Faculty of Natural and Applied Sciences Journal of Computing and Applications
Print ISSN: 3026-8133
www.fnasjournals.com
Volume 2; Issue 1; September 2024; Page No. 124-133.

Utilization of AI Detection Application Software Packages for Quality Research Integration and Content Assessment in Mathematics Education

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

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

Abstract
The effectiveness of AI-detection technologies in identifying variations in content generated by AI and humans is assessed in this study. In particular, fifteen human-paraphrased and fifteen AI-paraphrased human-generated and AI-generated contents were analyzed. To rate the documents, eight (8) free AI-detection tools were chosen at random. The findings show that, depending on the kind of information being examined (human-paraphrased or AI-paraphrased), there are significant variations in the effectiveness of AI-detection methods. While some tools, like OpenAI Text Classifier, are better at identifying human-paraphrased content, others, like Copyleak and Sampling, are excellent at identifying AI-paraphrased content. The study further shows that manual paraphrasing of human-generated content can reduce the accuracy of AI detection tools. Conversely, AI-generated content remains detectable even when paraphrased using AI, with no significant impact on detection tool performance. These results highlight how crucial it is to use the appropriate AI tool depending on the particular kind of content that needs to be examined to get reliable detection outcomes.

Keywords: AI-Detection Tools, Academic Integrity, Human-Paraphrased Content, AI-Paraphrased Contents.

Introduction

Artificial intelligence (AI) encompasses a broad range of computer science fields, including machine learning, natural language processing and algorithm development. These components work together to address a variety of problems efficiently. In recent years, there has been a lot of discussion about the role of artificial intelligence (AI) in academic research. The research landscape is being transformed by this disruptive technology, which is powered by machine learning algorithms and data analytics. AI has the potential to boost the pace of scholarly discovery and improve the quality of research findings by allowing researchers to handle massive amounts of data, extract important insights, and eliminate repetitive processes. Artificial intelligence (AI) has grown significantly in recent years, resulting in the development of a wide range of AI applications, including content generation. AI-generated content is any sort of content developed using machine learning algorithms or artificial intelligence. (Lavent Uzun, 2023).

Artificial intelligence (AI) plays a significant and growing role in education. One notable example is individualized teaching systems, which are already well-established and have rising proof of their effectiveness in improving learning. AI in education (AIED) systems may also make sophisticated and diversified use of AI to construct the interface that is so vital for the learning experience. The interface, for example, may employ natural language processing and production, speech interfaces, avatars, and video analysis of the learner to assess their attention and emotion. As learners utilize the system, these systems capture data about them. This can be gathered by interaction with the teaching interface as part of learning activities, or information can be gathered using other more recent systems (Ahmad el al., 2022). Furthermore, artificial intelligence has the ability to augment human capacities in academics. It has the ability to automate repetitive operations, allowing researchers to concentrate on higher-level cognitive activity. Data collection, analysis, and even creating manuscripts can all be automated. Scholars may give more time to analytical thinking, hypothesis creation, and investigating new research areas by expediting these processes.

Cite this article as:
Wonu, N., Martins, L.I., & Victor-Edema, U.A.(2024). Utilization of AI detection application software packages for quality research integration and content assessment in Mathematics Education. *FNAS Journal of Computing and Applications, 2*(1), 124-133.

Artificial Intelligence Research Tools:

1. Pictory: is an artificial intelligence-driven video generator that streamlines the process of making and editing videos of excellent quality.
2. Jasper: With its extraordinary features and exceptional quality, Jasper stands out to be the most outstanding AI writing aid on the market.
3. Murf: is a text-to-speech generator. It is recognized as one of the most popular and amazing AI voice generators on the market.
4. HitPaw Photo Enhancer: is an AI-powered tool for improving image quality and features.
5. ChatGPT: is a machine learning algorithm for linguistic processing that produces human-like text responses.
6. Lovo. AI: in its functionality, has received recognition as a prestigious voice creator and text-to-speech solution.
7. Reply.io: Reply is a comprehensive sales interaction system which allows the scalable generation of new possibilities while preserving a personalized touch (Abbadia,2023).

