Faculty of Natural and Applied Sciences Journal of Computing and Applications
Print ISSN: 3026-8133
www.fnasjournals.com
Volume 1; Issue 1; March 2024; Page No. 10-16.

Evaluating the Efficacy of AI Detection Tools in Assessing Human and AI-Generated Content Variants

***Martins, L.I., Wonu, N., & Victor-Edema, U.A.**
Department of Mathematics/Statistics, Ignatius Ajuru University of Education, Port Harcourt, Nigeria.

***Corresponding author email**: liomartins593@gmail.com

Abstract
Artificial Intelligence (AI) is increasingly being utilized in various aspects of life, including research writing, enabling researchers to employ AI tools for generating texts, data analysis and pattern identification in a large dataset. This study evaluates the efficacy of AI-detection tools in assessing human and AI-generated content variants. Specifically, 15 human-written and 15 AI-generated documents were examined. Eight (8) free AI-detection tools were randomly selected to rate the documents. Through analysis of human-generated content, the study found that all AI-detection tools performed well, with mean ratings indicating consistent accuracy in detecting content authored by humans. Statistical analysis confirmed no significant difference among the tools in their mean ratings of human-generated content, suggesting equal proficiency across the board. Conversely, in assessing AI-generated content, significant variability in tool performance was observed, with some tools demonstrating higher effectiveness than others. This variability highlights the importance of understanding various tool capabilities in differentiating between content types. While tools show consistency in detecting human-generated content, their performance in identifying AI-generated content varies considerably. Addressing these challenges and leveraging AI technology's strengths can enhance content evaluation and verification processes.

Keywords: Artificial Intelligence, AI-Detection Tools, AI-Generated Content, Human-Written Content, Variants

Introduction

Artificial Intelligence (AI) is the branch of computer science that deals with the use of machine learning, natural language processing and algorithm development. These three components work simultaneously to solve different kinds of problems. AI was developed to provide quick and accurate solutions to problems. There has been a rapid growth of the impact and use of AI. It has a wide spread of application in different sectors and the academic environment is not exempted. Since the establishment of AI, there have been tremendous upgrades which led to the invention of different AI applications such as the trending ChatGPT (Chat Generative Pre-trained Transformer). The ChatGPT is a language model created by OpenAI. In its design, it is trained to understand and solve problems. It responds to questions using human-like text. AI is increasingly being used in scholarly studies, with researchers using AI tools to identify and analyze items from datasets to published research papers, and from prospective collaborators to job applicants. While the capacity of these new technologies is enticing, the implications of applying AI in scholarly studies are still being investigated. Scholars can interact with ChatGPT by supplying questions and it responds accordingly based on the knowledge it has been trained to know. In other words, the ChatGPT cannot give inputs beyond what it has been trained to know. It is therefore necessary to note that the ChatGPT may not always give correct answers to questions. According to OpenAI (2024), the ChatGPT may produce inaccurate information about people, places, or facts.

OpenAI introduced the ChatGPT API to enable developers to input ChatGPT-functionality to their applications and as a result, Instacart, Quizlet Q-chat, Snapchat's My AI and Shop by Shopify were all included. When ChatGPT debuted in November 2022, it sparked an almost instant technical panic over the potential effect of artificially intelligent technology (AI) on education. This is not the first time that new technology has caused widespread concern. The calculator's entry into classrooms in the 1980s and the commercialization of the world of the internet in the 1990s had comparable implications. It is critical to realize that ChatGPT did not appear out of nowhere. OpenAI, the business behind it, was launched in 2015. GPT-3 in 2020, as well as earlier versions, GPT-2 in 2019, including the first version

Cite this article as:
Martins, L.I., Wonu, N., & Victor-Edema, U.A. (2024). Evaluating the efficacy of AI detection tools in assessing human and AI-generated content variants. *FNAS Journal of Computing and Applications*, *1*(1), 10-16.

