Regret being hostile online? AI tool guides users away from vitriol

February 14 2023, by Louis DiPietro

Credit: Pixabay/CC0 Public Domain

To help identify when tense online debates are inching toward

irredeemable meltdown, Cornell researchers have developed an artificial intelligence tool that can track these conversations in real-time, detect when tensions are escalating and nudge users away from using incendiary language.

Detailed in two recently published papers that examine AI's effectiveness in moderating online discussions, the research shows promising signs that conversational forecasting methods within the field of natural language processing could prove useful in helping both moderators and users proactively lessen vitriol and maintain healthy, productive debate forums.

"Well-intentioned debaters are just human. In the middle of a heated debate, in a topic you care about a lot, it can be easy to react emotionally and only realize it after the fact," said Jonathan Chang, a doctoral student in the field of computer science, and lead author of "Thread With Caution: Proactively Helping Users Assess and Deescalate Tension in Their Online Discussions," which was presented virtually at the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW) on Nov. 10.

The idea is not to tell users what to say, Chang said, but to encourage users to communicate as they would in-person.

The tool, named ConvoWizard, is a browser extension powered by a deep neural network. That network was trained on mountains of language-based data pulled from the subreddit Change My View, a forum that prioritizes good faith debates on potentially heated subjects related to politics, economics and culture.

When participating Change My View users enable ConvoWizard, the tool can inform them when their conversation is starting to get tense. It can also inform users, in real-time as they are writing their replies,

whether their comment is likely to escalate tension. The study suggests that AI-powered feedback can be effective in guiding the user toward language that elevates constructive debate, researchers said.

"ConvoWizard is basically asking, 'If this comment is posted, would this increase or decrease estimated tension in the conversation?' If the comment increases tension, ConvoWizard would give a warning," Chang said. The textbox would turn red, for example. "The tool toes this line of giving feedback without veering into the dangerous territory of telling them to do this or that."

To test ConvoWizard, Cornell researchers collaborated with the Change My View subreddit, where roughly 50 participating forum moderators and members put the tool to use. Findings were positive: 68% felt the tool's estimates of risk were as good as or better than their own intuition, and more than half of participants reported that ConvoWizard warnings stopped them from posting a comment they would have later regretted.

Chang also noted that, prior to using ConvoWizard, participants were asked if they ever posted something they regretted. More than half said yes.

"These findings confirm that, yes, even well-intentioned users can fall into this type of behavior and feel bad about it," he said.

"It's exciting to think about how AI-powered tools like ConvoWizard could enable a completely new paradigm for encouraging high-quality online discussions, by directly empowering the participants in these discussions to use their own intuitions, rather than censoring or constraining them," said Cristian Danescu-Niculescu-Mizil, associate professor of information science in the Cornell Ann S. Bowers College of Computing and Information Science and research co-author.

In a separate Cornell paper also presented at CSCW, "Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support," researchers—including Chang—explore how an AI tool powered by similar conversational forecasting technology might be integrated and used among moderators.

The research aims to find healthier ways to both address vitriol on forums in real-time and lessen the workload on volunteer moderators. Paper authors are Charlotte Schluger '22, Chang, Danescu-Niculescu-Mizil and Karen Levy, associate professor of information science, and associate member of the faculty of Cornell Law School.

"There's been very little work on how to help moderators on the proactive side of their work," Chang said. "We found that there is potential for algorithmic tools to help ease the burden felt by moderators and help them identify areas to review within conversations and where to intervene."

Looking ahead, Chang said the research team will explore how well a model like ConvoWizard generalizes to other online communities.

How conversation-forecasting algorithms scale is another important question, researchers said. Chang pointed to a finding from the ConvoWizard research that showed 64% of Change My View participants felt the tool, if broadly adopted, would improve overall discussion quality. "We're interested in finding out what would happen if a larger slice of an online community used this technology," he said. "What would be the long-term effects?"

Both papers have been published as part of the *Proceedings of the ACM on Human-Computer Interaction*.

More information: Jonathan P. Chang et al, Thread With Caution: