Busting anti-queer bias in text prediction

August 11 2022

Credit: Pixabay/CC0 Public Domain

Modern text prediction is far from perfect—take, for instance, when a search query suggests something completely different from your intention. But the trouble doesn't end at inaccuracy. Text prediction can also be extremely exclusive or biased when it comes to predicting results related to marginalized communities.

A team of researchers from the USC Viterbi School of Engineering Information Sciences Institute and the USC Annenberg School for Communication and Journalism, led by Katy Felkner, a USC Viterbi Ph.D. in computer science student and National Science Foundation Graduate Research Fellowship recipient, has developed a system to quantify and fix anti-queer bias in the artificial intelligence behind text prediction.

The project, presented by Felkner at the Queer in AI workshop at the North American Chapter of the Association for Computational Linguistics (NAACL) conference in July, looks at both detecting and reducing anti-queer bias in a large language model, which is used in everything from search bars to language translation systems.

The large language model, or LLM, is the "brain" behind the text prediction that pops up when we type something in a search bar—an artificial intelligence that "completes" sentences by predicting the most likely string of words that follows a given prompt.

However, LLMs must first be "trained" by being fed millions of examples of pre-written content so that they can learn what sentences typically look like. Like an energetic toddler, the LLM repeats what it hears, and what it hears can be heteronormative or even overtly discriminatory.

"Most LLMs are trained on huge amounts of data that's crawled from the internet," Felkner said. "They're going to pick up every kind of social bias that you can imagine is out there on the web."

Few words, big effect

The project found that a popular LLM called BERT showed significant homophobic bias. This bias is measured through Felkner's benchmark, which compares the likelihood that the LLM predicts heteronormative sentences versus sentences that include a queer relationship.

"A heteronormative output is something like 'James held hands with Mary,' versus 'James held hands with Tom,'" said Felkner. "Both are valid sentences, but the issue is that, across a wide variety of contexts, the model prefers the heteronormative output."

While the difference is just a few words, the effect is far from small.

Predicted outputs that talk about queer people in stereotypical ways can enforce users' biases, and the model's lack of 'experience' with queer voices can result in it looking at queer language as obscene.

"A persistent issue for queer people is that a lot of times, the words that we use to describe ourselves, or slurs that have been reclaimed, are still considered obscene or overly sexual," said Felkner, who is also the graduate representative for Queers in Engineering, Science and Technology (QuEST) chapter of Out in STEM at USC.

"If a model routinely flags these words, and these posts are then taken down from the platforms or forums they're on, you're silencing the queer community."

Community input

To tackle this problem, Felkner gave BERT a tune-up by feeding it Tweets and news articles containing LGBT+ keywords. This content used to "train" BERT came from two separate databases of Felkner's own creation, called QueerTwitter and QueerNews.

Although language processing requires extremely large amounts of data—the QueerTwitter database contained over 2.3 million

Tweets—she took care to single out hashtags that were being used primarily by queer and trans people, such as #TransRightsareHumanRights.

As the model was exposed to different perspectives and communities, it became more familiar with queer language and issues. As a result, it was more likely to represent them in its predictions.

After being trained with the new, more inclusive data, the model showed significantly less bias. The tweets from QueerTwitter proved the most effective of the two databases, reducing the prevalence of heteronormative results to almost half of all predictions.

"I think QueerTwitter's results being more effective than QueerNews speaks to the importance of direct community involvement, and that queer and trans voices—and the data from their communities—is going to be the most valuable in designing a technology that won't harm them," Felkner said. "We were excited about this finding because it's empirical proof of that intuition people already hold: that these communities should have an input in how technology is designed."

Going forward, the project will look to address bias that affects specific parts of the LGBT+ community, using more refined and targeted sets of data and more customized prompts for the model to work with—such as tackling harmful stereotypes around lesbians. Long term, Felkner hopes the project can be used to train other LLMs, help researchers test the fairness of their natural language processing, or even uncover completely new biases.

"We're dealing with how to fight against the tide of biased data to get an understanding of what 'unfair' looks like and how to test for and correct it, which is a problem both in general and for subcultures that we don't even know about," said Jonathan May, USC Viterbi research associate professor of computer science, Felkner's advisor and study co-author. "There's a lot of great ways to extend the work that Katy is doing."

More information: Conference: 2022.naacl.org/

Provided by University of Southern California