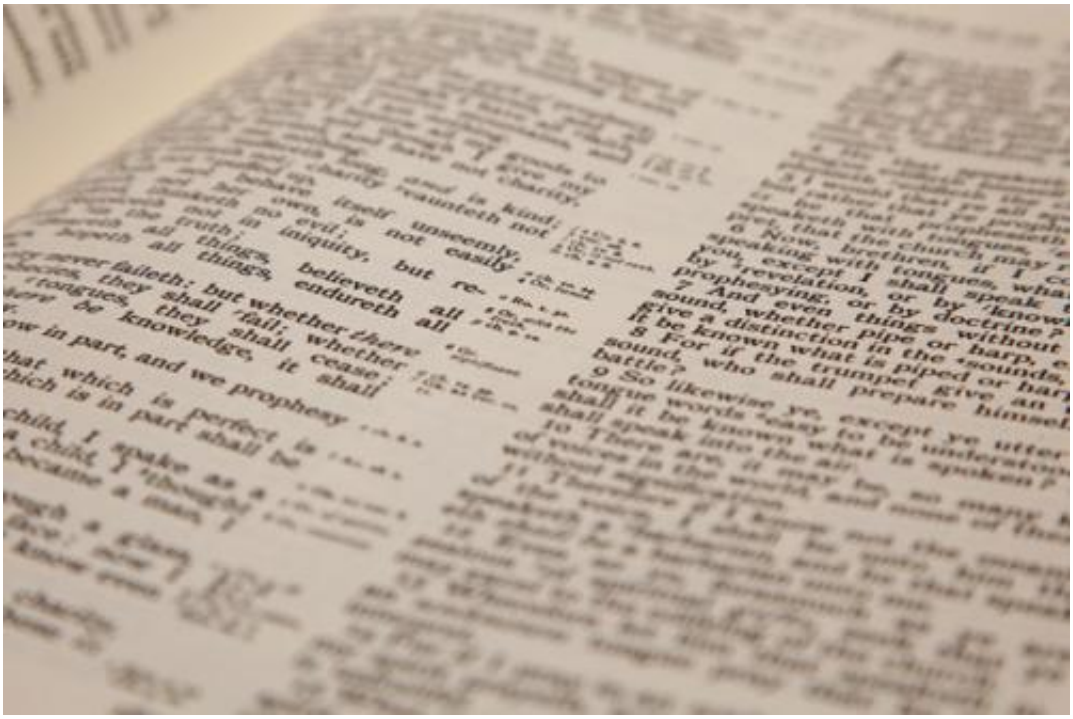
Researchers develop a method to identify computer-generated text

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."

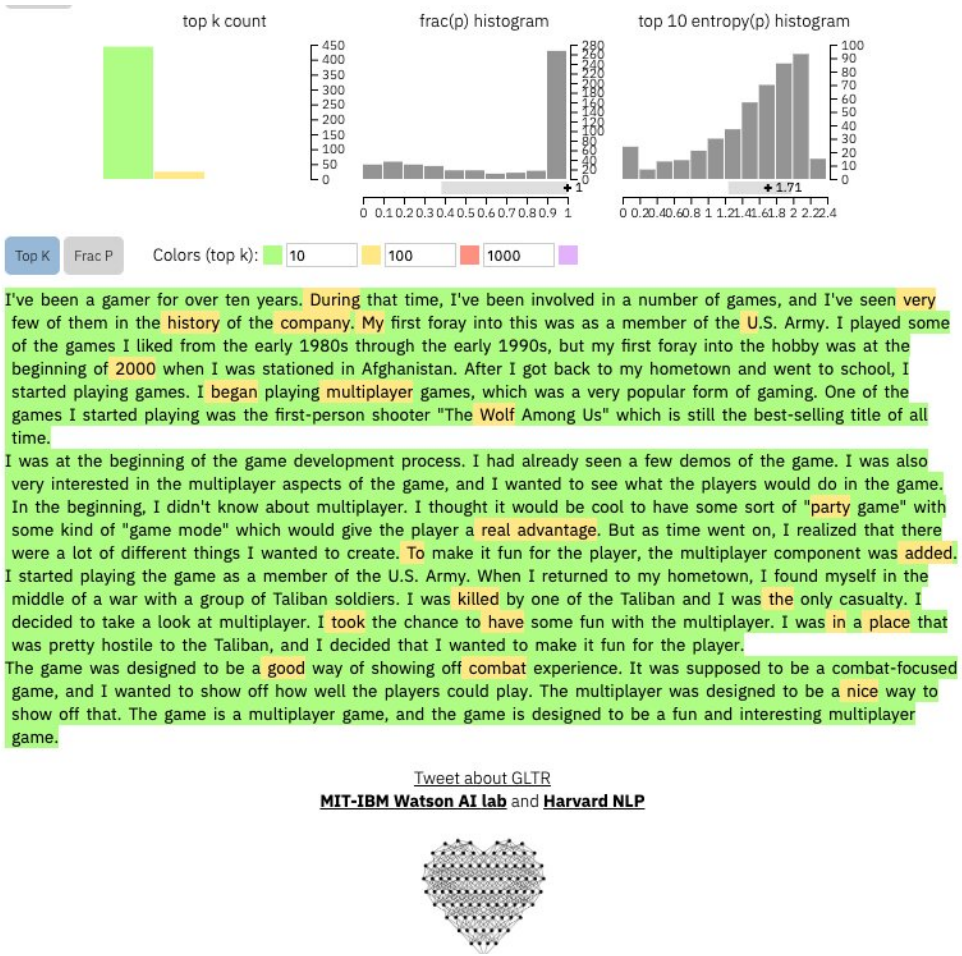
Gehrmann and Strobelt will present their research, which was co-authored by Alexander Rush, Associate in Computer Science at SEAS, at the Association for Computational Linguistics (ACL) conference on July 28th—Aug 2nd.