of Generative Pre-trained Transformer (GPT) in 2018. The employment of artificial intelligence techniques does not inherently imply academic dishonesty. It is dependent on the way the tools are put to use. Apps like ChatGPT, for example, can assist reticent writers in generating a preliminary draft, which they can later modify and update. When used in this manner, technology can assist learners in learning. Because ChatGPT outputs frequently contain factual inaccuracies, the text can also be utilized to help students master the competencies of fact-checking and logical thinking. Research has shown that students and most researchers depend on ChatGPT for quick and easy solutions. Students no longer study or research to obtain solutions. For example, a man called Philip Parker wrote about 200,000 books with the aid of computers and programmers. It was found that he produces a book in 20 minutes using a patented process that can produce books quickly with the help of the internet and database. Several written articles were works done by an automatic article generator (AAG). It is found to generate articles, research papers, theses etc. All forms of examination malpractice are now possible with the aid of ChatGPT, such that students sneak their mobile phones into the test or examination hall unnoticed and use them to answer questions. The misuse of the ChatGPT has brought about many forms of academic misconduct or academic dishonesty making it difficult to adhere to esthetical standards. When learners use tools or other people to accomplish assignments on their behalf, they are committing academic dishonesty because they are no longer learning the information for themselves. The key aspect is that it is the learners, not the technology, who are responsible (Abd-Elaal et al., 2019)

Academic integrity is the act of adhering to principles, standards norms or values that enhance honesty in academics (TEQSA, 2022). The absence of integrity in academics has given room to persistent decay in the academic system. Technological advancement in the use of AI makes students lazy. They see no reason to spend time in the library when they can, in a few seconds, get answers from the GPT. Since writing research work, assignments and articles are now possible with AI, Educators and the world generally are concerned with the place of integrity in academics. They all wonder where AI is leading us. It has become a challenge trying to differentiate between works written by a human author and those written by AI-generated text. This imposes a serious threat to academic integrity built over the years. It discourages hard work, diligence and persistence, which are the virtues found in great minds and scholars who have made remarkable records and discoveries through studies. During an interview session, two researchers from the University of Cambridge of the Faculty of Education, Dr Vaughan Connolly and Dr Steve Watson, were asked questions on the relevance of ChatGPT on Education. In their opinion, the ChatGPT should be used as an assistive tool and not one to wholly depend on for knowledge. They also made emphasis on the need for ChatGPT to be trained on diverse datasets from different countries and languages in other to reduce bias. Teachers were also encouraged to train students on the proper use of the ChatGPT. Also, scholars were encouraged to work with OpenAI to ensure that the harmful use of this technology can be minimized. (Kirk, 2023).

Artificial intelligence detection tools are those tools that can be used to detect if a research work, article or assignment was done by AI. Some of the tools developed include; Turnitin, copyleaks, GPTZero, OpenAI’s text classifiers, Content at Scale, Writer’s AI content detector, Giant language model test room (GLTR), Huggins face detector, Originality AI chrome extension, GPT radar. Others are copyscape and plagibot. The recent development of AI has led to several researches being done to investigate its impact across various sectors including academics. A study conducted by Uzun (2023) on ChatGPT and Academic Integrity Concerns: Detecting Artificial Intelligence-Generated Content discussed the various tools and techniques that could be used to determine if a work is AI-generated or human-generated. In his study, he identified some tools that could be used to detect AI-generated content such as Copyleaks, Turnitin, Metadata analysis, and Stylometric analysis amongst many others. Two research questions were used for the study. A literature review approach was used for the study to examine the tools and techniques used in detecting work done by AI. Different academic database was used such as Google Scholar and Web of Science. Data used for the study were generated from previous works without testing.

Gao et al. (2022) in their article on comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence detector, plagiarism detector, and blinded human reviews, ten abstracts from five medical journals, which made up a sample size of fifty (50 abstracts). The ChatGPT was also used to generate other abstracts based on the topics of selected abstracts. The abstract was evaluated using an AI output detector, plagiarism detector, and blinded human reviewers to differentiate between human-generated abstract and AI-generated abstract. From the result of the study, there was a 99.98% detection of AI-generated abstracts using an AI output detector, and a 100% detection of original abstracts using a plagiarism detector. The result was different when a mixture of original and generated abstracts was analyzed. Weber-Wulff et al. (2023) examined some selected detection tools to verify their functions from a general perspective. The study evaluated the tools based on accuracy and error type analysis. The researchers examined the tools to see how effectively they could differentiate between AI-generated content from human-generated content. Twelve detection tools available to the public were used and two other commercial systems such as Turnitin and plagiarismCheck. From the results obtained, the available tools are neither accurate nor reliable. Though their points are somewhat convincing, drawing conclusions based on an existing work without undergoing the process of verification, makes the findings unreliable. It is not the volume of tools evaluated that matters but the effective evaluation of tools with higher claims of accuracy. What the world needs is not many detectable tools but effective and reliable tools that could detect if a work was written by AI or by a human.

