Improving discourse quality in online discussions: A large-scale, longitudinal study of influencing speech characteristics

Keywords: hate speech, counter speech, discourse quality, collective moderation, social media

Extended Abstract

Introduction and General Aim

Political discussions on social media are often overshadowed by hate speech. Uncivilized practices can keep people with moderate attitudes from participating in online discussions, distorting public opinion (Ziegele & Jost, 2020). Citizen-organized counter speech can be a means against hate speech on the Internet (Friess et al., 2021). While several counter speech strategies have been identified (Buerger, 2021), their effectiveness is still largely unknown. In this study, we investigate how effective a variety of speech characteristics are for mitigating online hate.

Methods

Our data set includes 130,127 conversations on Twitter in response to posts of prominent German news outlets, journalists and politicians from 2015 to 2018 with 1,150,469 tweets and 130,548 unique users in total. We used autoregressive distributed lag (ARDL) modelling to analyze effects of four speech characteristics on the presence of hate speech, toxicity, and political extremity. Speech characteristics include a speaker's argumentation strategy, the presence of in- and outgroup thinking, the social psychological goal with respect to those groups, and the presence of emotions. We estimated models with lags of one and two units (i) on the level of discussion trees containing at least 50 Tweets and (ii) on the level of days to examine both short- and long-term effects, respectively. We interpret the findings in light of the presence of organized hate and counter speech groups that were active at that time (i.e., Reconquista Germanica - RG, and Reconquista Internet - RI) (Garland et al., 2022). We used expert annotations for a variety of speech characteristics and outcomes to train several deep-learning models to detect these variables in our large corpus of Tweets. We further used pre-existing classifiers for emotional tone (Widmann & Wich, 2022), as well as for political extremity and toxicity (Garland et al., 2022; Jigsaw, n.d.).

Results

We show descriptive results for the development of speech characteristics and indicators of incivility over time (see Fig. 1). Political extremity, toxicity and hate show a steep upwards trend after the start of the migrant crisis, a less severe increase after the establishment of RG (hate speech group), and a downwards trend after the launch of RI (counter speech group). Interestingly, different independently trained classifiers identify similar trends (Fig. 1B). While we can observe greater hate and toxicity on both ends of the political spectrum, the left-leaning group RI still shows overall lower levels compared to the right-leaning group RG (Fig. 1C).

Opinions (not necessarily factual, but also not insulting) and insults were on the rise, while sarcasm and other constructive comments (providing information, posing questions, pointing out negative consequences of words or actions, correction somebody, and pointing out hypocrisy or contradictions) decreased (Fig. 1D). Over time, exclusionary mentions of the outgroup grew (Fig. 1E). Lastly, anger is by far the most dominating emotion during the entire study period (Fig. 1F).

Results for the effectiveness of different strategies against hate, toxicity and extremity are shown in Figure 2. Offering opinions and sarcasm positively influence discourse quality in the long run over discussion trees or days. The effect of sarcasm is more pronounced when RG is present. Other constructive comments can reduce hate and toxicity especially on the tree level, but increase extremity of speech, possibly because of backfire effects. Mixed effects might also be observed due to the variety of argumentation strategies included in this class. Comments putting the outgroup down, but also comments portraying the own or both groups in a positive light, increased hate, toxicity and extremity compared to neutral speech. While comments about the outgroup may stress group divide, comments about the ingroup might predominantly include posturing. Interestingly, we found very few instances of speech re-conciliating in- and outgroup, even to the extend that those instances were not detectable by our classifier. Similarly, emotionally charged comments, be it positive (e.g., pride) or negative (e.g., anger), decreased discourse quality in the long run, although effects are inconsistent. These counter-intuitive findings might be explained by positive emotions such as pride being used to stress the superiority of the own group.

Conclusion

Building on an unprecedented large, longitudinal data set, we provide new and nuanced insights into the effects of speech characteristics on discourse quality useful for individuals and groups wishing to reduce hate in online spaces. In particular, we investigated the effects of in- and outgroup thinking on hate speech, constituting a new angle in this line of research. Furthermore, we incorporated the presence of organized hate and counter speech groups in our analyses, which gives a more fine-grained understanding of collective civic moderation.

References

Buerger, C. (2021). Counterspeech: A literature review. *Available at SSRN 4066882*.

