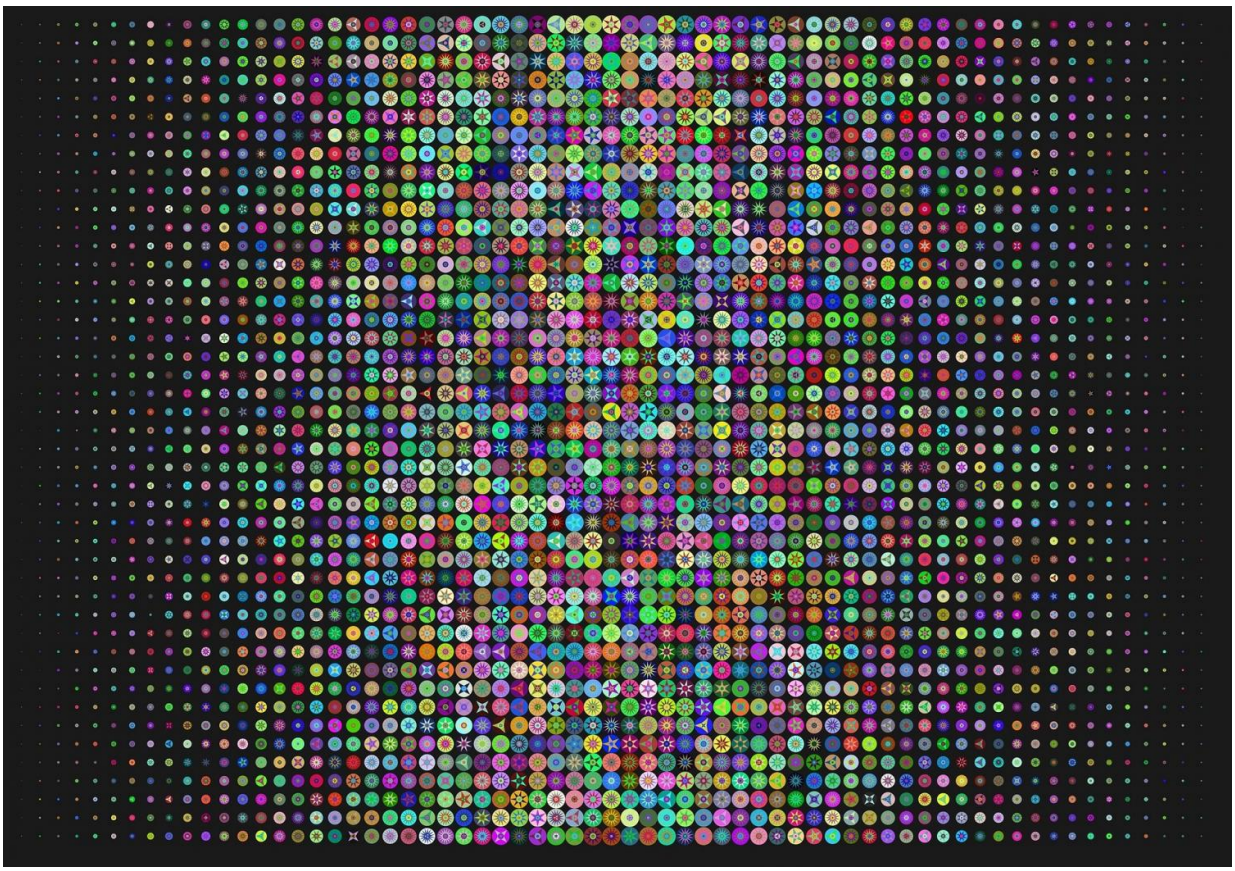
Linguistic changes in gender and ethnic stereotypes correlated with major social movements and demographic changes

April 4 2018, by Alex Shashkevich



Credit: CC0 Public Domain

Artificial intelligence systems and machine-learning algorithms have

come under fire recently because they can pick up and reinforce existing biases in our society, depending on what data they are programmed with.

But an interdisciplinary group of Stanford scholars turned this problem on its head in a new *Proceedings of the National Academy of Sciences* paper published April 3.

