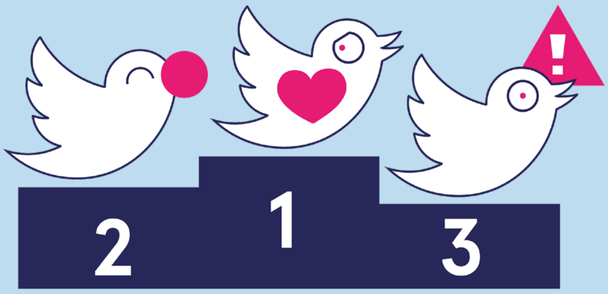
ipl immigration policy lab

ETH zürich



Empathy-based counter speech can reduce hate speech

Online hate speech can be curbed by inducing empathy for those affected. In contrast, the use of humour or warnings of possible consequences have little effects. A team of social scientists and 13 ETH students has demonstrated this in a new scientific publication. Their result forms part of a research collaboration between ETH and the University of Zurich, which also includes the women's umbrella organisation alliance F and two Swiss media companies.

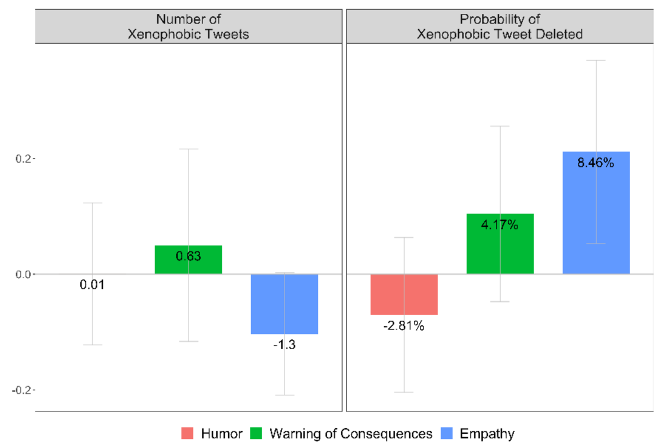
Online hate speech has become a pressing issue worldwide. On social networks, sexual minorities are vilified, members of particular religions are intimidated, and ethnic groups are discriminated. In addition, hate speech is a threat to democracy, as it can prevent those who are targeted from participating in public debate.

To moderate hateful comments, many social media platforms have developed sophisticated filters. However, these alone are not sufficient to fix the problem. For example, Facebook estimates (according to the internal documents leaked in October 2021) that it is not able to delete more than 5 percent of the hate comments posted. Furthermore, automatic filters are imprecise and could harm freedom of speech.

Inducing empathy with those affected

An alternative to deleting problematic comments is the use of targeted counterspeech. Counterspeech is used by numerous organisations aiming to tackle online hate speech. However, until now, little is known about which counterspeech strategies are most effective in addressing online hostility. A team of researchers led by Dominik Hangartner, Professor of Public Policy at ETH, have now joined forces with colleagues at the University of Zurich to investigate what kind of messages could encourage authors of hate speech to refrain from such postings in the future. Using machine learning methods, the researchers identified 1,350 English-speaking Twitter users who had published racist or xenophobic content. They randomly assigned these accounts to a control group or one of following three, often-used counterspeech strategies: messages that elicit empathy with the group targeted by racism, humour, or a warning of possible consequences.

The results, are clear: only counterspeech messages that elicit empathy with the people affected by hate speech are likely to persuade the senders to change their behaviour. An example of such a response could be: "Your post is very painful for Jewish people to read..." Compared to the control group, the authors of hate tweets posted around one-third fewer racist or xenophobic comments after such an empathy-inducing intervention. Additionally, the probability that a hate tweet was deleted by its author increased significantly. In contrast, the authors of hate tweets barely reacted to humorous counterspeech. Even a message that reminded the sender that their family, friends and colleagues could see their hateful comments, too, were not effective. This is striking because these two strategies are frequently used by organisations that are committed to combatting hate speech.



"We have certainly not found a panacea against hate speech on the internet, but we have uncovered important clues about which strategies might work, and which do not," says Hangartner. What remains to be studied is whether all empathy-based responses work similarly well, or whether particular messages are more effective. For example, hate speech authors could be encouraged to put themselves in the victim's shoe or be asked to adopt an analogous perspective ("How would you feel if people talked about you like that?").

Blending teaching and research

Alongside Professors Karsten Donnay and Fabrizio Gilardi from the University of Zurich's Digital Democracy Lab, 13 Master's students from the ETH Center for Comparative and International Studies (CIS) were also heavily involved in the project. The students participated in all phases of the project, from developing an algorithm to detect hate tweets, to testing the strategies on Twitter, to statistical analysis and project management. "To me, this new type of collaborative seminar exemplifies a form of education that not only equips students with important tools for data science and social science, but also for research ethics. My hope is that this hands-on education enables them to make a positive impact in the field of digitalisation and social media," says Hangartner.

The students involved take a similar view. "We haven't just read about other people's research; now we also know how a big research project works," says Laurenz Derksen. "Although there was a lot of work involved, this experiment lit a fire in me and got me excited about ambitious and collaborative research," Derksen continues.

Buket Buse Demirci, now a doctoral student, felt that the project went far beyond the normal scope of seminars. As an example, she cites the Pre-Analysis Plan: the public registration of every single research step before the start of the experiment, thus increasing the credibility of the statistical analyses as well as the reliability of the results. Another motivating factor, she says, is that all 13 students are listed as co-authors on the study detailing the results, which is published in one of the most prestigious interdisciplinary science journals. "I've contributed to a study that has not only been published in a scientific journal, but could also have an impact in the real world," says Demirci.

Practical applications through NGO and media

Hangartner is aware that this type of research, embedded in a seminar, may sometimes also yield null results. Yet the experience is valuable for the students in any case, he says. It can help them anticipate what to expect in case they embark on PhD studies and provides hands-on research experience, which is an asset for many different careers inside and outside of academia.

The collaborative research seminar is part of a more comprehensive project to develop algorithms that detect hate speech, and to test and refine further counterspeech strategies. To this end, the research team is collaborating with the Swiss women's umbrella organisation alliance F, which has initiated the civil society project Stop Hate Speech. Through this collaboration the scientists are able to directly translate their research insights into practice, and to provide an empirical basis for alliance F to optimise the design and content of their counterspeech messages.

"The research findings make me very optimistic. For the first time, we now have experimental evidence that show the efficacy of counterspeech in real-life conditions," says Sophie Achermann, Executive Director of alliance F and co-initiator of Stop Hate Speech. Also involved in the research project, which was sponsored by the Swiss innovation agency Innosuisse, are the media companies Ringier and TX Group via their newspapers Blick and 20 Minuten respectively.

Also published in ETHZ News, December 7, 2021

Details see "Hangartner, D, Gennaro, G, Alasiri, S, Bahrich, N, Bornhoft, A, Boucher, J, Demirci, BB, Derkse, L, Hall, A, Jochum, M, Murias Munoz, M, Richter, M, Vogel, F, Wittwer, S, Wüthrich, F, Gilardi, F, Donnay, K. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. Proceedings of the National Academy of Sciences.

Illustration: Claudia Blum