to train a system for incivility detection from 10,170 hand-coded comments [2], enabling the automatic classification of more than one million comments extracted from the same pages over a three-months-period in 2015. Beyond training and analyzing the model itself, research questions in this project include: i) will uncivil comments receive more "Likes" than other comments?, ii) does the share of incivility differ between reactive comments and interactive comments (the latter responding to one or more comments from other users)?, iii) does the type of news influence the prevalence of incivility?, and iv) does the prevalence of incivility vary between different news media outlets? To the best of our knowledge, this is the first project to address these questions on such a large and diverse sample of comments from online news media platforms.

II. Experiments

For experimenting with automatic incivility detection, we initially used a simple logistic regression classifier with lexical-semantic features as described in [2]. This classifier already outscored a baseline predicting the majority class ("no incivility") by a large margin—however, its overall performance was rather weak. As a step towards a more robust detection of uncivil comments, we further experimented with classifiers based on deep learning. State-of-the-art neural network classifiers for text classification typically use so called embeddings as input. Embeddings enable continuous representations of words or characters, encoding semantic and/or syntactic characteristics, e.g. based on their distributional properties. We initially tested with the Convolutional Neural Networks (CNN) proposed by Kim [3], with 128-dimensional word embeddings trained on the 10,170 hand-coded comments itself. Since this did not yield the desired improvements, we switched to the fastText system proposed by Joulin et al. [4], a simple but extremely fast text classification approach. Using the fastText system allowed us to test different word embeddings based on the approach proposed by Bojanowski et al. [5]. While word embeddings trained on articles from the English Wikipedia did not yield satisfying results due to the domain shift (encyclopedic articles vs. social media comments), classifying with 100-dimensional embeddings trained on all (one million) comments from our dataset outperformed all other results.

III. Results

With the approach described above and evaluating on 5-fold cross-validation, we achieved an overall accuracy of 75%, with a macro-F1 score over three classes (no, scattered, or predominant incivility) of 46%. We further trained the same classifier on only two classes (distinguishing comments containing any share of incivility and comments without incivility). For that case, the accuracy improves to 78% and the macro-F1 score to 68%, compared to a baseline of 42% for majority classification (all comments classified as containing no incivility). Given the class imbalance in the training data (only 27% of the 10,170 comments are coded as containing scattered or predominant incivility—thus, the classifier needs to learn incivility from less than 3,000 short text snippets), this is an encouraging result.

We then automatically classified all one million comments using this model. The result gave us a broad empirical foundation to answer the initially stated research questions. We found that (i) uncivil comments receive a significantly higher number of "Likes" as compared to comments without

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1816B (CEDIFOR).

incivility. This is a novel finding, given that previous work [6] did not find significant differences for this property. Further, (ii) interactive comments show a significantly higher degree of incivility as compared to reactive comments—a finding that even contradicts previous work [6]. With respect to iii), hard news (latest news from e.g. politics and economics) attract significantly more uncivil comments than soft news (news articles about e.g. arts and lifestyle). This difference was again significant and is in line with previous work [6]. Finally (iv), we found that public broadcasters (here: German television broadcasting companies "ARD" and "ZDF") attract significantly more uncivil comments than the Tabloid press (e.g. the German newspaper "Bild"). This finding is again novel.

Our study shows that automatic incivility detection in social media platforms is—to a certain extent—possible. With regard to future work, our study leaves open questions about the influence of cultural contexts or platform types, which might explain why some of our findings contradict previous research while others are consistent with previous findings.

REFERENCES

[1] H. Hwang, Y. Kim, and Y. Kim, "Influence of Discussion Incivility on Deliberation: An Examination of the Mediating Role of Moral Indignation," *Communication Research*, vol. 45, no. 2, pp. 213 – 240, 2016.

[2] M. Ziegele, J. Daxenberger, O. Quiring, and I. Gurevych, "Developing Automated Measures to Predict Incivility in Public Online Discussions on the Facebook Sites of Established News Media," 2018, Paper presented at the 68th Annual Conference of the International Communication Association (ICA).

[3] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1746–1751.

[4] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 427–431.

[5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[6] K. Coe, K. Kenski, and S. A. Rains, "Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comment," *Journal of Communication*, vol. 64, no. 4, pp. 658–679, 2014.