The researchers used word embeddings – an algorithmic technique that can map relationships and associations between words – to measure changes in gender and ethnic stereotypes over the past century in the United States. They analyzed large databases of American books, newspapers and other texts and looked at how those linguistic changes correlated with actual U.S. Census demographic data and major social shifts such as the women's movement in the 1960s and the increase in Asian immigration, according to the research.

"Word embeddings can be used as a microscope to study historical changes in stereotypes in our society," said James Zou, an assistant professor of biomedical data science. "Our prior research has shown that embeddings effectively capture existing stereotypes and that those biases can be systematically removed. But we think that, instead of removing those stereotypes, we can also use embeddings as a historical lens for quantitative, linguistic and sociological analyses of biases."

Zou co-authored the paper with history Professor Londa Schiebinger, linguistics and computer science Professor Dan Jurafsky and electrical engineering graduate student Nikhil Garg, who was the lead author.

"This type of research opens all kinds of doors to us," Schiebinger said. "It provides a new level of evidence that allow humanities scholars to go after questions about the evolution of stereotypes and biases at a scale that has never been done before."

The geometry of words

A word embedding is an algorithm that is used, or trained, on a collection of text. The algorithm then assigns a geometrical vector to every word, representing each word as a point in space. The technique uses location in this space to capture associations between words in the source text.

"Embeddings are a powerful linguistic tool for measuring subtle aspects of word meaning, such as bias," Jurafsky said.

Take the word "honorable." Using the embedding tool, previous research found that the adjective has a closer relationship to the word "man" than the word "woman."

In its new research, the Stanford team used embeddings to identify specific occupations and adjectives that were biased toward women and particular ethnic groups by decade from 1900 to the present. The researchers trained those embeddings on newspaper databases and also used embeddings previously trained by Stanford computer science graduate student Will Hamilton on other large text datasets, such as the Google Books corpus of American books, which contains over 130 billion words published during the 20th and 21st centuries.

The researchers compared the biases found by those embeddings to demographical changes in the U.S. Census data between 1900 and the present.

Shifts in stereotypes

The research findings showed quantifiable shifts in gender portrayals and biases toward Asians and other ethnic groups during the 20th century.

One of the key findings to emerge was how biases toward women changed for the better – in some ways – over time.

For example, adjectives such as "intelligent," "logical" and "thoughtful" were associated more with men in the first half of the 20th century. But since the 1960s, the same words have increasingly been associated with women with every following decade, correlating with the women's movement in the 1960s, although a gap still remains.

The research also showed a dramatic change in stereotypes toward Asians and Asian Americans.

For example, in the 1910s, words like "barbaric," "monstrous" and "cruel" were the adjectives most associated with Asian last names. By the 1990s, those adjectives were replaced by words like "inhibited," "passive" and "sensitive." This linguistic change correlates with a sharp increase in Asian immigration to the United States in the 1960s and 1980s and a change in cultural stereotypes, the researchers said.

"The starkness of the change in stereotypes stood out to me," Garg said. "When you study history, you learn about propaganda campaigns and these outdated views of foreign groups. But how much the literature produced at the time reflected those stereotypes was hard to appreciate."

Overall, the researchers demonstrated that changes in the word embeddings tracked closely with demographic shifts measured by the U.S. Census.

Fruitful collaboration

The new research illuminates the value of interdisciplinary teamwork between humanities and the sciences, researchers said.

Schiebinger said she reached out to Zou, who joined Stanford in 2016, after she read his previous work on de-biasing machine-learning algorithms.

"This led to a very interesting and fruitful collaboration," Schiebinger said, adding that members of the group are working on further research together.

"It underscores the importance of humanists and computer scientists working together. There is a power to these new machine-learning methods in humanities research that is just being understood," she said.

More information: Nikhil Garg et al. Word embeddings quantify 100 years of gender and ethnic stereotypes, *Proceedings of the National Academy of Sciences* (2018). DOI: 10.1073/pnas.1720347115

Provided by Stanford University