A New AI Tool Can Help Us Fight AI-Written Fake News, Reviews

Chris Young

Over the past few months, AI tools have raised serious concerns with the way they've been — and can be — used to manipulate the public.

What's the solution? According to researchers from Harvard and MIT, there's a better solution than simply unplugging Skynet.

The best way to beat AI is, in fact, with AI, they say.

RELATED: KEY TAKEAWAYS FROM ELON MUSK'S NEURALINK PRESENTATION: SOLVING BRAIN DISEASES AND MITIGATING AI THREAT

AI vs AI

AI can be used to spread fake news, write fake reviews and to create a pretend mob of social media users aimed at bombarding comments sections with specific agendas.

However, according to MIT researchers, it can now also be used to spot fake artificially-generated text — apparently it takes one to know one.

Though the technology for misinformation is advancing at a worryingly fast pace, the same broad toolset can thankfully be used to catch this type of misinformation. Fake news, deepfakes and twitter bots might have their days numbered by the very technology that helped create them.

Detecting statistical text patterns

Harvard University and MIT-IBM Watson AI Lab researchers recently developed a new tool that spots text that has been generated by AI.

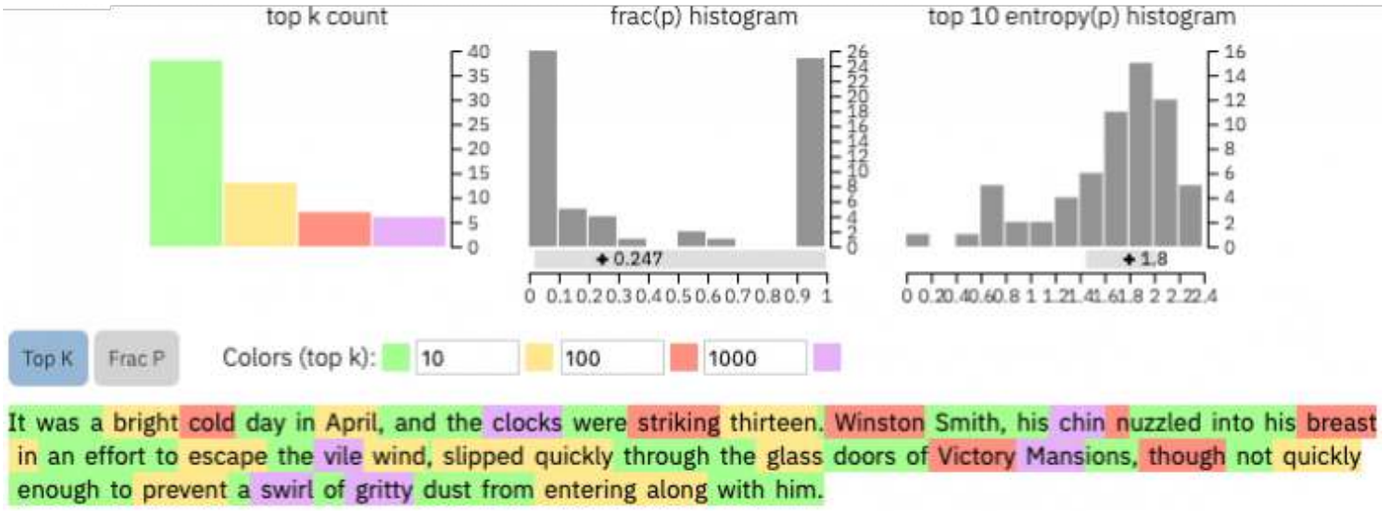
The tool, called the Giant Language Model Test Room (GLTR), takes advantage of the fact that AI text generators use fairly predictable statistical patterns in text.

While these patterns might not be easy to spot for your average reader, it seems that an algorithm can do a pretty good job at it. The AI tool, essentially, can tell if the text is too predictable to have been written by a human.

How does it work?

GLTR tests for predictability in texts by looking at the statistical probability of one word being chosen after another in a sentence.

We ran the opening passage of *1984* through the tool in order to put George Orwell through his paces — and to prove he wasn't really an AI sent back in time from the future.



Source: MIT IBM-Watson AI Lab

Less predictable words are flagged by the color purple — a sprinkling of these throughout a text shows that it was most likely written by a human hand.

Green words, on the other hand, are the most predictable, while yellow and red fall in between.

Of course, we can envision a future in which text-generating machine-learning tools are trained on these purple words in order to trick GLTR by coming across as more human — here's hoping the GLTR researchers can keep up.

AI vs humans?

To test GLTR, the researchers asked Harvard students to identify AI-generated text — firstly with the tool, and then without it.

The students only successfully spotted half of all fake texts on their own. With the tool, on the other hand, they spotted 72%.

While the experiment pitted humans versus AI — an eerie foreshadowing of the future perhaps — one of the researchers says their ultimate goal is different:

“Our goal is to create human and AI collaboration systems,” Sebastian Gehrmann, a Ph.D. student involved in the project, said to MIT in a release about GLTR.

If you want to try out the new tool yourself, you can find it here. A paper detailing the experiments and the new tool was released recently by the researchers.

In the future, huge political unrest could be caused with believable deepfake videos of world leaders, and AI-generated text might be utilized to spread misinformation en masse. Thankfully, the same tech might also provide a solution to the problem.