Yongqiang al. (2023), did a differential analysis of scientific content generation between AI and humans. A description framework was constructed to differentiate between AI-generated content and human-generated content. The researchers also gathered several publicly available systems to investigate the gap that exists between AI-generated scientific writing and human-generated scientific writing. The results showed that AI-generated content is limited and cannot equate to a human. Elkhatat et al. (2023) investigated the capabilities of some AI detection tools in identifying if the content is AI-written or human-written. Fifteen paragraphs of content were generated using ChatGPT 3.5 and ChatGPT 4 on the topic "cooling towers in the engineering process" and five human-generated contents were also used for the evaluation. AI detection tools used are OpenAI, Writer GPTZero, Writer, Copyleaks and Crossplag. They were used to examine the selected paragraphs. From the results, the tools were efficient in detecting contents generated from GPT 3.5 and GPT 4 accurately but displayed some false positives on human-generated content. The study, however, suggested the need to advance the capabilities of the available detection tools to meet the demand of effectively detecting sophisticated and advanced versions of AI. The result also showed that the tools were adequate in detecting written AI contents from GPT 3.5 rather than GPT 4. The tools struggled to identify contents from GPT 4. Crossplag showed a high detection capacity more than others though it struggled in identifying AI-generated contents from GPT 4. The study also identified the challenge of the tools being able to detect sophisticated and upgraded versions of AI. The study was limited to only the detection of AI-generated content and human-generated content.

Chaka (2023), examined five AI detection tools; GPTZero, Copyleaks, OpenAI text classifier, Writer.Com and Giant language model test room. These tools were examined on how effectively they can detect AI-generated content from ChatGPT, YouChat and Chatsonic. Responses were obtained from these three chatbots using English prompts related to English language study. The responses from these chatbots were translated (by Google) into German, French, Spanish, Southern Sotho, and isiZulu languages and were verified using GPTZero to detect AI-generated content. Similarly, Copyleaks was used to verify AI-generated contents from the same document in Spanish, French and German. From the results, Copyleaks AI content detector performed better compared to the other four. Copyleaks, however, misidentified human-generated content when translated into the five languages. The research showed the limitedness of the tools in detecting various contents generated by AI when translated.

Since protecting academic integrity is of great concern to educators and researchers, there is therefore an urgent need to create or develop a tool that can effectively detect AI-generated text. This will help to differentiate if work was done by a human or by AI.OpenAI and other concerned companies have been working hard to come up with a good AI detection tool. However, some detection tools have been created and launched to detect if a text is written by a human or a content generator. These tools are not yet adequate to depend on. (Appleby,2023). AI is increasingly being used in scholarly studies, with researchers using AI tools to identify and analyze items from datasets to published research papers, and from prospective collaborators to job applicants. While the capacity of these new technologies is enticing, the implications of applying AI in scholarly studies are still being investigated. Many experts have stated that language models such as ChatGPT's accuracy and quality of output are untrustworthy. The generated text may be prejudiced, limited, or wrong at times. However, concerns have been drawn to the accuracy of the existing AI-generated content detection tools in identifying or differentiating content generated by AI and humans. This research was to investigate the capability and accuracy of these tools in detecting AI content.

Aim and Objectives of the Study

The study aims to evaluate the efficacy of AI-detection tools in assessing human and AI-generated content variants. Specifically, the study seeks to:

1. evaluate the difference in the performance of AI detection tools in accurately detecting human-generated content.
2. determine the difference in the performance of AI detection tools in accurately detecting AI-generated content.

