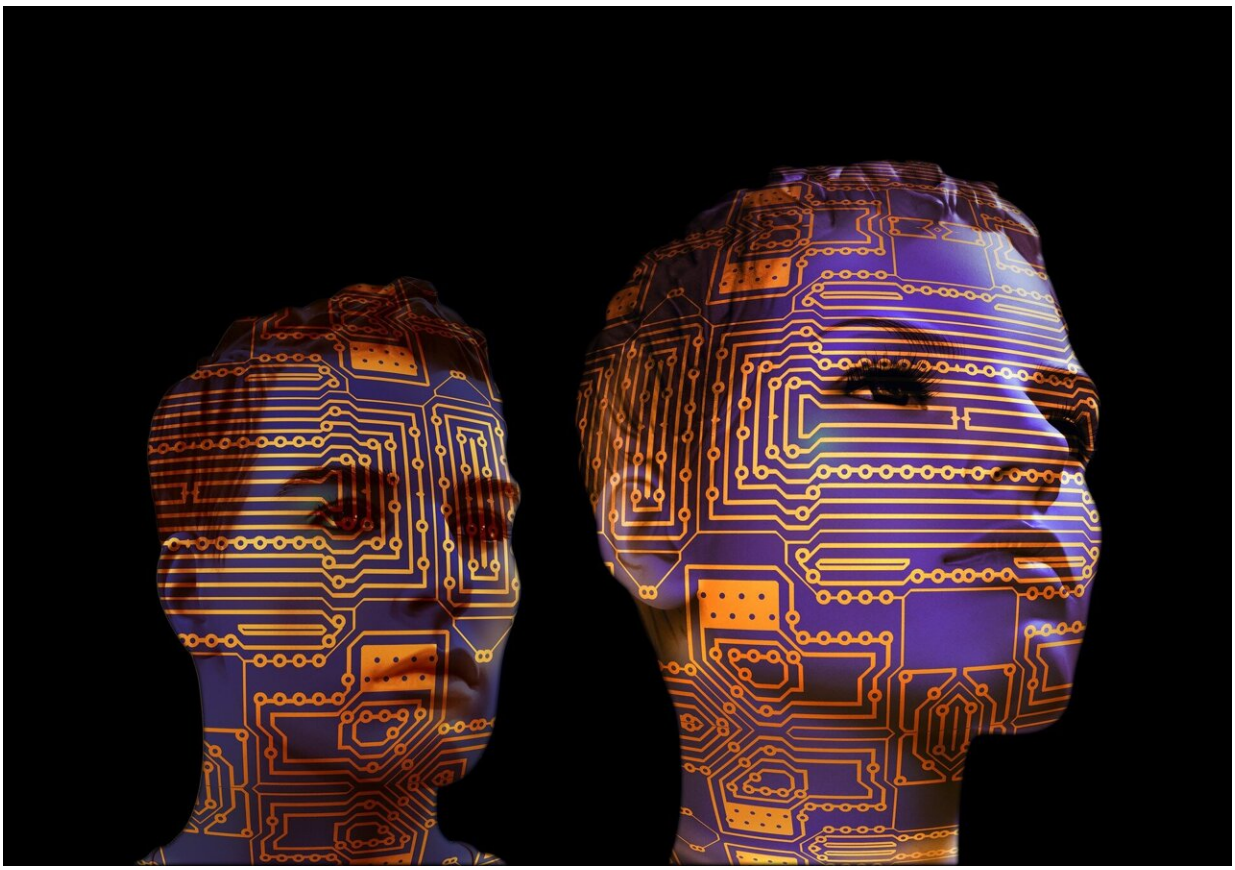
# AI researchers improve method for removing gender bias in machines built to understand and respond to text or voice data

September 8 2022, by Adrianna MacPherson



Credit: Pixabay/CC0 Public Domain

Researchers have found a better way to reduce gender bias in natural

language processing models while preserving vital information about the meanings of words, according to a recent study that could be a key step toward addressing the issue of human biases creeping into artificial intelligence.

While a computer itself is an unbiased machine, much of the data and programming that flows through computers is generated by humans. This can be a problem when conscious or unconscious human biases end up being reflected in the text samples AI models use to analyze and "understand" language.

Computers aren't immediately able to understand text, explains Lei Ding, first author on the study and graduate student in the Department of Mathematical and Statistical Sciences. They need words to be converted into a set of numbers to understand them—a process called word embedding.

"Natural language processing is basically teaching the computers to understand texts and languages," says Bei Jiang, associate professor in the Department of Mathematical and Statistical Sciences.

Once researchers take this step, they're able to then plot words as numbers on a 2D graph and visualize the words' relationships to one another. This allows them to better understand the extent of the gender bias, and later, determine whether the bias was effectively eliminated.

## All the meaning, none of the bias

Though other attempts to reduce or remove gender bias in texts have been successful to some degree, the problem with those approaches is that gender bias isn't the only thing removed from the texts.

"In many gender debiasing methods, when they reduce the bias in a word

vector, they also reduce or eliminate important information about the word," explains Jiang. This type of information is known as semantic information, and it offers important contextual data that could be needed in future tasks involving those word embeddings.

For example, when considering a word like "nurse," researchers want the system to remove any gender information associated with that term while still retaining information that links it with related words such as doctor, hospital and medicine.

"We need to preserve that semantic information," says Ding. "Without it, the embeddings would have very bad performance [in natural language processing tasks and systems]."

## Fast, accurate—and fair

The new methodology also outperformed leading debiasing methods in various tasks that evaluated based on word embedding.

As it becomes refined, the methodology could offer a flexible framework other researchers could apply to their own word embeddings. As long as a researcher has guidance on the right group of words to use, the methodology could be used to reduce bias linked with any particular group.

While at this stage the methodology still requires researcher input, Ding explains it may be possible in the future to have some sort of built-in system or filter that could automatically remove gender bias in a variety of contexts.

Published in the *Proceedings of the AAAI Conference on Artificial Intelligence*, the new methodology is part of a larger project, entitled BIAS: Responsible AI for Gender and Ethnic Labor Market Equality,

that is looking to solve real-world problems.

For example, people reading the same job advertisement may respond differently to particular words in the description that often have a gendered association. A system using the methodology Ding and his collaborators created would be able to flag the words that may change a potential applicant's perception of the job or decision to apply because of perceived gender bias, and suggest alternative words to reduce this bias.

Though many AI models and systems are focused on finding ways to perform tasks with greater speed and accuracy, Ding notes the team's work is part of a growing field that seeks to make strides regarding another important aspect of these models and systems.

"People are focusing more on responsibility and fairness within artificial intelligence systems."

**More information:** Lei Ding et al, Word Embeddings via Causal Inference: Gender Bias Reducing and Semantic Information Preserving, *Proceedings of the AAAI Conference on Artificial Intelligence* (2022). DOI: 10.1609/aaai.v36i11.21443

Provided by University of Alberta