Trained AI models exhibit learned disability bias, researchers say

December 1 2023, by Mary Fetzer

A paper authored by researchers in the College of IST, including doctoral students Mukund Srinath and Pranav Narayanan Venkit, received the Best Short Paper Award from the 2023 Workshop on Trustworthy Natural Language Processing. The research was led by Shomir Wilson, assistant professor of IST. Credit: Jena Soult / Penn State

A growing number of organizations are using sentiment analysis tools from third-party artificial intelligence (AI) services to categorize large amounts of text into negative, neutral or positive sentences for social applications ranging from health care to policymaking. These tools, however, are driven by learned associations that often contain biases against persons with disabilities, according to researchers from the Penn State College of Information Sciences and Technology (IST).

In the paper "Automated Ableism: An Exploration of Explicit Disability Biases in Artificial Intelligence as a Service (AIaaS) Sentiment and Toxicity Analysis Models," researchers detailed an analysis of biases against people with disabilities contained in the natural language processing (NLP) algorithms and models they tested.

The work, led by Shomir Wilson, assistant professor in IST and director of the Human Language Technologies Lab, received the Best Short Paper Award from the 2023 Workshop on Trustworthy Natural Language Processing at the 61st Annual Meeting of the Association for Computation Linguistics, held July 9–14 in Toronto, Canada.

"We wanted to examine whether the nature of a discussion or an NLP model's learned associations contributed to disability bias," said Pranav Narayanan Venkit, a doctoral student in the College of IST and first author on the paper. "This is important because real-world organizations that outsource their AI needs may unknowingly deploy biased models."

The researchers defined disability bias as treating a person with a disability less favorably than someone without a disability in similar circumstances and explicit bias as the intentional association of stereotypes toward a specific population.

A growing number of organizations are using AIaaS, or Artificial Intelligence as a Service, for easy-to-use NLP tools that involve little

investment or risk for the organization, according to the researchers. Among these tools are sentiment and toxicity analyses that enable an organization to categorize and score large volumes of textual data into negative, neutral or positive sentences.

Sentiment analysis is the NLP technique for extracting subjective information—thoughts, attitudes, emotions and sentiments—from social media posts, product reviews, political analyses or market research surveys. Toxicity detection models look for inflammatory or content—such as hate speech or offensive language—that can undermine a civil exchange or conversation.

The researchers conducted a two-stage study of disability bias in NLP tools. They first studied social media conversations related to people with disabilities, specifically on Twitter and Reddit, to gain insight into how bias is disseminated in real-world social settings.

They crawled blog posts and comments from a one-year period that specifically addressed perspectives on people with disabilities or contained the terms or hashtags "disability" or "disabled." The results were filtered and categorized and then statistically analyzed with popular sentiment and toxicity analysis models to quantify any disability bias and harm present in the conversations.

"Statements referring to people with disabilities versus other control categories received significantly more negative and toxic scores than statements from other control categories," said contributing author Mukund Srinath, a doctoral student in the College of IST.

"We wanted to test whether these biases arise from discussions surrounding conversations regarding people with disabilities or from associations made within trained sentiment and toxicity analysis models and found that the main source of bias disseminated from the models

rather than the actual context of the conversation."

The researchers then created the Bias Identification Test in Sentiment (BITS) corpus to help anyone identify explicit disability bias in in any AIaaS sentiment analysis and toxicity detection models, according to Venkit. They used the corpus to show how popular sentiment and toxicity analysis tools contain explicit disability bias.

"All of the public models we studied exhibited significant bias against disability," Venkit said. "There was a problematic tendency to classify sentences as negative and toxic based solely on the presence of disability-related terms, such as 'blind,' without regard for contextual meaning, showcasing explicit bias against terms associated with disability."

According to the researchers, identifying explicit bias in large-scale models may help us to understand the social harm caused by training models from a skewed dominant viewpoint—for developers as well as users.

"Nearly everyone, at some point in their life, experiences a disability that could lead to their being socially marginalized," Venkit said. "This work represents an important step toward identifying and addressing disability bias in sentiment and toxicity analysis models and raising awareness of the presence of bias in AIaaS."

More information: Automated Ableism: An Exploration of Explicit Disability Biases in Artificial Intelligence as a Service (AIaaS) Sentiment and Toxicity Analysis Models. trustnlpworkshop.github.io/papers/5.pdf

Provided by Pennsylvania State University