Best AI Detection Tools

The more sophisticated AI generating tools becomes, the more the need to improve on the detecting tools. (Hasker,2023). Though the available detection tools are not yet adequate to be relied on, there are a few of them listed below (Appleby, 2023):

1. COPYLEAKS AI CONTENT DETECTOR: CopyleaksAI Content Detector; was designed to detect GPT-2, GPT-3,GPT-4, ChatGPT, T5. Copyleaks developers claim an accurate detection of 99.1%.
2. TURNLTLIN'SAI DETECTION MODEL: Turnltlin's AI Detection Model is a tool that detects GPT-3 and ChatGPT. It reveals the percentage of AI-generated text contained in a document. This tool was released by Turnitin April 2023. It was basically for the use of educators to enable adequate assessment of students possible. There are two platforms in Turnitin, one for educators and another for students. There are features students can access that educators can. Turnitin claims 98% accuracy in detecting content generated using AI.
3. GPTZERO: GPTZero; in its design, detects ChatGPT, GPT-4, BARD, Llama and other AI models. It detects perplexity (i.e. to measure text's complexity) and burstiness (to compare the length of sentence and variation in complexity) It was designed by Edward Tian on the 3rd of January 2023.
4. OPENAI'S AI TEXT CLASSIFIER: The Open AI's AI text classifier was designed to detect ChatGPT can detect longer texts and can be used by anyone. It is only able to detect about 26% of AI-generated text. This tool was released by OpenAI31st January 2023 and it's still being worked on.
5. CONTENT AT SCALE AI DETECTOR; is a tool that detects GPT-2, GPT-3,GPT- 3.5, ChatGPT and GPT-4. It can take up to 25,000 characters and measures human-test content score. It is considered one of the most popular free detectors.
6. WRITER'S AI CONTENT DETECTOR; is designed to detect GPT-2, GPT-3, GPT3.5 and ChatGPT. It can accommodate about 1,500 inputted characters at a time. It assists writers in publishing their works online.
7. GIANT LANGUAGE MODEL TEST ROOM (GLTR); is designed to detect GPT-2. It was designed by Harvard Natural Language Processing Group and MIT-IBM Waston AI Lab. The tool uses the same model used to generate fake text for detection.
8. HUGGINS FACE OPENAI DETECTOR; was designed in 2019 by OpenAI. It detects AI-generated content, AI images GPT-2, and plagiarism. Its developers claim a detection of 99% of AI-generated text.
9. ORIGINALITY.AI CHROME EXTENSION: This detection tool was developed to detect GPT-3. It is able to detect 94% of GPT-3 generated contents. It was made basically for writers and publishers. It can also detect plagiarism.
10. GPTRADAR; trained its tool using GPT-3. It can detect GPT-3 and GPT-2. It uses the GPT3 language model to determine if a work is AI-generated or by a human.
11. COPYSCAPE, PLAGIBOT: corrector App AI content Detector, Crossplag AI Content; are tools designed to detect plagiarism by showing related text online. It is mostly used by bloggers and website owners to ensure the originality of their work.
12. GRAMMARLY: is a writing tool that uses a machine learning algorithm to help users write well. Its advanced version detects plagiarism.

The advancement of AI has prompted a surge in research aimed at understanding its implications across various sectors, including academia. A study conducted by LeventUzun (2023) on ChatGPT and Academic Integrity

Cite this article as:
Wonu, N., Martins, L.I., & Victor-Edema, U.A.(2024). Utilization of AI detection application software packages for quality research integration and content assessment in Mathematics Education. *FNAS Journal of Computing and Applications, 2*(1), 124-133.

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. This is not helpful as it may cause a serious limitation to the study. From the results of the findings, detecting AI-generated content can lead to ethical issues which I totally disagree with, because other means could be used in such a way that it does not conflict with the ethical policy. Catherine 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 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 plagiarism check. 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. Also, the research did not consider the detection of paraphrased AI-generated content. Effective and reliable tools should not only be able to differentiate between AI-generated and human-generated content but also be able to detect AI content that has been paraphrased or rewritten. Yongqiang et 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. The study is limited in scope and does not show how it could solve the issue of academic integrity.

Elkhatat et al. (2023) investigated the capabilities of some AI detection tools in identifying if 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 contents. 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. It did not investigate paraphrased AI-generated content and rewritten AI-generated content. The need for detection tools that could be able to simultaneously show sensitivity and specificity at a glance, was also suggested for more accurate input.

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) to 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

Cite this article as:
Wonu, N., Martins, L.I., & Victor-Edema, U.A.(2024). Utilization of AI detection application software packages for quality research integration and content assessment in Mathematics Education. *FNAS Journal of Computing and Applications, 2*(1), 124-133.

the results, Copyleaks AI content detector performed better compared to the other four. Copyleaks, however, misidentified the human-generated contents when translated into the five languages. The research showed the limitedness of the tools in detecting various contents generated by AI when translated. The research is okay but did not evaluate paraphrased and rewritten AI-generated content. An exploratory design was adopted for the study. Martins et al. (2024) established that on human-generated content, 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.

