UNIVERSITY OF
SAN FRANCISCO

CHANGE THE WORLD FROM HERE

The Failures of LLM-Generated Text Detection

False Positives Dilute the Efficacy of AI Detection

Themes:

Academic Integrity & Truth

Prerequisites:

- None for the Case Study section
- None for the Technical Exercises

Owner:

Center for AI and Data Ethics at University of San Francisco

Author(s):

Hadley Dixon and Robert Clements

License:

Citation:

Dixon, Hadley and Clements, Robert. (2024). The Failures of LLM Detection: False Positives Dilute the Efficacy of AI Detection.

Objective:

The purpose of this case study is to expose the risks in relying on the use of AI-detection tools to mitigate the issues that have been introduced by the widespread accessibility and use of generative text or large language models (LLMs) through the use of tools such as ChatGPT.

Instructions:

1. Read through the case study individually and then answer the discussion questions as a group, or in small groups.
2. Individually complete the exercises.

Case Study:

Over the past year, there has been a notable rise in the use of large language models (LLMs), like Open AI's GPT-4 and Google's Gemini, in academic settings. These models have been rapidly improving due to advances in natural language processing techniques and generative models and have gained popularity for various academic applications including course content creation, study guides, and grading assistance. After the public launch of ChatGPT in November 2022, the chatbot quickly attracted over one million users, including students of all ages who recognized how it may be used in the classroom for their own benefit. While ChatGPT can be a helpful resource, educators have raised concerns about ethical usage, in the context of academic integrity and plagiarism (Gegg-Harrison, 2023). These issues take two prominent forms. One, instructors are concerned about students using ChatGPT or similar LLMs to generate and submit work as their own without proper citation. This includes explicit copy and pasting, as well as submission without understanding or synthesizing the material themselves. Two, is the concern that ChatGPT generates (copies) text from public works, such as published articles, books, or websites, without needed acknowledgement of its sources. In the process, ChatGPT may inadvertently replicate phrases, sentences, or ideas from existing works that require citation. Speaking generally, teachers are worried about the spread of unoriginal content, the failure to develop critical thinking, research, writing, and reading comprehension skills, and with instilling in students the importance of honesty and the benefits of hard work.

As a quick response to these trends, several LLM detection services have been created to combat plagiarism and ensure academic integrity. However, such services, designed to identify content generated by models like ChatGPT, have exhibited inadequacies and unfair outcomes. In an effort to uphold academic integrity, unintended consequences of relying solely on LLM detection services have surfaced. Imagine the following scenario:

A prestigious university adopted a state-of-the-art LLM detection service mid-spring 2023, called ChatterGuard. This was in response to numerous students caught for violating academic integrity policies during finals week of the previous semester. Instructors, aware of increased use of ChatGPT, recognized obvious changes in students' writing patterns and subsequently questioned them, revealing that many students did in fact turn in final assignments written either entirely or partially by ChatGPT. ChatterGuard was implemented campus-wide to automatically scan and identify potential instances of content generated by LLMs, executed similarly to other plagiarism detectors, such as Turnitin, which look for similarities and matches to preexisting written content.

Right away, ChatterGuard's deployment seemed to be a success; swaths of students were flagged as turning in unoriginal work. However, soon after the implementation, students began contesting these claims and an unexpected issue surfaced–a surge in false positives, with non-native English speakers disproportionately affected. Many students, including those who have not engaged in plagiarism, found themselves accused of academic dishonesty, a serious policy which has grounds for expulsion. Students felt unfairly targeted, leading to a breakdown in trust and communication within their academic community.

Unsure how to move forward, the university turned to both the developer of ChatterGuard and other universities for counsel. Surprisingly, the university found this to be a common trend across campuses that deployed similar technologies, with one west coast school raising, "serious questions about the objectivity of AI detectors and… the potential that foreign-born students and workers might be unfairly accused of or, worse, penalized for cheating" (Myers, 2023). In fact, across 14 different AI detection tools available for use, all scored below 80% accuracy, with only five above 70% (Ramel, 2023).

Discussion Questions:

1. Who are all of the invested parties involved and what are the intentions of each group? Which of these parties have a responsibility to address false-positives?
2. In what ways should the use of LLMs and other generative AI tools, as well as the use of AI-detectors, be acceptable in an academic setting by both students and instructors?
3. What effect does the use of LLMs and corresponding detection tools have on the notions of truth and ownership? What consequences might this have in an academic setting?

Following its research, the university acknowledged the issues and initiated a thorough reassessment of ChatterGuard's detection criteria. This involved collaborating across departments to revise each student's claim, making a more nuanced and accurate assessment of the student's submission. To address and rebuild trust, the university implemented educational initiatives to inform both students and instructors about the limitations of AI-based plagiarism detection services, proper citation practices and responsible use of AI models.

On a larger scale, students' complaints accumulated across the country, pressuring adjustments from the company itself, resulting in a shift in the marketing of ChatterGuard. An official statement was released, stating that the assessment capability reflected *how likely it is* that the input text was AI-generated, not *if* the input text was AI-generated. The only valid conclusion from using ChatterGuard is that the given text *may* have been written by AI, with no guarantee. Instead of simply flagging texts as positive matches, a "likeness score" is outputted, measuring the randomness of the text and the variation in complexity, leaving the onus on the user to determine if further action is necessary. The press release also added disclaimers, cautioning against the use of ChatterGuard on texts (1) under 1,000 words, (2) written by children under the age of 12, and (3) written by non-native English speakers.

