Can Twitter anticipate attacks against Asians and Asian Americans?

April 21 2020, by Bob Marcotte

Credit: CC0 Public Domain

University of Rochester computer scientists are gleaning a wealth of information from Twitter users to document the social impacts of the novel coronavirus pandemic.

For example, a new study by the research group of Jiebo Luo, a professor of computer science, and posted to the scholarly website *ArXiv*, finds that the increased use of terms like "Chinese virus" and "Wuhan virus" on the social media platform correlated strongly with a rise in media reports of attacks on Chinese and other Asians.

The researchers were also able predict with more than 80 percent accuracy which Twitter users are more likely to use the terms based on their age, gender, geographic location, "social capital," and political affiliation. The terms used to refer to the source of the pandemic has sparked controversy in some media circles between those who consider a geographic description an accurate reflection of where the virus originated while others consider the geographic terms to be pejorative.

A real-time look at a large-scale crisis

"To the best of our knowledge, this is the first large-scale social media-based study to characterize users with respect to their usage of controversial terms during a major crisis," writes lead author Hanjia Lyu, a Ph.D. student. Long Chen '20, an undergraduate in the group, is a co-author along with Luo.

Luo's group is also using Twitter data to explore other aspects of the coronavirus pandemic, including its impact on mental health, on the success of crowd-funding platforms, on how college students react to social distancing, and the relationship between hoarding and scarcity.

"The data captured in social media platforms can provide an important real-time look into how people communicate and what they think is important to talk about," Luo says.

The researchers gathered more than 17 million tweets—about 1.5 terabytes of data—from March 23 to 26. They then applied a facial

recognition platform to help determine which Twitter users could be confidently characterized by age, gender, and race. Users who followed candidates from both parties were excluded.

This produced a working database of 593,233 tweets using "controversial terms" and 490,168 tweets using "noncontroversial terms."

The researchers then used machine-learning classifier techniques to predict which users would be most likely to use either controversial or noncontroversial terms.

Suburban as well as rural users most likely to use controversial terms

The researchers were able to draw a number of conclusions based on their analysis of the more than one million tweets. Among them:

- Males were responsible for 61 percent of tweets using controversial terms.
- Females were responsible for to 56.2 percent of tweets using noncontroversial terms.
- More than half of those using noncontroversial terms were under 35 years of age; users older than 45 are more likely to use controversial terms.
- Controversial terms were more likely to be used by Twitter users in rural and suburban areas.
- Among Twitter users whose political following could be determined, followers of President Donald Trump were more likely to use controversial terms. Followers of Elizabeth Warren and Pete Buttegieg were most likely to use noncontroversial terms.

- Twitter users who have had accounts longer—and who have more followers, friends, favorites and other "social capital"—were more likely to use noncontroversial terms.

Luo's group used similar methodologies to track the 2016 presidential campaign and offer clues as to why the race turned out the way it did.

 More information: Sense and Sensibility: Characterizing Social Media Users Regarding the Use of Controversial Terms for COVID-19: arxiv.org/abs/2004.06307

Provided by University of Rochester