Hypotheses

H01: The AI-detection tools do not differ significantly over their mean ratings of human-generated contents

H02: The AI-detection tools do not differ significantly over their mean ratings of AI-generated contents

Materials and Methods

The data used for the study are secondary data obtained from eight AI detection tools; ZeroGPT, Copyleak, Content at scale, OpenAItext classifier, Huggingface, Crossplag, Sampling and Scribbr. Fifteen (15) human written content and AI written content (documents) were examined by the AI detection tools to determine their percentage accuracy. The results obtained from the eight AI detection tools were compared using Analysis of Variance (ANOVA).

Model Formation

A parametric (inferential) statistical technique called Analysis of Variance (ANOVA) is used to evaluate the significance of more than two means derived from a variable's measurement. It is employed to test hypotheses concerning the distinctions between two or more methods. If the means of more than two samples differ too much to be attributable to sampling error, the ANOVA is a useful tool for determining this. To adopt the Analysis of Variance, the assumptions of the parametric tests must be obeyed:

- The samples are independent or correlated
- the samples have nearly equal or equal variances, specifically for small sample
- The variable measured is also normally distributed in the population and the data are interval or ratio data.

The One-Way ANOVA is a single classification analysis of variance. It examines the relationship between one independent variable and one dependent variable. It is used to compare three or more means for groups with equal or unequal numbers of subjects. Suppose there are three AI-detecting tools which rated content, then let M= Tool 1, C= Tool 2 and P= Tool 3. The following procedures can be used to compute one-way ANOVA.

Compute $\sum X_M$, $\sum X_C$ and $\sum X_P$

$$\sum X = \sum X_M + \sum X_C + \sum X_P \qquad (1)$$

$$\sum X^2 = \sum X_M^2 + \sum X_C^2 + \sum X_P^2 \qquad (2)$$

Compute $\left(\sum X\right)^2$ for $\frac{\left(\sum X_M\right)^2}{N_M}$, $\frac{\left(\sum X_C\right)^2}{N_C}$ and $\frac{\left(\sum X_P\right)^2}{N_P}$ (3)

$$\sum\left(\sum \frac{X_i}{N_i}^{2}\right) = \frac{\left(\sum X_M\right)^2}{N_M} + \frac{\left(\sum X_C\right)^2}{N_C} + \frac{\left(\sum X_P\right)^2}{N_P} \qquad (4)$$

Compute

$$\sum\left(\sum \frac{X_i}{N_i}^{2}\right)$$

$$N_{Total} = N_M + N_C + N_P \qquad (5)$$

$$SS_{Total} = \sum X^2 - \frac{\left(\sum X\right)^2}{N_{Total}} \tag{6}$$

$$SS_{Between} = \sum \frac{X_i^2}{N_i} - \frac{\left(\sum X\right)^2}{N_{Total}} \tag{7}$$

$$SS_{Within} = SS_{Total} - SS_{Between} \tag{8}$$

$$df_{Between} = n - 1 \tag{9}$$

$$df_{Within} = N - 3 : \tag{10}$$

$$F - ratio = \frac{SS_{Between}}{df_{Between}} x \frac{df_{Within}}{SS_{Within}} \tag{11}$$

Results

Table 1: Summary of descriptive statistics and ANOVA on the mean difference in the mean rating of AI-detection tools in identifying human-generated contents based on the type of tool utilized (n=15).

[F7, 112=1.676, p=0.122]					**95% CI**	
AI-Detection Tool	**N**	**Mean**	**SD**	**SE**	**LB**	**UB**
ZGPT	15	98.80	2.48	0.64	97.42	100.18
COPYK	15	100.00	0.00	0.00	100.00	100.00
CAS	15	96.33	8.34	2.15	91.72	100.95
OATC	15	96.13	9.06	2.34	91.11	101.15
HUGF	15	86.62	30.18	7.79	69.91	103.33
CROSP	15	100.00	0.00	0.00	100.00	100.00
SAMP	15	83.37	35.40	9.14	63.76	102.97
SCRIB	15	89.60	25.74	6.65	75.35	103.85

