Context reduces racial bias in hate speech detection algorithms

7 July 2020

Credit: CC0 Public Domain

Understanding what makes something harmful or offensive can be hard enough for humans, never mind artificial intelligence systems.

So, perhaps it's no surprise that social media hate speech detection algorithms, designed to stop the spread of hateful speech, can actually amplify racial bias by blocking inoffensive tweets by black people or other minority group members.

In fact, one previous study showed that AI models were 1.5 times more likely to flag tweets written by African Americans as "offensive"—in other words, a false positive—compared to other tweets.

Why? Because the current automatic detection models miss out on something vital: context. Specifically, hate speech classifiers are oversensitive to group identifiers like "black," "gay," or "transgender," which are only indicators of hate speech when used in some settings.

Now, a team of USC researchers has created a hate speech classifier that is more context-sensitive, and less likely to mistake a post containing a group identifier as hate speech.

To achieve this, the researchers programmed the algorithm to consider two additional factors: the context in which the group identifier is used, and whether specific features of hate speech are also present, such as dehumanizing and insulting language.

"We want to move hate speech detection closer to being ready for real-world application," said Brendan Kennedy, a computer science Ph.D. student and co-lead author of the study, published at ACL 2020, July 6.

"Hate speech detection models often 'break,' or generate bad predictions, when introduced to real-world data, such as social media or other online text data, because they are biased by the data on which they are trained to associate the appearance of social identifying terms with hate speech."

Additional authors of the study, titled "Contextualizing Hate Speech Classifiers with Post-Hoc Explanation," are co-lead author Xisen Ji, a USC computer science Ph.D. student, and co-authors Aida Mostafazadeh Davani, a Ph.D. computer science student, Xiang Ren, an assistant professor of computer science and Morteza Dehghani, who holds joint appointments in psychology and computer science

Why AI bias happens

Hate speech detection is part of the ongoing effort against oppressive and abusive language on social media, using complex algorithms to flag racist or violent speech faster and better than human beings alone. But machine learning models are prone to learning human-like biases from the training data that feeds these algorithms.

For instance, algorithms struggle to determine if

group identifiers like "gay" or "black" are used in offensive or prejudiced ways because they're trained on imbalanced datasets with unusually high rates of hate speech (white supremacist forums, for instance). As a result, the models find it hard to generalize to real-world applications.

"It is key for models to not ignore identifiers, but to match them with the right context," said Professor Xiang Ren, an expert in natural language processing.

"If you teach a model from an imbalanced dataset, the model starts picking up weird patterns and blocking users inappropriately."

To test the systems, the researchers accessed a large, random sample of text from "Gab," a social network with a high rate of hate speech, and "Stormfront," a white supremacist website. The text had been hand-flagged by humans as prejudiced or dehumanizing.

They then measured the state-of-the-art model's tendencies, versus their own model's, towards inappropriately flagging non-hate speech, using 12,500 New York Times articles devoid of hate speech, excepting quotation. State-of-the-art models achieved a 77 % accuracy of identifying hate versus non-hate. The USC model was able to boost this to 90 %.

"This work by itself does not make hate speech detection perfect, that is a huge project that many are working on, but it makes incremental progress," said Kennedy.

"In addition to preventing social media posts by members of protected groups from being inappropriately censored, we hope our work will help ensure that hate speech detection does not do unnecessary harm by reinforcing spurious associations of prejudice and dehumanization with social groups."

Provided by University of Southern California