Positive triggering method reduces nationality bias in large text generators

April 27 2023

Large language models that use internet files to learn how to respond to user prompts about different countries worldwide repeat biased ideas – both positive and negative – found online. Using positive trigger words, like “hopeful” and “hardworking,” when entering prompts can retrain the models and result in less biased responses, according to Penn State researchers. Credit: Pixabay

Humans aren't the only ones learning toxic ideas online. New research led by Penn State researchers reveals that large language models that use internet files to learn how to respond to user prompts about different countries worldwide repeat biased ideas—both positive and negative—found online.

For example, asking for information about higher income countries yields responses with words such as "good" and "important," while asking about lower income countries yields words such as "terrorist" and "dangerous." The team found that using positive trigger words, like "hopeful" and "hardworking," when entering prompts can retrain the models and result in less biased responses.

"Large language models like GPT-2 are becoming a big deal in language technologies and are working their way into consumer technologies," said Shomir Wilson, assistant professor of information sciences and technology. "All language models are trained on large volumes of texts that encode human biases. So, if we're using them as tools to understand and generate text, we should be aware of the biases that come with them as they sort of place a lens on how we view the world or speak to the world."

The researchers asked OpenAI's GPT-2, a precursor to ChatGPT and GPT-4, to generate 100 stories about the citizens of each of the 193 countries recognized by the United Nations to understand how the language model looks at nationality. They chose GPT-2 because its training data is freely available for analysis, unlike later models whose training data has yet to be released. They found that a country's population of internet users and economic status had a significant impact on the types of adjectives used to describe the people.

"Part of my enthusiasm for this research direction comes from the geopolitical implications," Wilson said. "One aspect that my research

team and I discussed early on was: what perspective of the world would this data represent? Would it be an amalgamation of multiple perspectives and, if so, how would they come together? Language technologies are becoming part of the lens of how we understand the world and have many social implications."

Large language models like GPT-2 work by analyzing training data—in this case, web pages linked on the social media platform Reddit—to learn how to respond to user prompts. The language models create responses by taking one word and trying to predict the next word that would logically follow.

The research team used a simple prompt—"[Demonym] people are"—to generate the stories. A demonym is a noun that describes the citizens or inhabitants of a country, such as American or French. The scientists analyzed each batch of 100 stories to identify the most common adjectives associated with each demonym. They compared the AI-written stories to news stories composed by humans to measure the machine model's bias.

They found that the language model used more positive adjectives to describe nations with higher populations of internet users and economic statuses than those with fewer internet users and lower economic statuses. For instance, GPT-2 repeatedly used "good," "important" and "better" to describe the highest scoring countries—France, Finland, Ireland, San Marino and the United Kingdom. The language model used words such as "terrorist," "dangerous" and "poor" to describe the lowest scoring countries.

"Our findings suggest that any text generation model almost always mimics the human biases learned from the training data," said Pranav Venkit, a doctoral candidate in information sciences and technology and lead author of the study. "It's important for software engineers to

understand the data sets they're using to train language models to ensure that the model doesn't have a skewed perception because some group always gets the short end of the stick, which can translate into social harms."

The team also found that adversarial triggering, a method that uses "trigger" words to break a machine learning model, can de-bias the model.

"We used two positive adjectives, hopeful and hardworking, to see how those words affected the model," said Ruchi Panchanadikar, a master's student in information sciences and technology. "For example, instead of giving GPT-2 the prompt 'American people are,' we used 'the hardworking American people are.'"

The trigger words forced GPT-2 to think about what the words "hardworking" and "hopeful" meant in the context of each demonym. The researchers found that the trigger words not only elevated the results of the lowest scoring demonyms, but they caused the demonyms with overly positive scores to dip, resulting in a more unbiased view of each country.

The researchers focused on GPT-2 as a use case, but the findings and adversarial triggering method are likely applicable to any language model trained on large collections of online texts, according to the researchers.

The next step in the research is to study how humans perceive machine-generated biases, said doctoral candidate Sanjana Gautam.

"Assuming GPT-2 and its successors are used in social scenarios, how do people consume the AI-generated data?" she said. "How do these data affect how individuals perceive a country if there's an inherent bias in

the machine model?"

The researchers will present their findings at the 17th Conference of the European Chapter of the Association for Computational Linguistics, which will take place May 2—6 in Dubrovnik, Croatia.

The work is published on the *arXiv* preprint server.

More information: Pranav Narayanan Venkit et al, Nationality Bias in Text Generation, *arXiv* (2023). DOI: 10.48550/arxiv.2302.02463

Conference: 2023.eacl.org/

Provided by Pennsylvania State University