INSIGHTS
Cyber made simple.

one brightly cyber

AI TEXT DETECTORS – DO THEY WORK?

BY JAY BORDEN

It is no surprise that Artificial Intelligence, AI, has been growing by leaps and bounds into almost every area of life.

That includes writing texts, reports, letters, commentaries, opinions, social media posts, theses, etc., etc. With so much being written by AI, many people want to know if something they are reading was written by a real person or an AI system.

Where there is a need someone will jump in to try and fill it. In this case many have jumped in offering apps or services claiming to be able to identify AI written material.

Are they accurate? Good question. First, remember that AI system developers are working hard to make their offerings as human-like as possible so that you can't tell.

The AI detection systems look for patterns, phrases, use of certain words and other identifiers that indicate something was generated by an AI system. It works in theory but like with many things, the hype doesn't always match reality.

Researchers from the University of Pennsylvania looking at these systems found that the claims for the AI identifying systems are often exaggerated. The claimed accuracy figures are much higher than the test results.

But detecting AI generated material is not easy. Circumventing some of the current AI detection systems is relatively simple, just insert something unexpected into the text. A graphic or symbol will often do it. However, this technique is not a good fit for all types of writing. It might work for technical writing, or a social media post, but isn't a good fit for school essays for English or history for example, where graphics or symbols would be out of place. The same is true for papers for publication in those subjects.

AI TEXT DETECTORS – DO THEY WORK?

CONTINUED

Another way to fool the detection systems is the use of alternative words from another language. It can be as simple as substituting British English for American English, for instance colour for color.

The researchers looking at these apps found that the samples the AI detection system was trained on had an impact on accuracy.

Some of the AI identification systems even mistakenly identified human-written material for AI generated material.

To improve the accuracy of the AI detection systems and vendor claims of accuracy, the researchers want to add a degree of objectivity to the tests. A uniform way of testing any app professing to identify AI generated material is being proposed. Their intended approach is to create a huge database of 10 million samples in different categories as a standardized test for any AI generated text detector.

To be sure, the vendors are working to improve their systems and not be fooled by the simple tricks given above.

Whether this testing model will come to fruition and how accurate it will be only time will tell. But at least the problem of AI generated material being mistaken for human generated is being considered.

Until then, be skeptical of what you read.

To learn all the ways we can help make your company and family safer, visit onebrightlycyber.com, contact OneBrightlyCyber at info@onebrightlycyber.com, or call (888) 773-1920.