Unmasking the Source: Identifying Human vs. ChatGPT-Generated Text through Machine Learning

Prerana Singh,
Student, CSE
B.M. Institute of Engineering And Technology
prernasingh260428@gmail.com

Sameer
Student, CSE
B.M. Institute of Engineering And Technology
sameerrathi0302@gmail.com

Aditya Pratap Singh
Student, CSE
B.M. Institute of Engineering And Technology
pratap.aditya7777@gmail.com

Sonika Vasesi
Assistant Professor, CSE
B.M. Institute of Engineering And Technology
sonikavasesi@gmail.com

Abstract—ChatGPT, a conversational Artificial Intelligence, has the capacity to produce grammatically accurate and persuasively human responses to numerous inquiry types from various fields. Its consumers and applications are both expanding at an unparalleled rate. Sadly, abuse and usage often go hand in hand. Since the words produced by AI are nearly comparable to those produced by humans, the AI model can be used to influence people or organizations in a variety of ways. In this paper, we test the accuracy of various online tools widely used for the detection of AI-generated and Human-generated texts or responses.

Keywords— ChatGPT, Artificial Intelligence, Text Recognition, AI Tools, Human-Generated Text, Information Security.

I. INTRODUCTION

In 2022, Open Artificial Intelligence introduced ChatGPT [1], on November 30, a sizable language model that has demonstrated previously unheard-of performance in comprehending user inquiries and producing text that resembles human speech. The interactive design of ChatGPT gained so much excitement within a couple of days of its release, that many of users all over the globe tested it out. However, the arrival of ground-breaking AI-based chatbots like ChatGPT emphasizes how crucial it is to be able to tell if a piece of writing was produced by an AI or an actual person. Information security and digital forensics may suffer as a result, among other professions. A harmful application of AI, such as the distribution of false information and disinformation or social engineering attempts, must be recognized and safeguarded against, for example, in information security, where the capacity to recognize AI-generated text is crucial. The development of techniques for spotting AI-generated text is crucial to ensuring the reliability and accuracy of the information, especially given that it may be used in sensitive contexts like elections, reports on finances, legal paperwork, or consumer feedback (such as reviews of products, eateries, or movies).

A large body of research is attributed to building detectors for the text generated by AI bots [2], [3], [4], [5], [6], [17], [18], [19]. Furthermore, some claim that their AI-text detector can distinguish the AI-generated text from the human-generated text [11], [8], [9], [10], [14], [15], [17], [12], [13], [16]. On that account, our motivation is to test the various tools (generalized AI-text detectors plus detectors targeting ChatGPT-generated text) available. We will elaborate on each tool and its functionality in the following section.

We assess the accuracy and reliability of different technologies that are designed to differentiate between AI-generated and human-generated answers. Tools that assert to be able to recognize ChatGPT prompts and other AI-generated text detection techniques that do not focus on ChatGPT-generated content are all included in our evaluation. The objective of this evaluation is to compare the various technologies and determine how well these technologies work in spotting AI-generated content.

II. TOOL ANALYSIS

A summary of recent studies on how to tell AI-generated text from human-generated language is given in this section. Various tools have been developed to help discern between human-written

language and text produced by AI; the majority of these tools are free, while some are paid.
Below, we analyze the multiple existing free online tools and some of the alternative paid ones, focusing on their effectiveness in detecting AI-generated text. We examine these tools and delineate their brief functionality.

1) ZeroGPT [8]: Although the software was designed primarily to recognize Open AI text, it has restricted functionality with smaller text. Though the developers claim around 98% accuracy of the tool, on self-evaluation, its accuracy amounts to about 50% only.

2) Hugging Face [13]: The software was made available for the purpose of identifying text produced by ChatGPT. However, it frequently overclassifies material as ChatGPT-written. Moreover, on evaluating the tool, it turned out to have much lower accuracy than the other available tools.

3) Perplexity (PPL) [20] Now a paid tool for some premium features, Perplexity is a commonly used statistic for evaluating the effectiveness of large language models (LLM). It is determined by multiplying the negative average log-likelihood of text under the LLM by an exponential.
A lower PPL number indicates increased predictability for the language model. Large text corpora are used during LLM training to teach them typical language patterns and text structures. Therefore, PPL can be used to assess how well a given text adheres to such common traits.

4) Writefull GPT Detector [15]: This tool, which is mostly used to find plagiarism, can tell whether a passage of text was produced using ChatGPT or GPT-3. Likewise, there is some confusion in the tool's percentage-based system for detecting if the text was generated by AI for both samples produced by people and those produced by ChatGPT.

5) Copyleaks [11]: The software promises to be able to tell whether a text was produced by GPT-3, ChatGPT, humans, or a mix of AI and humans. Text must be 150 characters or more to be accepted by the tool.

6) Writer AI Content Detector [16]: It's a software that may be used with ChatGPT and GPT-3 models. The maximum number of letters that can be evaluated in each test due to this restriction is 1500. However, its usage showed that it wrongly classified ChatGPT-generated text as a human written one.