Statement of the problem

AI has a huge impact on academia, altering how research is done, information is created, and education is provided. The incorporation of AI technology in academia can expedite operations, improve research outcomes, and stimulate creativity. Data analysis is one of the key ways AI is altering academics. Researchers can use AI algorithms to swiftly and efficiently evaluate massive amounts of data. This allows them to find patterns, correlations, and trends that would be difficult to detect using traditional methods. Furthermore, AI is revolutionizing the research process. It can help researchers with literature reviews and knowledge synthesis by scanning automatically and extracting pertinent information from a broad spectrum of scientific papers. This is not just as AI evolves, academics must adjust and adopt this powerful technology while remaining aware of its limitations and ethical consequences. Researchers can open novel opportunities, enhance scientific knowledge, and contribute to AI's transformational potential in academic research by establishing an equilibrium between AI-driven automation and human brilliance. Education is another area where AI is having a huge impact in academia. Intelligent tutoring platforms, personalized education platforms, and tailored educational experiences can be developed using AI-powered technology. These technologies are capable of analyzing students' learning patterns and providing personalized feedback, help, and resources. The creation of content has grown easier and more accessible as AI technology has advanced. This has resulted in an upsurge in the production of AI-generated content, such as text, photos, and videos. However, it may be necessary to discern between content provided by people and that generated by AI. This study investigates the utilization of AI detection application software packages for quality research integration and content assessment in Mathematics Education.

Aim and Objectives of the Study

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

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

Hypotheses

$H0_1$: The AI-detection tools do not differ significantly over their mean ratings of human-paraphrased human-generated contents

$H0_2$: The AI-detection tools do not differ significantly over their mean ratings of AI-paraphrased 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-paraphrased human-generated content and AI-paraphrased AI-generated content (MS Word documents) were examined by the AI detection tools to determine their percentage accuracy. We used the terms human-paraphrased human-generated content and AI-paraphrased AI-generated content because this study is a follow-up of an earlier study by Martins et al. (2024) were 15 Human-generated, and 15 AI-generated MS Word documents were examined. The results obtained from the eight AI detection tools were compared using Analysis of Variance (ANOVA).

Cite this article as:
Wonu, N., Martins, L.I., & Victor-Edema, U.A.(2024). Utilization of AI detection application software packages for quality research integration and content assessment in Mathematics Education. *FNAS Journal of Computing and Applications, 2*(1), 124-133.

Model Structure

For a one-way ANOVA, we assume that we have *k* different groups, and we want to compare the means of these groups to determine if there is a statistically significant difference between them. The model can be written as:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

Where:

- Y_{ij} is the observation for *j*-th subject in the *i*-th group.
- μ is the overall mean of the data (the grand mean).
- τ_i is the effect of the *i*-th group (also called the treatment effect).
- ϵ_{ij} is the random error term for the *j*-th subject in the *i*-th group, assume to be normally distributed with mean 0 and variance σ^2.

Hypothesis:

- Null hypothesis (H_o): All group means are equal, i.e., $\mu_1 = \mu_2 \ldots = \mu_k$.
- Alternative hypothesis (H_A): At least one of the group's means is different from the others.

Assumptions:

- The observations within each group are normally distributed.
- Homogeneity of variation (equal variance across groups).
- Independence of observations.

Steps:

1. Estimate the grand mean μ.
2. Partition the total variation into:
    - Between-group variation: how much the group mean differ from the grand mean.
    - Within-group variation: how much individual observations differ from their respective group mean.

3. **F-statistics:**

$$F = \frac{\text{Between-group variance}}{\text{Within-group variance}}$$

4. compare the critical F-statistics to a critical value from the F-distribution to determine significance

Cite this article as:
Wonu, N., Martins, L.I., & Victor-Edema, U.A.(2024). Utilization of AI detection application software packages for quality research integration and content assessment in Mathematics Education. *FNAS Journal of Computing and Applications, 2*(1), 124-133.

Results

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

[F7, 112=16.72, p=0.00]					95% CI	
AI-Detection Tool	**N**	**Mean**	**SD**	**SE**	**LB**	**UB**
ZGPT	15	39.07	35.49	9.16	19.42	58.72
COPYK	15	13.33	35.19	9.09	-6.15	32.82
CAS	15	41.66	34.74	8.97	22.42	60.90
OATC	15	93.80	11.82	3.05	87.25	100.35
HUGF	15	88.62	25.22	6.51	74.66	102.58
CROSP	15	61.53	49.09	12.67	34.35	88.72
SAMP	15	0.03	0.08	0.02	-0.02	0.07
SCRIB	15	75.13	38.76	10.01	53.67	96.60

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 indicate that the mean rating of human-paraphrased human-generated contents by OpenAItext Classifier (OATC) was 93.80 ±11.82, which was followed by Huggingface (HUGF) with a mean rating of 88.62 ±25.22 and the least was sampling (SAMP) with a mean rating of 0.03± 0.08. The result of the ANOVA shows that AI-detection tools differ significantly over their mean ratings of human-paraphrased human-generated contents (F7, 112=16.72, p=0.00). This led credence to the rejection of the null hypothesis. This indicates that the eight AI-detection tools were not equally accurate in detecting human-paraphrased human-generated contents.

