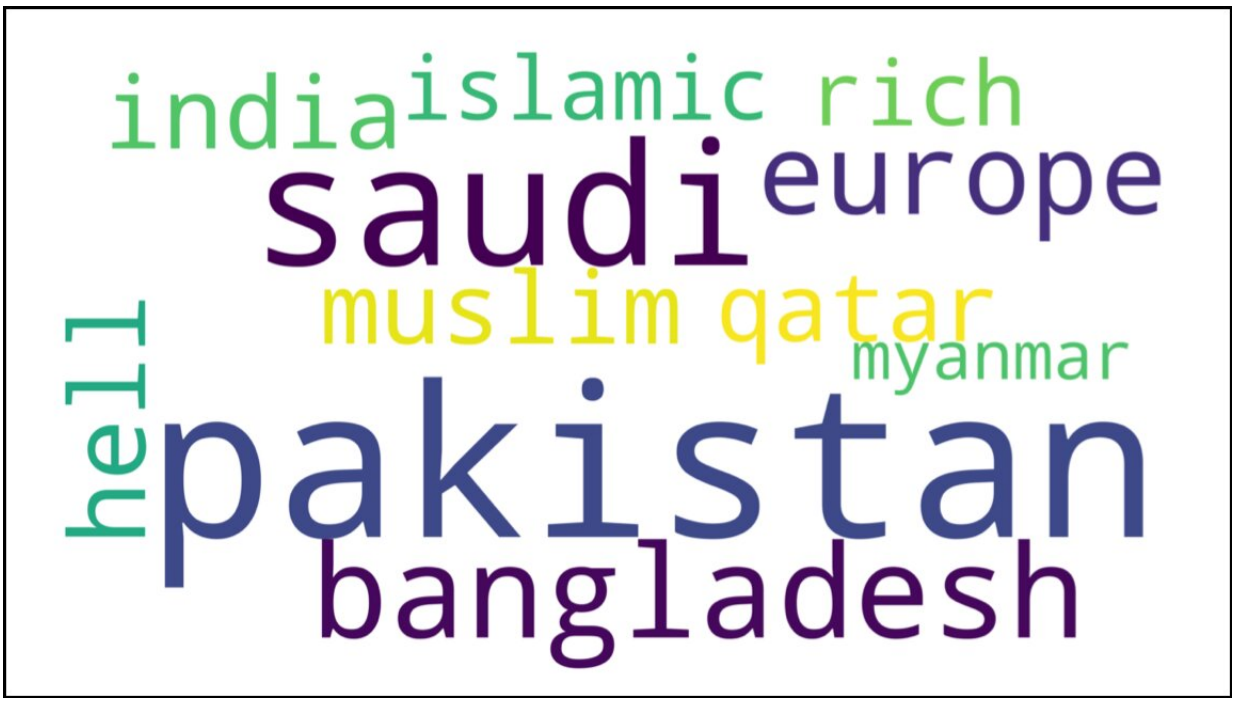
Hate speech countered by detecting, highlighting 'help speech'

January 13 2020



This word cloud depicts responses on social media to the question of where Rohingya refugees should go. Credit: Carnegie Mellon University

Complete the following sentence: Rohingya refugees should go to ... —

A. Pakistan.
B. Bangladesh.
C. Hell.

These aren't good choices, but all are sentiments that have been expressed repeatedly on social media. The Rohingyas, who began fleeing Myanamar in 2017 to avoid ethnic cleansing, are ill-equipped to defend themselves from these online attacks, but innovations from Carnegie Mellon University's Language Technologies Institute (LTI) could help counter the hate speech directed at them and other voiceless groups.

The LTI researchers have developed a system that leverages artificial intelligence to rapidly analyze hundreds of thousands of comments on social media and identify the fraction that defend or sympathize with disenfranchised minorities such as the Rohingya community. Human social media moderators, who couldn't possibly manually sift through so many comments, would then have the option to highlight this "help speech" in comment sections.

"Even if there's lots of hateful content, we can still find positive comments," said Ashiqur R. KhudaBukhsh, a post-doctoral researcher in the LTI who conducted the research with alumnus Shriphani Palakodety. Finding and highlighting these positive comments, they suggest, might do as much to make the internet a safer, healthier place as would detecting and eliminating hostile content or banning the trolls responsible.

Left to themselves, the Rohingyas are largely defenseless against online hate speech. Many of them have limited proficiency in global languages such as English, and they have little access to the internet. Most are too busy trying to stay alive to spend much time posting their own content, KhudaBukhsh said.

To find relevant help speech, the researchers used their technique to search more than a quarter of a million comments from YouTube in what they believe is the first AI-focused analysis of the Rohingya refugee crisis. They will present their findings at the Association for the

Advancement of Artificial Intelligence annual conference, Feb. 7-12, in New York City.

Similarly, in an as-yet unpublished study, they used the technology to search for antiwar "hope speech" among almost a million YouTube comments surrounding the February 2019 Pulwama terror attack in Kashmir, which enflamed the longstanding India-Pakistan dispute over the region.

The ability to analyze such large quantities of text for content and opinion is possible because of recent major improvements in language models, said Jaime Carbonell, LTI director and a co-author on the study. These models learn from examples so they can predict what words are likely to occur in a given sequence and help machines understand what speakers and writers are trying to say.

But the CMU researchers developed a further innovation that made it possible to apply these models to short social media texts in South Asia, he added. Short bits of text, often with spelling and grammar mistakes, are difficult for machines to interpret. It's even harder in South Asian countries, where people may speak several languages and tend to "code switch," combining bits of different languages and even different writing systems in the same statement.

Existing machine learning methods create representations of words, or word embeddings, so that all words with a similar meaning are represented in the same way. This technique makes it possible to compute the proximity of a word to others in a comment or post. To extend this technique to the challenging texts of South Asia, the CMU team obtained new embeddings that revealed language groupings or clusters. This language identification technique worked as well or better than commercially available solutions.

This innovation has become an enabling technology for computational analyses of social media in that region, Carbonell noted.

Samplings of the YouTube comments showed about 10% of the comments were positive. When the researchers used their method to search for help speech in the larger dataset, the results were 88% positive, indicating that the method could substantially reduce the manual effort necessary to find them, KhudaBukhsh said.

"No country is too small to take on refugees," said one text, while another argued "all the countries should take a stand for these people."

But detecting pro-Rohingya texts can be a double-edged sword: some texts can contain language that could be considered hate speech against their alleged persecutors, he added.

Antagonists of the Rohingya are "really kind of like animals not like human beings so that's why they genocide innocent people," said one such text. Though the method reduces manual efforts, comments such as this indicate the continuing need for human judgment and for further research, the scientists concluded.

Provided by Carnegie Mellon University