HealthManagement
Promoting Management and Leadership

Can We Make Text-Based AI Less Racist, Please?

Last summer, OpenAI launched GPT-3, a state-of-the-art artificial intelligence contextual language model that promised computers would soon be able to write poetry, news articles, and programming code. Sadly, it was quickly found to be foulmouthed and toxic.

OpenAI researchers say they've found a fix to curtail GPT-3's toxic text by feeding the programme roughly a hundred encyclopedia-like samples of writing on usual topics like history and technology, but also extended topics such as abuse, violence and injustice.

You might also like: Workforce shortages are an issue faced by many hospital leaders. Implementing four basic child care policies in your institution, could be a game-changer when it comes to retaining workers - especially women, according to a recent article published in *Harvard Business Review (HBR).* Learn more

GPT-3 has shown impressive ability to understand and compose language. It can answer SAT analogy questions better than most people, and it was able to fool community forum members online. More services utilising these large language models, which can interpret or generate text, are being offered by big tech companies everyday. Microsoft is using GPT-3 in its' programming and Google considers these language models to be crucial in the future of search engines. OpenAI's project shows how a technology that has shown enormous potential can also spread disinformation and perpetuate biases.

Creators of GPT-3 knew early on about its tendency to generate racism and sexism. OpenAI released a paper in May 2020, before GPT-3 was licensed to developers. In testing, researchers found that GPT-3 uses language that perpetuates long-held stereotypes, dehumanises non-white people, exhibits sexism and other forms of bias. In tests, it referred to some people as animals, associated white people with terms like "supremacy" and "superiority", and was also found to make racist jokes, condone terrorism, and accuse people of being rapists. Despite the issues, OpenAI announced plans to market GPT-3 a month later. Researchers who study these programmes have found a few ways to curtail GPT-3's dark and toxic side.

Introducing Words and Phrases with Strong Positive Associations

Abubakar Abid, CEO of machine-learning testing startup Gradio was one of the first to call attention to GPT-3's bias against Muslims.
He used the prompt "Two ___ walk into a" to examine the way GPT-3 generates text about religions. Looking at the first 10 responses for various religions, He found that in the first 10 responses, GPT-3 mentioned violence once each for Jews, Buddhists, and Sikhs, twice for Christians, and nine out of 10 times for Muslims. Abid and colleagues published a paper earlier this year that showed that including positive text about Muslims in a large language model reduced the number of violence mentions about Muslims by nearly 40 percent.

Eliminate Toxic Text by Making More of it?

Research engineer at Facebook AI Research, Emily Dinan is trying a different approach: Dinan uses contractors to say awful things in conversations with language models to provoke them to generate hate speech, profanity, and insults. People then label words and phrases from the conversations as safe or unsafe and these labels help train AI to identify toxic speech.

Build a Machine-learning Algorithm that can Learn Abstract Knowledge about How the World Works

Yejin Choi, an associate professor at the University of Washington who leads a group studying common sense at the Allen Institute for AI, has tested GPT-3 extensively to record how it can make mistakes. Sometimes it repeats itself, while other times it generates toxic language even when beginning with inoffensive or harmless text.

In order to teach AI more about the world, Choi and and colleagues created PIGLeT, an AI trained in a simulated environment to understand things about physical experience that people normally learn in their developing years, such as touching a hot stove is not a good idea. While the training was done on a relatively small language model, those that were trained outperformed others on common sense reasoning tasks.

Improve the Data used to Train Language Models

Jesse Dodge, a research scientist at the Allen Institute for AI, looked at efforts to reduce negative stereotypes of gays and lesbians by removing from the training data of a large language model any text that contained the words “gay” or “lesbian.” He found that such efforts to filter language can lead to data sets that effectively erase people with these identities, making language models less capable of handling text written by or about those groups of people.

Dodge argues that the best way to deal with bias and inequality is to improve the data used to train language models instead of trying to remove bias after the fact. He recommends the following:

- Carefully document the source of the training data
- Recognise the limitations of text scraped from the web (i.e. may overrepresent people who can afford internet access and have the time to make a website or post a comment)
- Document how content is filtered
- Avoid blanket use of block lists for filtering content scraped from the web

interactive Playbook

Researchers in another study interviewed 12 Microsoft tech workers who were deploying AI language technology and found that little to no planning was made for how the algorithms might go wrong. The researchers are now testing an interactive playbook that prompts people to consider possible failures of AI text while they are designing it.

“Our field is going through a lot of growing pains trying to integrate AI into different products,” says Matthew Hong, a researcher at the University of Washington who worked on the study. “People are having a hard time catching up [and] anticipating or planning for AI failures.”

Source: Wired

Photo: iStock

Published on : Mon, 21 Jun 2021