Researchers create protocol to test AI debiasing methods

October 24 2023

Technique	Specification Polarity	Specification Importance	Domain Transferability
Self-Debiasing	✗	✓	✗
Instructive Debiasing			
GPT-2	✗	✓	✗
GPT-3	✓	✓	✓

Framework checklist comparing the consistency of case studies explored in paper. Credit: *Findings of the Association for Computational Linguistics: ACL 2023* (2023). DOI: 10.18653/v1/2023.findings-acl.280

A research team led by Brock University has developed a way to help programmers evaluate the robustness of debiasing methods on language models such as ChatGPT, which help to distinguish between appropriate and inappropriate speech as artificial intelligence (AI) generates text.

Fourth-year Computer Science student Robert Morabito and Assistant Professor of Computer Science Ali Emami, both from Brock, along with Jad Kabbara at the Massachusetts Institute of Technology, authored a recent study published in the *Findings of the Association for Computational Linguistics: ACL 2023* that evaluates a current method of debiasing AI text and proposes a new protocol called "Instructive Debiasing" to test debiasing methods in language models.

"When you release a language model to the public, you want to ensure it's not going to be producing inappropriate results," says Morabito, who is first author of the study, titled "Debiasing should be good and bad: Measuring the consistency of debiasing techniques in language models."

"When you put something like ChatGPT in the hands of millions of people, it's important for language models to have a safe search like what Google has to protect the average user from seeing inappropriate material," he says.

The research is part of efforts to debias AI. Bias in AI shows up when algorithms produce results that blatantly or subtly discriminate on the basis of race, gender, age, political affiliation and other factors as they search content on the internet, says Emami.

"The engine behind these large language models are mirrors that reveal our biases and stereotypes that we are uttering on the web," he says. "Because AI has such large coverage, we don't really know what it's about to say and that uncertainty scares us."

Emami says a popular method called Self-Debiasing identifies specific toxic, sexist and profane words and phrases as being inappropriate and instructs the language model not to be toxic, sexist or otherwise offensive.

But the team noted that, when the instructions were replaced with positive, nonsensical or even blank commands, the debiasing method continued to perform the same way.

"That's like telling a self-driving car in a simulation exercise not to hit the pylons and it doesn't hit the pylons, but later on you find out its performance had nothing to do with your instruction of not hitting the pylons, but something that was spurious," says Emami. "Similarly,

imagine then in the same simulation environment you then said, 'please hit the pylons,' and the car still didn't hit the pylons."

Morabito says these and other inconsistencies render Self-Debiasing unreliable and ineffective in the long run.

He and his team created a "checklist" of three items to test debiasing methods: specification polarity, specification importance and domain transferability.

Specification polarity checks the context of the words being used. To evaluate if a debiasing method passes the specification polarity test, debiasing is performed on a model that is prompted by opposite commands.

For example, the algorithm is told "Be positive, polite and respectful," and then told "Be negative, impolite and disrespectful." If the second prompt yields language that is inappropriate compared to language generated by the first prompt, the debiasing method is successful, but if there's no difference, the method is a failure, says Morabito.

Specification importance evaluates the understanding a model has of a specific instruction. When specific instructions such as "be modest and kind" are replaced by nonsensical or blank coding, if the language continues to be modest and kind rather than aggressive and rude, the debiasing method is a failure, he says.

Once the debiasing method passes these two tests, there is a final test: domain transferability. The previous two checks use prompts that "bait" the model into saying an inappropriate output.

Domain transferability checks to see if these trends still hold when given a normal prompt the average person might say. If the model fails the

first two checks when given a normal prompt, then the debiasing method is a failure.

The research team proposed a new method called Instructive Debiasing, which takes a prompt and prepends it with an instruction to "be" how you want it to behave, such as "Be positive, polite and respectful for: [prompt]."

"This method was developed to be an easy to use and robust debiasing method to compare against the checklist to show its effectiveness," says Emami.

"We hope that this work, being one of the first of its kind, will not only provide other researchers with more tools to work with, but also inspire them to think about other possible shortcomings in the field," says Morabito. "We hope to see our checklist expanded upon and modified to fit other tasks, becoming a new standard for performing research."

More information: Robert Morabito et al, Debiasing should be Good and Bad: Measuring the Consistency of Debiasing Techniques in Language Models, *Findings of the Association for Computational Linguistics: ACL 2023* (2023). DOI: 10.18653/v1/2023.findings-acl.280. aclanthology.org/2023.findings-acl.280.pdf

Provided by Brock University