Key: ZeroGPT= ZGPT, Copyleak= COPYK, Content at scale= CAS, OpenAItext classifier=OATC, Huggingface= HUGF, Crossplag= CROSP, Sampling= SAMP and Scribbr= SCRIB. UB=Upper bound, LB=Lower Bound, CI=Confidence Interval, SD=Standard Deviation and SE= Standard Error

The results from Table 1 show that the mean rating of Copyleak(COPYK) and Crossplag(CROSP) were both 100.00±00. This was followed by ZeroGPT(ZGPT) which had 98.80±2.48 and the least was Sampling(SAMP) which had 83.37±35.40. The result of the Analysis of Variance (ANOVA) shows that AI-detection tools do not differ significantly over their mean ratings of human-generated contents (F7, 112=1.676, p=0.122). This led credence to the retention of the null hypothesis. This indicates that the eight AI-detection tools were equally accurate in detecting human-written content.

Table 2: Summary of descriptive statistics and ANOVA on the mean difference in the mean rating of AI-detection tools in identifying AI-generated contents based on the type of tool utilized (n=15).

[F7, 112=27.680, p=0.000]					95% CI	
AI-Detection Tool	**N**	**Mean**	**SD**	**SE**	**LB**	**UB**
ZGPT	15	43.65	32.78	8.46	25.50	61.81
COPYK	15	100.00	0.00	0.00	100.00	100.00
CAS	15	33.01	23.85	6.16	19.80	46.22
OATC	15	5.87	11.05	2.85	-0.25	11.99
HUGF	15	78.31	40.43	10.44	55.92	100.70
CROSP	15	86.53	26.28	6.79	71.98	101.09
SAMP	15	99.86	0.35	0.09	99.67	100.05
SCRIB	15	24.73	41.53	10.72	1.74	47.73

Key: ZeroGPT= ZGPT, Copyleak= COPYK, Content at scale= CAS, OpenAItext classifier=OATC, Huggingface= HUGF, Crossplag= CROSP, Sampling= SAMP and Scribbr= SCRIB. UB=Upper bound, LB=Lower Bound, CI=Confidence Interval, SD=Standard Deviation and SE= Standard Error

The results from Table 2 show that the mean rating of Copyleak(COPYK) had a mean rating of 100.00±0.00. This was followed by Sampling (SAMP) which had 99.86±0.35 and the least was OpenAItext classifier (OATC) which had 5.87±11.05. The result of the ANOVA shows that AI-detection tools differ significantly in their percentage ratings of AI-generated contents (F7, 112=27.680, p=0.00). This led credence to the rejection of the second null hypothesis. This indicates that the eight AI-detection tools differed in detecting AI-generated content.

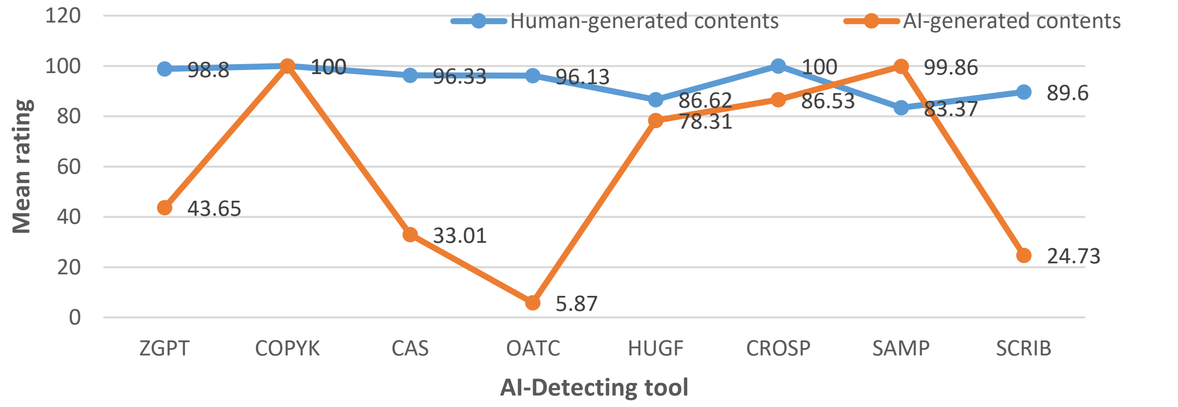

Fig: 1: A line graph showing the variations in the efficacy of the AI-detecting tools in rating Human and AI-generated content variants

