Predicting when online conversations turn toxic

July 13 2018, by Melanie Lefkowitz

Credit: CC0 Public Domain

The internet offers the potential for constructive dialogue and cooperation, but online conversations too often degenerate into personal

attacks. In hopes that those attacks can be averted, Cornell researchers have created a model to predict which civil conversations might take a toxic turn.

After analyzing hundreds of exchanges on Wikipedia, the researchers developed a computer program that scans for red flags – such as repeated, direct questioning and use of the word "you" in the first two posts – to predict which initially civil conversations would go awry.

Early exchanges that included greetings, expressions of gratitude, hedges such as "it seems," and the words "I" and "we" were more likely to remain civil, the study found.

"There are millions of these discussions, and you can't possibly monitor all of them live. This system might help human moderators better direct their attention," said Cristian Danescu-Niculescu-Mizil, assistant professor of information science and co-author of the paper "Conversations Gone Awry: Detecting Early Signs of Conversational Failure."

"We as humans have an intuition of how to detect whether something is going bad, but it's just a suspicion. We can't do it 100 percent of the time. Therefore, we wonder if we can build systems to replicate this intuition, because humans are expensive and busy, and we think this is the type of problem where computers have the potential to outperform humans," Danescu-Niculescu-Mizil said.

The computer model, which also considered Google's Perspective, a machine-learning tool for evaluating "toxicity," was correct around 65 percent of the time. Humans guessed correctly 72 percent of the time.

People can test their own ability to guess which conversations will derail at an online quiz.

The study analyzed 1,270 conversations that began civilly but degenerated into personal attacks, culled from 50 million conversations across 16 million Wikipedia "talk" pages, where editors discuss articles or other issues. They examined exchanges in pairs, comparing each conversation that ended badly with one that succeeded on the same topic, so the results weren't skewed by sensitive subject matter such as politics.

The paper, co-written with Cornell Ph.D. information science student Justine Zhang; Ph.D. computer science students Jonathan P. Chang, and Yiqing Hua; Lucas Dixon and Nithum Thain of Jigsaw; and Dario Taraborelli of the Wikimedia Foundation, will be presented at the Association for Computational Linguistics' annual meeting, from July 15 to 20 in Melbourne, Australia.

The researchers hope this model can be used to rescue at-risk conversations and improve online dialogue, rather than for banning specific users or censoring certain topics. Some online posters, such as nonnative English speakers, may not realize they could be perceived as aggressive, and warnings from such a system could help them self-adjust.

"If I have tools that find personal attacks, it's already too late, because the attack has already happened and people have already seen it," Chang said. "But if you understand this conversation is going in a bad direction and take action then, that might make the place a little more welcoming."

More information: Conversations Gone Awry: Detecting Early Signs of Conversational Failure. www.cs.cornell.edu/~cristian/C … ations_gone_awry.pdf