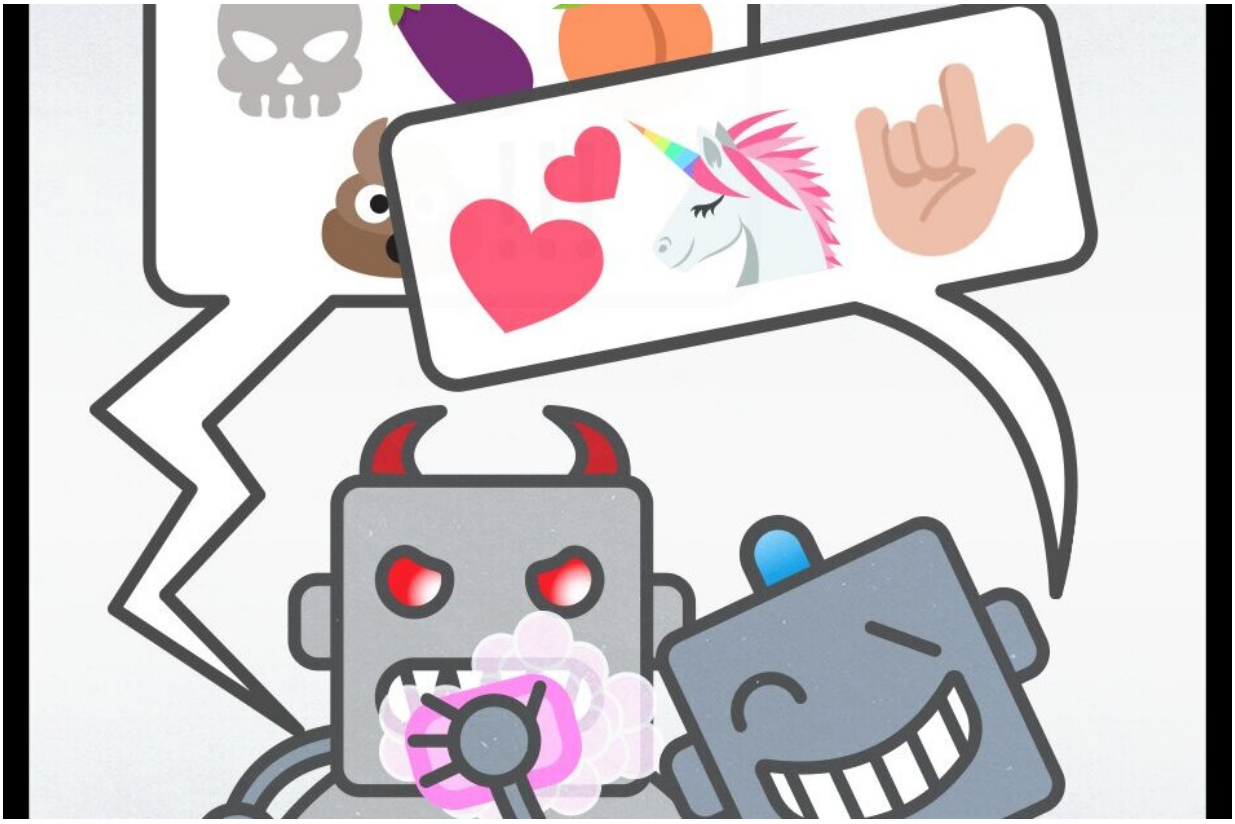
Researchers develop a method to keep bots from using toxic language

April 21 2022



Researchers at the University of California San Diego have developed algorithms to rid speech generated by online bots of offensive language, on social media and elsewhere. Credit: University of California San Diego

Researchers at the University of California San Diego have developed algorithms to rid speech generated by online bots of offensive language,

on social media and elsewhere.

Chatbots using toxic language is an ongoing issue. But perhaps the most famous example is Tay, a Twitter chatbot unveiled by Microsoft in March 2016. In less than 24 hours, Tay, which was learning from conversations happening on Twitter, started repeating some of the most offensive utterances tweeted at the bot, including racist and misogynist statements.

The issue is that chatbots are often trained to repeat their interlocutors' statements during a conversation. In addition, the bots are trained on huge amounts of text, which often contain toxic language and tend to be biased;certain groups of people are overrepresented in the training set and the bot learns language representative of that group only. An example is a bot producing negative statements about a country, propagating bias because it's learning from a training set where people have a negative view of that country.

"Industry is trying to push the limits of language models," said UC San Diego computer science Ph.D. student Canwen Xu, the paper's first author. "As researchers, we are comprehensively considering the social impact of language models and addressing concerns."

Researchers and industry professionals have tried several approaches to clean up bots' speech–all with little success. Creating a list of toxic words misses words that when used in isolation are not toxic, but become offensive when used in combination with others. Trying to remove toxic speech from training data is time consuming and far from foolproof. Developing a neural network that would identify toxic speech has similar issues.

Instead, the UC San Diego team of computer scientists first fed toxic prompts to a pre-trained language model to get it to generate toxic

content. Researchers then trained the model to predict the likelihood that content would be toxic. They call this their "evil model." They then trained a "good model," which was taught to avoid all the content highly ranked by the "evil model."

They verified that their good model did as well as state-of-the-art methods–detoxifying speech by as much as 23 percent.

They presented their work at the AAAI Conference on Artificial Intelligence held online in March 2022.

Researchers were able to develop this solution because their work spans a wide range of expertise, said Julian McAuley, a professor in the UC San Diego Department of Computer Science and Engineering and the paper's senior author.

"Our lab has expertise in algorithmic language, in natural language processing and in algorithmic de-biasing," he said. "This problem and our solution lie at the intersection of all these topics."

However, this language model still has shortcomings. For example, the bot now shies away from discussions of under-represented groups, because the topic is often associated with hate speech and toxic content. Researchers plan to focus on this problem in future work.

"We want to make a language model that is friendlier to different groups of people," said computer science Ph.D. student Zexue He, one of the paper's co-authors.

The work has applications in areas other than chatbots, said computer science Ph.D. student and paper co-author Zhankui He. It could, for example, also be useful in diversifying and detoxifying recommendation systems.

More information: Leashing the Inner Demons: Self-Detoxification for Language Models, arXiv:2203.03072 [cs.CL]
arxiv.org/abs/2203.03072

Provided by University of California - San Diego