

Researchers develop a method to identify computer-generated text

July 26 2019, by Leah Burrows



Credit: Petr Kratochvil/public domain

In a world of Deep Fakes and far too human natural language AI, researchers at the Harvard John A. Paulson School of Engineering and Applied Sciences (SEAS) and IBM Research asked: Is there a better way to help people detect AI-generated text?

That question led Sebastian Gehrmann, a Ph.D. candidate at SEAS, and

Hendrik Strobelt, a researcher at IBM, to develop a [statistical method](#), along with an [open access interactive tool](#), to detect AI-generated text.

Natural-language generators are trained on tens of millions of online texts and mimic [human language](#) by predicting the words that most often come after one another. For example, the words "have" "am" and "was" are statically most likely to come after the word "I."

Using that idea, Gehrmann and Strobelt developed a method that, rather than identify errors in text, identifies text that is too predictable.

"The idea we had is that as models get better and better, they go from definitely worse than humans, which is detectable, to as good as or better than humans, which may be hard to detect with conventional approaches," said Gehrmann.

"Before, you could tell by all the mistakes that text was machine-generated," said Strobelt. "Now, it's no longer the mistakes but rather the use of highly probable (and somewhat boring) words that call out machine-generated text. With this tool, humans and AI can work together to detect fake text."

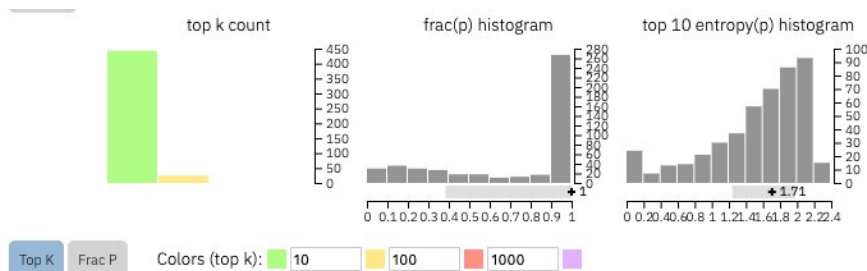
Gehrmann and Strobelt will present their research, which was co-authored by Alexander Rush, Associate in Computer Science at SEAS, at the Association for Computational Linguistics (ACL) conference on July 28th—Aug 2nd.

Gehrmann and Strobelt's method, known as GLTR, is based on a model trained on 45 million texts from websites—the public version of the OpenAI model, GPT-2. Because it uses GPT-2 to detect generated text, GLTR works best against GPT-2, but also does well against other models.

Here's how it works:

If you feed a passage of text into the tool, it highlights the text in green, yellow, red or purple, each color signifying the predictability of the word in the context of the word before it. Green means the word was very predictable, yellow, moderately predicable, red not very predictable and purple means the model wouldn't have predicted the word at all.

So a paragraph of text generated by GPT-2 will look like this:



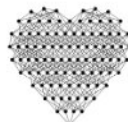
I've been a gamer for over ten years. During that time, I've been involved in a number of games, and I've seen very few of them in the history of the company. My first foray into this was as a member of the U.S. Army. I played some of the games I liked from the early 1980s through the early 1990s, but my first foray into the hobby was at the beginning of 2000 when I was stationed in Afghanistan. After I got back to my hometown and went to school, I started playing games. I began playing multiplayer games, which was a very popular form of gaming. One of the games I started playing was the first-person shooter "The Wolf Among Us" which is still the best-selling title of all time.

I was at the beginning of the game development process. I had already seen a few demos of the game. I was also very interested in the multiplayer aspects of the game, and I wanted to see what the players would do in the game. In the beginning, I didn't know about multiplayer. I thought it would be cool to have some sort of "party game" with some kind of "game mode" which would give the player a real advantage. But as time went on, I realized that there were a lot of different things I wanted to create. To make it fun for the player, the multiplayer component was added. I started playing the game as a member of the U.S. Army. When I returned to my hometown, I found myself in the middle of a war with a group of Taliban soldiers. I was killed by one of the Taliban and I was the only casualty. I decided to take a look at multiplayer. I took the chance to have some fun with the multiplayer. I was in a place that was pretty hostile to the Taliban, and I decided that I wanted to make it fun for the player.

The game was designed to be a good way of showing off combat experience. It was supposed to be a combat-focused game, and I wanted to show off how well the players could play. The multiplayer was designed to be a nice way to show off that. The game is a multiplayer game, and the game is designed to be a fun and interesting multiplayer game.

[Tweet about GLTR](#)

MIT-IBM Watson AI lab and Harvard NLP



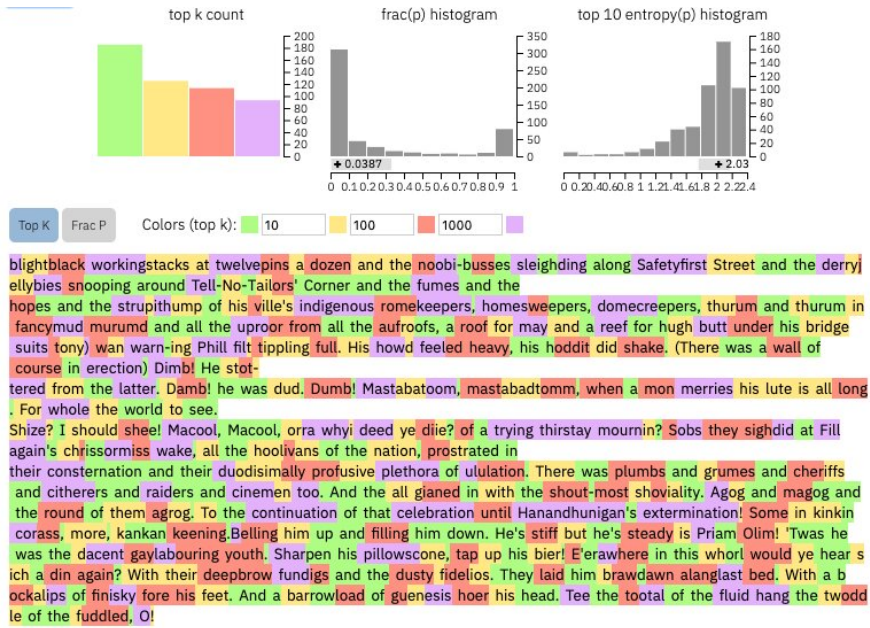
Credit: Harvard University

To compare, this is a real *New York Times* article:

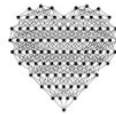


Credit: Harvard University

And this is an excerpt from arguably the most unpredictable human text ever written, James Joyce's *Finnegans Wake*:



[Tweet about GLTR](#)
 MIT-IBM Watson AI lab and Harvard NLP



Credit: Harvard University

The method isn't meant to replace humans in identifying fake texts but rather to support human intuition and understanding. The researchers tested the model with a group of undergraduates in a SEAS Computer Science class.

Without the model, the students could identify about 50 percent of AI-generated [text](#). With the color overlay, the students were able to identify 72 percent.

Gehrmann and Strobel say that with a little training and experience with the program, the number could improve even further.

"Our goal is to create human and AI collaboration systems," said Gehrman. "This research is targeted at giving humans more information so that they can make an informed decision about what's real and what's fake."

Provided by Harvard University

Citation: Researchers develop a method to identify computer-generated text (2019, July 26)
retrieved 2 June 2023 from

<https://techxplore.com/news/2019-07-method-computer-generated-text.html>

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.