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*, *38*(5), 624–646. https://doi.org/10.1080/10584609.2020.1830322

Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ data science*, *11*(1), 3.

Jigsaw. (n.d.). Perspective api [Accessed: 2023-02-27].

Widmann, T., & Wich, M. (2022). Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in german political text. *Political Analysis*, 1–16. https://doi.org/10.1017/pan.2022.15

Ziegele, M., & Jost, P. B. (2020). Not funny? the effects of factual versus sarcastic journalistic responses to uncivil user comments. *Communication Research*, *47*(6), 891–920. https://doi.org/10.1177/0093650216671854

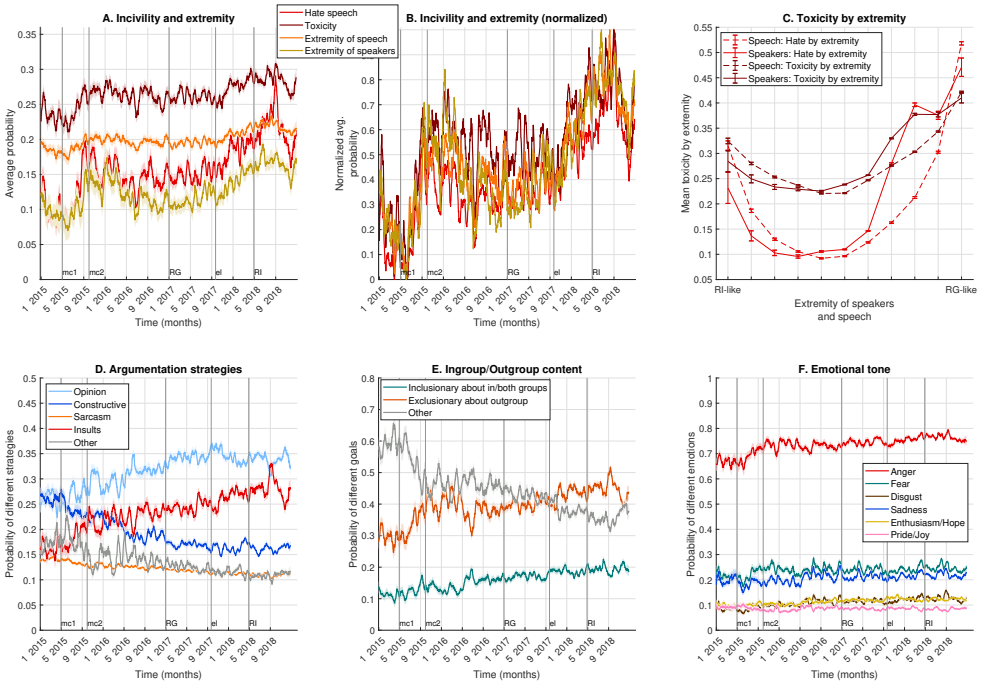

Figure 1: Trends of speech characteristics and indicators of incivility and extremity over time. Error bands reflect standard errors. All trends are smoothed over a two-week window. A. Hate, toxicity and extremity over time. B. Normalized (min-max) hate, toxicity and extremity over time. C. Mean probability of hate speech and toxicity for speech and speakers that resemble RG or RI. D. Probability of argumentation strategies over time. E. Probability of social psychological goals with respect to in- and outgroup over time. F. Probability of emotional tones over time. *Note*. mc1=beginning and mc2=peak of the migrant crisis, RG=start of Reconquista Germanica, el=2017 German elections, RI=start of Reconquista Internet.

