Google's Artificial Intelligence Hate Speech Detector Is 'Racially Biased,' Study Finds

Nicole Martin

Twitter

Google created an artificial intelligence algorithm in 2016 meant to monitor and prevent hate speech on social media platforms and websites. However, the machine learning tool may be having the opposite outcome as it seems to be biased against black people.

In order for the algorithm to search for hate speech, developers taught the tool to go through a database of over 100,000 tweets that were labeled "toxic" by Google's API called Perspective. The machine learning tool was used to flag content from healthy to toxic. Perspective defines toxic as “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.” From this learning, the algorithm was able to evaluate new content being produced on the scale of how toxic it would be perceived.

Recently, a group of researchers at the University of Washington discovered that the tool was profiling tweets posted by African-Americans as hate speech. The algorithm had a high rate of flagging content posted by African-Americans on social media, such as twitter, as toxic content when most of the language and content in those tweets were not harmful.

The study was lead by Maarten Sap, a Ph.D. student at the university, and found that tweets were written in African-American Vernacular English (AAVE) were often flagged as offensive and therefore labeled as "hate speech." This made the algorithm grow to be inherently biased towards African-Americans. When the team tested the algorithm against 5.4 million tweets, they found that the tool was twice as likely to flag posts by those who identified in the study as African-American in the database as toxic speech than those who it identified as other races.

According to the study, the use of the "n-word" online used by African-Americans was flagged even though its use is culturally more acceptable and a term often used in AAVE as a non-hate speech by other African- Americans. However, there are instances where the "n-word" is used in hateful terms and the algorithm is currently unable to tell the difference at this time.

A recent report from IBM research explained that AI is only as good as the data that is input into the algorithm.

"Bad data can contain implicit racial, gender, or ideological biases. Many AI systems will continue to be trained using bad data, making this an ongoing problem. But we believe that bias can be tamed and that the AI systems that will tackle bias will be the most successful," said the report.

Time will tell how Google will shift its algorithm to be less biased against AAVE. There will

need to be some monitoring of the systems to ensure there is no bias against different cultures of speech when monitoring hate speech online.

"A crucial principle, for both humans and machines, is to avoid bias and therefore prevent discrimination. Bias in [the] AI system mainly occurs in the data or in the algorithmic model. As we work to develop AI systems we can trust, it’s critical to develop and train these systems with data that is unbiased and to develop algorithms that can be easily explained."