Study finds racial bias in tweets flagged as hate speech

6 August 2019, by Melanie Lefkowitz

Credit: CC0 Public Domain

Tweets believed to be written by African Americans are much more likely to be tagged as hate speech than tweets associated with whites, according to a Cornell study analyzing five collections of Twitter data marked for abusive language.

All five datasets, compiled by academics for research, showed bias against Twitter users believed to be African American. Although social media companies—including Twitter—probably don't use these datasets for their own hate-speech detection systems, the consistency of the results suggests that similar bias could be widespread.

"We found consistent, systematic and substantial racial biases," said Thomas Davidson, a doctoral candidate in sociology and first author of "Racial Bias in Hate Speech and Abusive Language Datasets," which was presented at the Annual Meeting of the Association for Computational Linguistics, July 28-Aug. 2 in Florence, Italy.

"These systems are being developed to identify language that's used to target marginalized populations online," Davidson said. "It's extremely concerning if the same systems are themselves discriminating against the population they're designed to protect."

As internet giants increasingly turn to artificial intelligence to flag hateful content amid millions of posts, concern about bias in machine learning models is on the rise. Because bias often begins in the data used to train these models, the researchers sought to evaluate datasets that were created to help understand and classify hate speech.

To perform their analysis, they selected five datasets—one of which Davidson helped develop at Cornell—consisting of a combined 270,000 Twitter posts. All five had been annotated by humans to flag abusive language or hate speech.

For each dataset, the researchers trained a machine learning model to predict hateful or offensive speech.

They then used a sixth database of more than 59 million tweets, matched with census data and identified by location and words associated with particular demographics, in order to predict the likelihood that a tweet was written by someone of a certain race.

Though their analysis couldn't conclusively predict the race of a tweet's author, it classified tweets into "black-aligned" and "white-aligned," reflecting the fact that they contained language associated with either of those demographics.

In all five cases, the algorithms classified likely African American tweets as sexism, hate speech, harassment or abuse at much higher rates than those tweets believed to be written by whites—in some cases, more than twice as frequently.

The researchers believe the disparity has two causes: an oversampling of African Americans'

tweets when databases are created; and inadequate training for the people annotating tweets for potential hateful content.

"When we as researchers, or the people we pay online to do crowdsourced annotation, look at these tweets and have to decide, "Is this hateful or not hateful?" we may see language written in what linguists consider African American English and be more likely to think that it's something that is offensive due to our own internal biases," Davidson said. "We want people annotating data to be aware of the nuances of online speech and to be very careful in what they're considering hate speech."

More information: Racial Bias in Hate Speech and Abusive Language Detection Datasets. arxiv.org/pdf/1905.12516.pdf

Provided by Cornell University