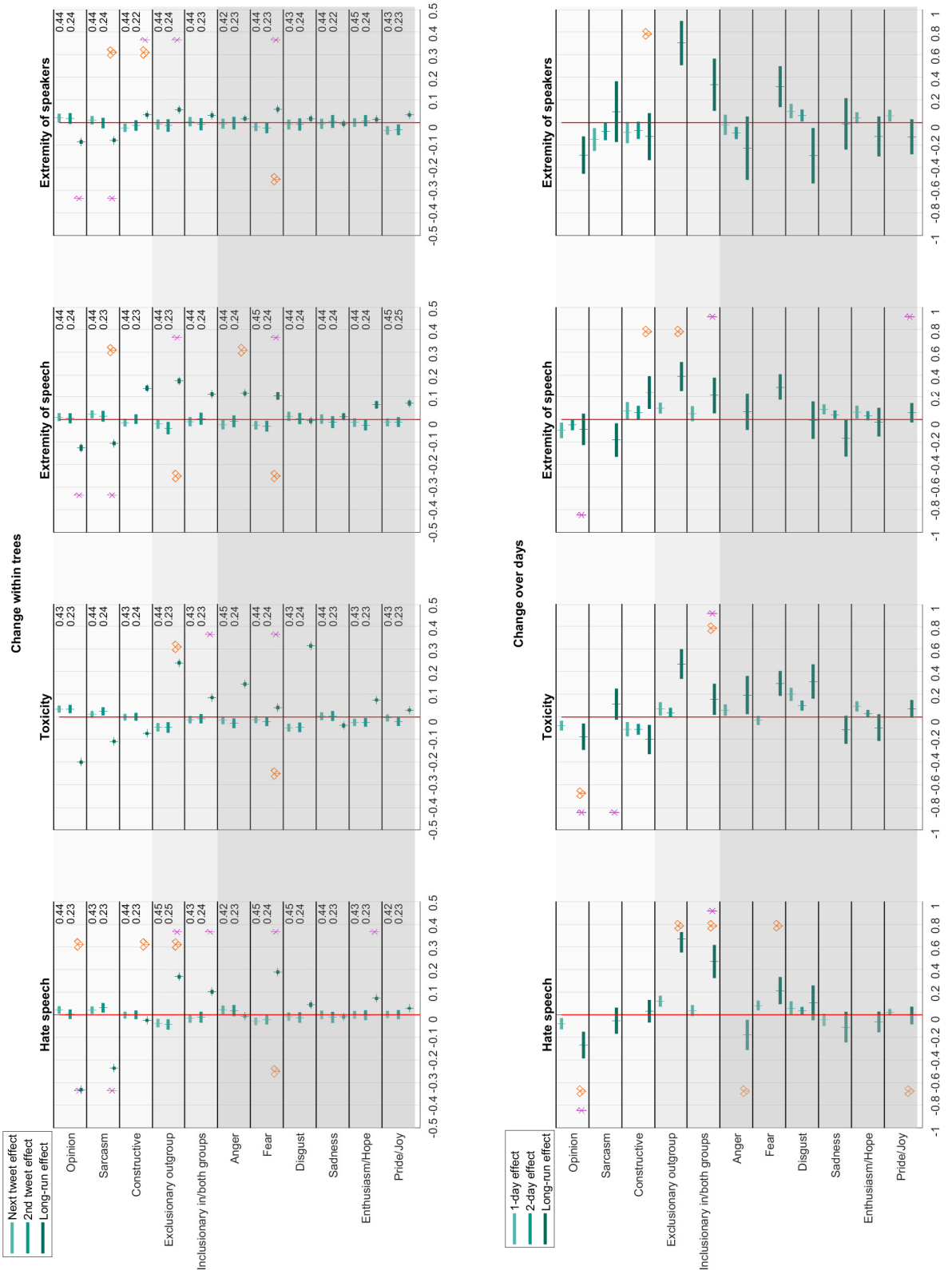
Change within trees
Change over days
Hate speech
Toxicity
Extremity of speech
Extremity of speakers
Next tweet effect
2nd tweet effect
Long-run effect
1-day effect
2-day effect
Long-run effect
Opinion
Sarcasm
Constructive
Exclusionary outgroup
Inclusionary in/both groups
Anger
Fear
Disgust
Sadness
Enthusiasm/Hope
Pride/Joy

Figure 2: Results of ARDL models predicting changes in hate, toxicity, and extremity from different speech characteristics. Top row: Analysis on the level of discussion trees (meta-analytic effects for 3,569 trees with at least 50 Tweets). Numbers denote the proportion of trees with significant short-run effects. Bottom row: Analysis on the level of days (1,461 days between January 1, 2015 and December 31, 2018). *Note.* Logos of RG (purple) and RI (orange) indicate interaction effects of speech characteristics and percentage of extreme speakers (in the tree level analysis), or presence of RG and/or RI (in the day level analysis). Logos are on the left/right hand side, if effects become more negative/positive.