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:

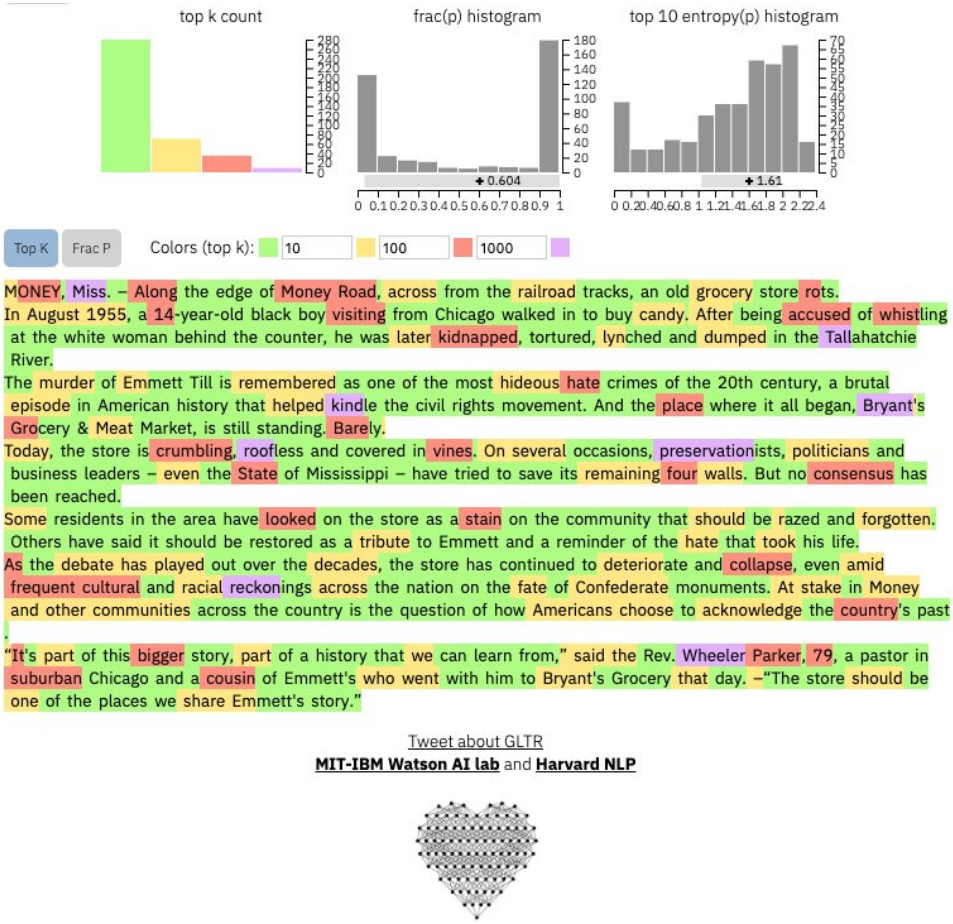
If you feed a passage of text into the tool, it highlights the text in green, yellow, red or purple, each color signifying the predictability of the word in the context of the word before it. Green means the word was very predictable, yellow, moderately predicable, red not very predictable and purple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

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



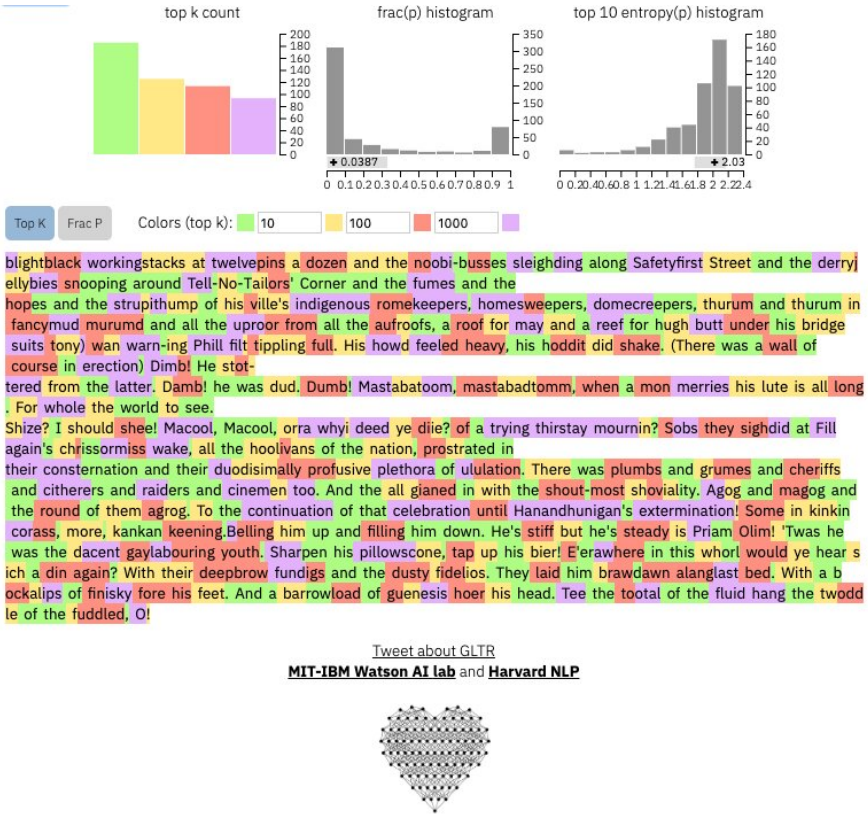
Credit: Harvard University

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*:



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 Strobelt 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 Gehrmann. "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