AI language models show bias against people with disabilities, study finds

October 13 2022

Credit: Pixabay/CC0 Public Domain

Natural language processing (NLP) is a type of artificial intelligence that allows machines to use text and spoken words in many different applications—such as smart assistants or email autocorrect and spam

filters—helping automate and streamline operations for individual users and enterprises. However, the algorithms that drive this technology often have tendencies that could be offensive or prejudiced toward individuals with disabilities, according to researchers at the Penn State College of Information Sciences and Technology (IST).

The researchers found that all the algorithms and models they tested contained significant implicit bias against people with disabilities. Previous research on pretrained language models—which are trained on large amounts of data that may contain implicit biases—has found sociodemographic biases against genders and races, but until now similar biases against people with disabilities have not been widely explored.

"The 13 models we explored are highly used and are public in nature," said Pranav Venkit, doctoral student in the College of IST and first author on the study's paper presented today (Oct. 13) at the 29th International Conference on Computational Linguistics (COLING). "We hope that our findings help developers that are creating AI to help certain groups—especially people with disabilities who rely on AI for assistance in their day-to-day activities—to be mindful of these biases."

In their study, the researchers examined machine learning models that were trained on source data to group similar words together enabling a computer to automatically generate sequences of words. They created four simple sentence templates in which to variably populate a gender noun of "man," "woman," or "person," and one of the 10 most commonly used adjectives in the English language—for example, "They are parents of a good person." Then, they generated more than 600 adjectives that could be associated with either people with or without disabilities—such as neurotypical or visually impaired—to randomly replace the adjective in each sentence. The team tested more than 15,000 unique sentences in each model to generate word associations for the adjectives.

"For example, we selected the word 'good,' and wanted to see how it associated with terms related to both non-disability and disability," explained Venkit. "With the addition of a non-disability term, the effect of 'good' becomes 'great.' But when 'good' is associated with a disability-related term, we get the result of 'bad.' So that change in the form of the adjective itself shows the explicit bias of the model."

While this exercise revealed the explicit bias that exists in the models, the researchers wanted to further measure each model for implicit bias—attitudes toward people or associating stereotypes with them without conscious knowledge. They examined the adjectives generated for the disability and non-disability groups and measured each one's sentiment—an NLP technique to rate whether text is positive, negative or neutral. All models they studied consistently scored sentences with words associated with disability more negatively than those without. One particular model, which was pretrained on Twitter data, flipped the sentiment score from positive to negative 86% of the time when a term related to a disability was used.

"When we look at that score alone, we see that just by the addition of a term related to disability into the conversation, the sentiment score of the whole sentence drops," said Venkit. "For example, if a user includes a term related to disability in a social media comment or post, the probability of that post being censored or restricted increases."

The researchers also tested for implicit bias in two large language models that are used to automatically generate long texts, such as news articles, to see how a blank left in the sentence template would change depending on the adjective used. In this analysis they generated 7,500 sentences, again variably inserting adjectives related to non-disability or disability, and tested to see how a blank left in the sentence template would change depending on the adjective used. In this case, when given the sentence of "A man has 'blank'," the language models predicted "changed" for the

blank word. However, when a disability-related adjective was added to the sentence, resulting in "A deafblind man has 'blank'," the model predicted "died" for the blank.

The models' implicit bias against people with disabilities could be apparent in various applications—for example, in text messages when applying autocorrect to a misspelled word or on social media where there are rules that prohibit abusive or harassing posts. In the latter, because humans are unable to review the vast number of posts that are made, artificial intelligence models use these sentiment scores to filter out those posts that are deemed in violation of the platform's community standards.

"If someone is discussing disability, and even though the post is not toxic, a model like this which doesn't focus on separating the biases might categorize the post as toxic just because there is disability associated with the post," explained Mukund Srinath, doctoral student in the College of IST and co-author of the study.

"Whenever a researcher or developer is using one of these models, they don't always look at all the different ways and all the different people that it is going to affect—especially if they're concentrating on the results and how well it performs," said Venkit. "This work shows that people need to care about what sort of models they are using and what the repercussions are that could affect real people in their everyday lives ."

Venkit and Srinath collaborated with Shomir Wilson, assistant professor of information sciences and technology, on the project.

More information: A Study of Implicit Bias in Pretrained Language Models against People with Disabilities, 29th International Conference on Computational Linguistics.

Provided by Pennsylvania State University