References:

Gegg-Harrison, W. (2023, February 27). *Against the use of GPTZero and other LLM-detection tools on student writing*. Medium. https://writerethink.medium.com/against-the-use-of-gptzero-and-other-llm-detection-tools-on-student-writing-b876b9d1b587

Myers, A. (2023, May 15). *AI-Detectors Biased Against Non-Native English Writers*. Human-Centered Artificial Intelligence. https://hai.stanford.edu/news/ai-detectors-biased-against-non-native-english-writers

Ramel, D. (2023, July 10). *Researchers: Tools to Detect AI-Generated Content Just Don't Work*. Virtualization & Cloud Review. https://virtualizationreview.com/articles/2023/07/10/ai-detection.aspx

Raicu, I. (2023, September 19). *AI-Writing Detectors; A Tech Ethics Case Study*. Markkula Center. https://www.scu.edu/ethics/focus-areas/internet-ethics/resources/ai-writing-detectors/

Walker, P., Syal, R., & Stewart, H. (2019, February 13). *Brexit minister plays down prospect of article 50 extension*. The Guardian. https://www.theguardian.com/politics/2019/feb/13/brexit-minister-downplays-article-50-extension

Technical Exercises:

Let's evaluate the ease or difficulty of detecting AI-generated text. We'll do this in multiple ways.

1) Read the following block of text about Brexit. Do you believe it is (a) completely AI-generated; (b) partially AI-generated; or (c) not at all AI-generated? Why? If you answer (b), highlight the sentences you believe to be AI-generated.

    "The Brexit secretary, Stephen Barclay, has played down the possibility of an extension to article 50 as the UK prepares to leave the European Union. His comments follow reports on Tuesday night that Theresa May's chief negotiator, Olly Robbins, was overheard in a Brussels bar saying MPs would be given a last-minute choice between her deal and a lengthy delay to Britain's departure from the EU.

The prime minister has repeatedly insisted the government intends to leave the EU as planned on 29 March, and on Tuesday urged MPs to hold their nerve while she tried to renegotiate changes to the Irish backstop. An ITV reporter overheard Robbins, the most senior civil servant involved in the Brexit process, appearing to suggest in a late-night conversation that May would wait until March – and then give MPs the choice between backing her, or accepting a long extension to article 50.

In an appearance on BBC Radio 4’s Today programme, Barclay declined to comment on Robbins’s comments, which he said were overheard in a “noisy bar”, but added that an extension was not the government’s plan and would not only be a decision for the UK government.

As the United Kingdom braces itself for its imminent departure from the European Union, Brexit Secretary Stephen Barclay has sought to allay concerns regarding the possibility of extending Article 50. His reassurances come in the wake of reports emerging on Tuesday night, indicating that Theresa May's chief negotiator, Olly Robbins, was overheard discussing potential scenarios in a Brussels bar.

While the British Prime Minister has steadfastly maintained her commitment to the scheduled departure on March 29th, rumors circulated suggesting a last-minute ultimatum for Members of Parliament: either approve May's deal or face a significant delay in Brexit proceedings. The reported remarks attributed to Robbins, the principal civil servant involved in Brexit negotiations, hinted at a strategy where May would defer crucial decisions until March, presenting MPs with the dilemma of supporting her deal or accepting a prolonged extension of Article 50.

During an interview on BBC Radio 4's Today program, Barclay refrained from directly addressing Robbins's purported comments, citing the ambient noise of the bar as a factor, but reiterated the government's stance against an extension. Emphasizing that such a decision would ultimately rest with the UK government, Barclay sought to downplay speculation surrounding any potential postponement of the Brexit deadline.

The incident has further fueled the already intense debate surrounding Britain's withdrawal from the EU, with critics accusing the government of resorting to brinkmanship tactics in a bid to secure parliamentary approval for May's embattled Brexit deal. Against a backdrop of political uncertainty and mounting pressure, May continues to urge MPs to stand firm while she endeavors to renegotiate amendments to the contentious Irish backstop arrangement.

As the countdown to March 29th accelerates, the prospect of an extension to Article 50 remains a subject of conjecture, reflecting the deep divisions and complex challenges that continue to define the Brexit process.”

2) Input the above text into an AI-detector, such as https://gptzero.me/. What result did you get?
   a) When all students have completed this exercise, the instructor will reveal if the text was AI-written or not.

3) Ask ChatGPT (or your favorite LLM) to write a five paragraph report on any topic you want.
   a) Input the text into an AI-detector. What result did you get?
   b) If the AI-detector guessed correctly that the text was AI-generated, input the text into an LLM and ask it to rewrite it for you in a way that it will not be detected by an AI-detector. Input the resulting text into the AI-detector and comment on the response. You may need to iterate on this, and provide the LLM with different instructions, before it is able to fool the AI-detector.
4) (Optional) Input something you've personally written without the aid of AI into an AI-detector and comment on the results.
   a) Did it detect that it was written by a human? If so, is there something about your writing style that might have made it obvious that it was human-written?
   b) Input your text into an LLM and ask it to rewrite it for you in a way that it will not be detected by an AI-detector. Input the resulting text into the AI-detector and comment on the response.
5) There are essentially three approaches that teachers can take when it comes to AI in the classroom:
   a) Ban any use of AI. Enforce the ban through the use of AI detection, supervision or by designing assignments that are difficult for AI to help with or by focusing on in-class hand-written assignments and exams.
   b) Allow the use of AI, but only as an aide, and specify the exact acceptable and unacceptable uses of AI. For example, all writing should be done by the student, but they can use AI for checking grammar and spelling. Enforce this policy using the same tactics from the first approach.
   c) Allow the use of AI with no restrictions.

   Given the results in Questions 1-4, comment on the pros and cons of each of the three approaches above.