7) Draft and Goal [12]: The software aims to identify text produced by the GPT-3 or ChatGPT designs, and it can do so in two languages which is English and French.
The input text must be no less than 600 characters long for it to function properly, though. The tool's average detection score is about 50%.

8) Originality.ai [9]: The GPT-3, GPT 3.5 (DaVinci-003), and ChatGPT models can be used with this premium utility.
The tool, however, is limited to sentences of at least 100 words, and it has a tendency to label ChatGPT-generated information as authentic. A paid tool, Originality.ai achieves about 80% accuracy for AI-generated text detection.

9) Content at Scale [10]: The accuracy of the AI checker amounts to about 95%. Its premium version also includes advanced paraphrasing tool, that emulates human-level skills.

III. EVALUATION

This section evaluates publicly accessible tools that can differentiate between AI-generated and human-generated responses.

A. DATASETS

A set of annotated text data used to train and evaluate machine learning models for the task of recognizing and localizing text within documents or images is the dataset for AI text identification, also known as a text detection dataset. Numerous applications, such as optical character recognition (OCR), analyzing documents, scene text identification, and others, depend heavily on text detection. In order to teach machine learning algorithms how to make predictions, it serves as an example. The chosen dataset is used as a standardized benchmark for assessing how accurately the various methods can identify content generated by AI.

B. Evaluation Metrics

We used the following measures to assess and contrast the potency of each strategy:

• True Positive Rate (TPR): The indicator illustrates how sensitively the program detects ChatGPT-generated text. The total number of samples that were correctly identified as being generated text is known as True Positive (TP), whilst the number of samples that were wrongly categorized as being human text is known as False Negative (FN). Consequently, TPR = TP/ TP + FN.

• True Negative Rate (TNR): This statistic shows how accurately the technology can identify texts that were produced by humans. The total number of accurately detected samples is known as True Negatives (TN), while the number of samples that ChatGPT mistakenly identified as being generated is known as False Positives (FP). TNR = TN/TN + FP.

C. Evaluated Tools

On comparing the various tools listed in section II, we find that none of them can accurately identify text that has been generated by AI. Analysis reveals that even for particular categories of prompts and topics, the most efficient online tools for identifying created content can only obtain a success rate of about 90%.

IV. Conclusion

In-depth examination of the various techniques created for identifying ChatGPT-generated text has been done in this study. We evaluate the capacity of various tools to distinguish between responses made by ChatGPT as well as human-generated text through an in-depth examination of the current technologies. The majority of the investigated detectors, with an overall high TNR of over 80% and low TPR, are prone to categorizing any text as being produced by a human, according to our trials, which also demonstrate AI's extraordinary capacity to fool detectors. These discoveries have important repercussions for raising the caliber and reliability of online tools.

This study seeks to stimulate additional research in this important field of study and to encourage the creation of more efficient and precise detection techniques for AI-generated text. Furthermore, our results highlight the significance of careful testing and verification when using AI systems. Ultimately, given the sophistication of AI-generated material, our findings highlight the necessity for ongoing work to increase the precision and durability of text identification systems.

References

1. OpenAI: Chatgpt: Optimizing language models for dialogue, https://openai.com/blog/chatgpt/ (2022).
2. Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, pages 1–12, 2023.
3. Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu ChatGPT Detector Using Linguistic Features.
4. Mohammad Khalil and Erkan Er. Will ChatGPT get you caught? rethinking of plagiarism detection. arXiv preprint arXiv:2302.04335, 2023.
5. Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. Stylometric detection of ai-generated text in twitter timelines. arXiv preprint arXiv:2303.03697, 2023.
6. https://github.com/TSKumarage/Stylo-Det-AI-Gen-Twitter-Timelines.
7. Edward Tian (Princeton). GPTZero. https://gptzero.me/, 2023.
8. ZeroGPT. https://www.zerogpt.com, January 2023.
9. Originality.ai. https://originality.ai/, 2022.
10. Content at Scale. AI DETECTOR. https://contentatscale.ai/ai-content-detector/, 2023. Copyleaks AI Content Detector. Copyleaks.
11. https://copyleaks.com/ai-content-detector, 2023.
12. Draft and Goal. ChatGPT - GPT3 Content Detector. https://detector.dng.ai/, 2023.
13. Hugging Face. Hugging Face ChatGPT-Detection. https://huggingface.co/spaces/imseldrith/ChatGPT-Detection, 2023.
14. OpenAI. https://beta.openai.com/ai-text-classifier, January 2023.
15. Writefull. GPT Detector. https://x.writefull.com/gpt-detector, 2023.
16. Writer.com. AI Content Detector. https://writer.com/ai-content-detector/, 2023.
17. Alessandro Pegoraro et al, To ChatGPT or not to ChatGPT: That is the question, 2023
18. Sandra Mitrovic, Davide Andreoletti, Omran Ayoub: ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for detecting short ChatGPT Generated Text, 2023
19. Niful Islam et al: Distinguishing Human Generated Text From ChatGPT Generated Text using Machine Learning
20. Perplexity.ai: https://www.perplexity.ai/, 2023