Q&A: The increasing difficulty of detecting AI- versus human-generated text

May 14 2024, by Mary Fetzer

Credit: Pixabay/CC0 Public Domain

Generative artificial intelligence (AI) tools are used to create text, images and videos, impacting the way society consumes and produces online content. As that technology continues to evolve, it is becoming

increasingly difficult to tell the difference between AI-generated and human-generated content.

Determining the integrity of online information—particularly text—is the current research focus in the Penn State Information Knowledge and wEb (PIKE) Lab, led by Dongwon Lee, professor in the College of Information Sciences and Technology at Penn State.

In a discussion with Penn State News, Lee spoke about the importance of examining the integrity of AI-generated text found on the internet.

Tell us about the motivation for your research.

Broadly speaking, I am interested in the quality of information. AI tools are continually becoming more powerful in terms of generation quality and are capable of creating text that is nearly indistinguishable from human-made content. While there are good uses for such tools, there are concerning implications as well.

In situations that involve privacy and security, for example, it's critical that we know whether something has been written by a human, by AI or by some kind of hybrid.

Further, the rise of fake news and disinformation in recent years makes it important to know where the written content we see on the web is coming from, particularly if we are making decisions based upon that information, and whether such AI-generated content is truthful and fact-grounded or not.

How does AI-generated text compare to text written by humans?

Text generated by AI often exhibits what we have established to be telltale non-human characteristics, but our research shows that people cannot always determine this on their own. In fact, experiments conducted by our lab revealed that humans can distinguish AI-generated text only about 53% of the time in a setting where random guessing achieves 50% accuracy.

When people first get trained on how to differentiate these two types, or even when multiple people work as a team to detect AI-generated text better, the final accuracy does not improve much. Hence, by and large, people cannot really distinguish AI-generated text well.

On the other hand, the best AI solution that we built analyzes text and gives a confident answer—with 85% to 95% accuracy—as to whether content was written by a human or made with AI.

What does that solution look like?

Simplifying it grossly, our solution is a binary classifier, which is a machine learning algorithm that categorizes data into two mutually exclusive groups based on a classification rule. Text is presented, and our software analyzes the text to give us a yes or no answer: yes, it is human; no, it is AI, with some probability score indicating the confidence of the answer.

Our earlier AI solution was largely informed by the linguistic patterns we saw when we looked collectively at human-generated text, such as the frequency with which humans use certain adjectives, formal words and emotional words. When the classifier identifies language patterns that differ from what human writers typically use, we deduce that they are more likely made by AI.

How will your solutions address the evolving improvements in the way AI is used to generate content?

As generative AI tools such as OpenAI ChatGPT and Google Gemini rapidly improve, the quality of the texts generated by these tools also rapidly improves, making it more and more difficult for humans to detect AI-generated text and the integrity of the information in it.

Our latest AI detection solution that achieves the best detection accuracy is made by fine-tuning the most state-of-the-art neural network model. Such a model is called a black box solution, meaning that it functions very well, but we don't fully understand why it is operating well and why AI sees certain text as AI-generated, not human-written.

For simple tasks, it might be okay to not be able to explain the solution's effectiveness. However, for mission-critical tasks in health or military domain, we need to know how an AI model concludes. Therefore, currently, we have a reasonably accurate tool to detect AI-generated text but cannot really explain why it does so. Mitigating this issue and improving our understanding is one of the open challenges for AI researchers.

In the meantime, we are playing cat and mouse with AI tool builders who are creating increasingly sophisticated content generators. The people who are doing the building are not necessarily operating with bad intent, but the things they are creating can be misused and abused, both by curious but honest users as well as malicious adversaries. In the political sector, there is fake news, for example; while in education, students may use AI as a substitute for learning.

As security researchers, we are often one step behind, responding and

reacting to evolving technologies. As we work to develop solutions, we want to position ourselves to head off potential attacks by trying to anticipate our adversary's next move.

AI tools are ubiquitous, and society has to learn to use them in the right way. While solutions to identify AI-generated text continue to evolve, individual users should be mindful about the veracity of the content they encounter and the source of the content, including whether the content was written by humans or AI. We can stave off harm—caused by fake news or misinformation, for example—by asking ourselves if what we're reading makes sense and by checking sources to see if it's true or not.

Provided by Pennsylvania State University

Citation: Q&A: The increasing difficulty of detecting AI- versus human-generated text (2024, May 14) retrieved 29 June 2024 from https://techxplore.com/news/2024-05-qa-difficulty-ai-human-generated.html