MediaWell

News Item

Can Facebook use AI to Fight Online Abuse? | The Conversation US

By Daniel Lowd

June 13, 2018

Daniel Lowd, Associate Professor of Computer and Information Science at the University of Oregon, argues that there’s no silver bullet for fighting abuse and hate speech online. He says the definition of what constitutes abuse changes as society does and that because of this, necessary solutions should integrate human, AI, and regulatory efforts.

> It can be difficult to determine when human-generated data are causing an AI to perform poorly. When possible, the best defense is for humans to add constraints to the system, such as removing language patterns that are considered sexist. Data poisoning can also be detected by measuring accuracy on a separate, curated data set: If a new model performs poorly on trusted data, then that could mean the new training data are bad. Finally, poisoning can be made less effective by removing outliers, data points that are very different from the rest of the training data.
>
> Of course, no machine learning system will ever be perfect. Like humans, computers should be used as part of a larger effort to fight abuse. Even email spam, a major success for machine learning, relies on more than just good algorithms: New internet communications standards make it harder for spammers to hide their identities when sending messages. In addition, federal law, such as the 2003 CAN-SPAM Act, sets standards for commercial email, including penalties for violations. Similarly, addressing online abuse may require new standards and policies, not just smarter artificial intelligence.

Source: Can Facebook use AI to Fight Online Abuse?