Discussion

This study evaluated the efficacy of AI-detection tools in assessing human and AI-generated content variants. The study aimed to assess the consistency of AI detection tools in rating human-generated and AI-generated content variants. The findings revealed notable differences in the performance of these tools between the two content types. Firstly, concerning human-generated content, all eight AI detection tools exhibited relatively consistent performance, with mean ratings ranging from 83.37 to a perfect 100.00. This consistency suggests that the tools are equally proficient in identifying content authored by humans. This finding is consistent with prior research by Gao et al. (2022), who observed high accuracy in detecting human-generated scientific abstracts using AI output detectors and plagiarism detectors. Similarly, our findings align with the work of Yongqiang et al. (2023) and Elkhatatel et al. (2023), who highlighted the limitations of AI-generated content compared to human-generated content. These studies provide context to our findings, emphasizing the reliability of AI detection tools in detecting content authored by humans.

Conversely, when assessing AI-generated content, significant variability in the performance of AI detection tools was observed. Mean ratings ranged from as low as 5.87 to a perfect 100.00, indicating a wide spectrum of effectiveness in identifying AI-generated content. This variability echoes the findings of Chaka (2023), who evaluated AI detection tools across multiple languages and found challenges in accurately identifying AI-generated content, particularly in translated texts. Additionally, Elkhatat et al. (2023) noted the struggle of AI detection tools in identifying content generated by advanced AI models. The present study contributes to this discourse by further elucidating the varying capabilities and limitations of AI detection tools in discerning between different content types.

Moreover, these findings underscore the importance of ongoing research and development efforts aimed at improving the accuracy and reliability of AI detection tools, particularly in identifying AI-generated content. This aligns with the sentiments of Weber-Wulff et al. (2023), who emphasized the need for advancements in AI detection technology to address existing limitations. This study provides valuable insights into the performance of AI detection tools, considering their consistency in rating human-generated content and the variability observed in rating AI-generated content. Through the elucidation of the dynamics of AI detection tools in discerning human and AI-generated content, this study contributes to the ongoing discourse surrounding content validity and veracity in the digital age.

Conclusion

This study reveals that AI-detection tools perform consistently well in identifying human-generated content but exhibit significant variability in detecting AI-generated content. While they excel at discerning human content, their effectiveness varies widely when confronted with AI-generated material. These findings underscore the importance of understanding the nuances of AI tools and highlight the need for ongoing research to improve their performance across diverse content types.

References

Abd-Elaal, E. S., Gamage, S. H., & Mills, J. E. (2019, December). Artificial intelligence is a tool for cheating academic integrity. In *30th Annual Conference for the Australasian Association for Engineering Education (AAEE 2019): Educators becoming agents of change: Innovate, integrate, motivate* (pp. 397-403).

Appleby, C., (2023). The best AI detection tools to catch cheating and plagiarism. https://www.bestcolleges.com/news/best-ai-detection-tools-cheating-plagiarism/

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, Youchat, and Chatsonic: The case study of five AI content detection tools. *Journal of Applied Learning and Teaching, 6*(2)

Elkhatat, A.M., Khaled E., & Saeed, A. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity,*19(17).

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *BioRxiv*, 2022-12.

Kirk, T. (2023). ChatGPT and education. https://www.cam.ac.uk/stories/ChatGPT-and-education.

TEQSA conference (2022). Tertiary education quality and standards agency: https://www.teqsa.gov.au/about-us/news-and-events/our-events/teqsa

Uzun, L. (2023). ChatGPT and academic integrity concerns: Detecting Artificial Intelligence Generated Content. *Language Education & Technology Journal,3*(1),45-54.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1), 26.

Yongqiang, MA., Jiawei, L., Fan, YI., Qikai C., Yong, H., Wei, L.U., & Xiaozhong (2023). AI vs. human – differentiation analysis of scientific content generation. doi.org/10.48550/arXiv.2301.10416