Artificial intelligence favors white men under 40

November 18 2021

Credit: CC0 Public Domain

"Insert the missing word: I closed the door to my ____." It's an exercise that many remember from their school days. Whereas some societal groups might fill in the space with the word "holiday home", others may

be more likely to insert "dorm room" or "garage". To a large extent, our word choice depends on our age, where we are from in a country and our social and cultural background.

However, the language models we put to use in our daily lives while using search engines, machine translation, engaging with chatbots and commanding Siri, speak the language of some groups better than others. This has been demonstrated by a study from the University of Copenhagen's Department of Computer Science, which has for the first time studied whether language models favor the linguistic preferences of some demographic groups over others—referred to in the jargon as sociolectal biases. The answer? Yes.

"Across language models, we are able to observe systematic bias. Whereas white men under the age of 40 with shorter educations are the group that language models align best with, the worst alignment is with language used by young, non-white men," says Anders Søgaard, a professor at UCPH's Department of Computer Science and the lead author of the study.

What's the problem?

The analysis demonstrates that up to one in ten of the models' predictions are significantly worse for young, non-white men compared to young white men. For Søgaard, this is enough to pose a problem:

"Any difference is problematic because differences creep their way into a wide range of technologies. Language models are used for important purposes in our everyday lives—such as searching for information online. When the availability of information depends on how you formulate yourself and whether your language aligns with that for which models have been trained, it means that information available to others, may not be available to you."

Professor Søgaard adds that even a slight bias in the models can have more serious consequences in contexts where precision is key:

"It could be in the insurance sector, where language models are used to group cases and perform customer risk assessments. It could also be in legal contexts, such as in public casework, where models are sometimes used to find similar cases rather than precedent. Under such circumstances, a minor difference can prove decisive," he says.

Most data comes from social media

Language models are trained by feeding enormous amounts of text into them to teach models the probability of words occurring in specific contexts. Just as with the school exercise above, models must predict the missing words from a sequence. The texts come from what is available online, most of which have been downloaded from social media and Wikipedia.

"However, the data available on the web isn't necessarily representative of us as tech users. Wikipedia is a good example in that its content is primarily written by young white men. This matters with regards to the type of language that models learn," says Søgaard.

The researchers remain uncertain as to why precisely the sociolectal characteristics of young white men are represented best by the language models. But they do have a educated guess:

"It correlates with the fact that young white men are the group that has contributed most to the data that models are trained on. A preponderance of data originates from social media. And, we know from other studies that it is this demographic that contributes most in writing in these types of open, public fora," explains Anders Søgaard.

If we do nothing, the problem will grow

The problem appears to be growing alongside digital developments, explains Professor Søgaard:

"As computers become more efficient, with more data available, language models tend to grow and be trained on more and more data. For the most prevalent type of language used now, it seems—without us knowing why—that the larger the models, the more biases they have. So, unless something is done, the gap between certain social groups will widen."

Fortunately, something can be done to correct for the problem:

"If we are to overcome the distortion, feeding machines with more data won't do. Instead, an obvious solution is to train the models better. This can be done by changing the algorithms so that instead of treating all data as equally important, they are particularly careful with data that emerges from a more balanced population average," concludes Anders Søgaard.

The research article "Sociolectal Analysis of Pretrained Language Models" is included at the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2021.

More information: Paper: aclanthology.org/2021.emnlp-main.375/

Provided by University of Copenhagen