Study: Countering hate on social media

20 November 2020

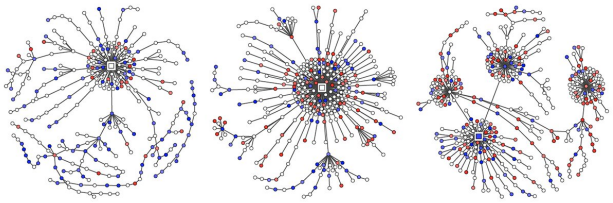

Figure 1 from the paper: Examples of Twitter conversations (reply trees) with labeled hate (red), counter (blue), and neutral speech (white). The root node is shown as a large square. Credit: Garland et al, EMNLP 2020

The rise of online hate speech is a disturbing, growing trend in countries around the world, with serious psychological consequences and the potential to impact, and even contribute to, real-world violence. Citizen-generated counter speech may help discourage hateful online rhetoric, but it has been difficult to quantify and study. Until recently, studies have been limited to small-scale, hand-labeled endeavors.

A new paper published in the proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) offers a framework for studying the dynamics of online hate and counter speech. The paper offers the first large-scale classification of millions of such interactions on Twitter. The authors developed a learning algorithm to assess data from a unique situation on German Twitter, and the findings suggest that organized movements to counteract hate speech on social media are more effective than individuals striking out on their own.

The authors will present their paper, "Countering hate on social media: Large-scale classifications of hate and counter speech" during the November 20, 2020, Workshop on Online Abuse and Harms, which is running in conjunction with EMNLP 2020.

"I've seen this big shift in civil discourse in the last two or three years towards being much more hateful and much more polarized," says Joshua Garland, a mathematician and Applied Complexity Fellow at the Santa Fe Institute. "So, for me, an interesting question was: what's an appropriate response when you're being cyber-bullied or when you're receiving hate speech online? Do you respond? Do you try to get your friends to help protect you? Do you just block the person?"

To study such questions scientifically, researchers must first have access to a wealth of real-world data on both hate speech and counter-speech, and the ability to distinguish between the two. That data existed, and Garland and collaborator Keyan Ghazi-Zahedi at the Max Planck Institute in Germany found it in a five-year interaction that played out over German Twitter: As an alt-right group took to the platform with hate speech, an organized movement rose up to counter it.

"The beauty of these two groups is they were self-labeling," explains Mirta Galesic, the team's social scientist and a professor of human social dynamics at SFI. She says researchers who study counter-speech usually have to employ hundreds of students to hand-code thousands of posts. But Garland and Ghazi-Zahedi were able to input the self-labeled posts into a machine-learning algorithm to automate large swaths of the classification. The team also relied on 20-30 human coders to check that the machine classifications matched up with intuition about what registers as hate and counter-speech.

The result was a dataset of unprecedented size that allows the researchers to analyze not just isolated instances of hate and counter speech, but also compare long-running interactions between the two.

The team collected one dataset of millions of tweets posted by members of the two groups, using these self-identified tweets to train their classification algorithm to recognize hate and counter speech. Then, they applied their algorithm to study the dynamics of some 200,000 conversations that occurred between 2013 and 2018. The authors plan to soon publish a follow-up paper analyzing the dynamics revealed by their algorithm.

"Now we can resolve a massive data set from 2016 to 2018 to see how the proportion of hate and counter-speech changed over time, who gets more likes, who is retweeted, and how they replied to each other" Galesic says.

The quantity of data, a tremendous boon, also makes it "incredibly complex," Garland notes. The researchers are in the process of comparing tactics for both groups and pursuing broader questions such as whether certain counter-speech strategies are more effective than others.

"What I'm hoping is that we can come up with a rigorous social theory that tells people how to counter hate in a productive way that's non-polarizing," Garland says, "and bring the Internet back to civil discourse."

More information:
www.aclweb.org/anthology/2020.alw-1.13/

Provided by Santa Fe Institute