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

[F7, 112=25.27, p=0.00]					95% CI	
AI-Detection Tool	**N**	**Mean**	**SD**	**SE**	**LB**	**UB**
ZGPT	15	90.60	22.24	5.74	78.28	102.92
COPYK	15	100.00	0.00	0.00	100.00	100.00
CAS	15	61.27	21.86	5.65	49.16	73.38
OATC	15	5.13	11.49	2.97	-1.23	11.50
HUGF	15	31.54	42.23	10.90	8.16	54.93
CROSP	15	84.00	27.31	7.05	68.88	99.12
SAMP	15	99.76	0.58	0.15	99.44	100.08
SCRIB	15	43.53	47.10	12.16	17.45	69.62

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 AI-paraphrased AI-generated contents by Copyleak (COPYK) was 100.00±0.00, which was followed by sampling (SAMP) with a mean of99.76 ±0.58 and the least was OpenAItext Classifier (OATC) with a mean of 5.13 ±11.49. The result of the ANOVA shows that AI-detection tools differ significantly over their mean ratings of AI-paraphrased AI-generated content contents(F7, 112=25.27, p=0.00). This led credence to the rejection of the null hypothesis. This indicates that the eight AI-detection tools were not equally accurate in detecting human-paraphrased human-generated contents.

From Figure 1, the results from Martins et al. (2024) highlight the varying performance of different AI-detection tools when applied to human-generated content. The data were used to compare with the paraphrased versions of the same content. The tools demonstrate significant fluctuations in accuracy when content is paraphrased, revealing that paraphrasing can drastically impact detection reliability. For human-generated content, most tools perform remarkably well, with ZGPT, COPYK, CAS, OATC, CROSP, and SCRIB exceeding 80% accuracy, indicating that these tools are adept at recognizing genuine human writing. COPYK stands out with a perfect 100% detection rate, followed closely by ZGPT (98.8%) and CAS (96.33%).

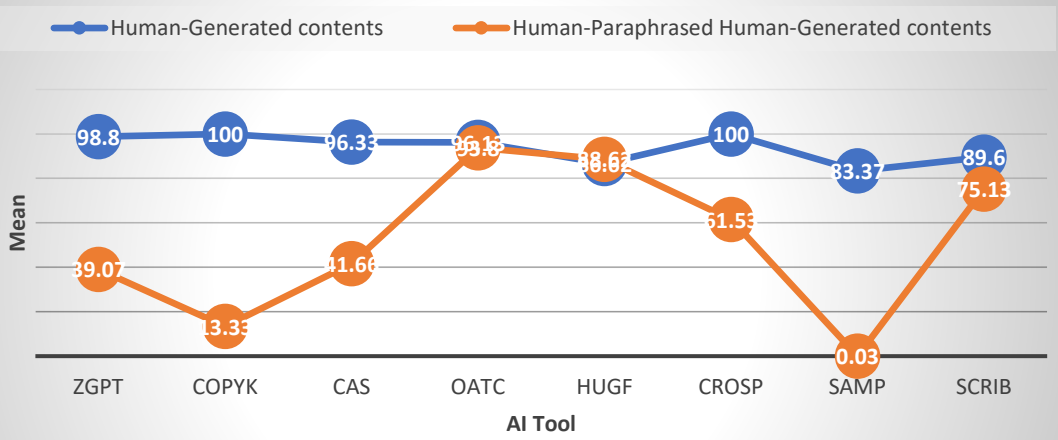

Fig.1: A comparison of Human-paraphrased Human-Generated contents result with data from Martins et al. (2024)

Tools like SAMP (83.37%) and HUGF (86.62%) show relatively lower accuracy but still maintain robust performance. However, paraphrasing severely affects some of these tools' effectiveness. COPYK, for instance, experiences a drastic drop from 100% accuracy to just 13.33%, and ZGPT's detection falls sharply from 98.8% to 39.07%. SAMP is almost completely ineffective against paraphrased content, plummeting from 83.37% to a mere 0.03%. In contrast, certain tools like OATC (93.8%) and HUGF (88.62%) maintain strong performance even with paraphrased content, showcasing greater resilience to content manipulation. CROSP and SCRIB also fare better than most, with accuracy rates of 61.53% and 75.13%, respectively. The analysis reveals that while some AI-detection tools remain reliable after content has been paraphrased, many others suffer significant performance degradation, highlighting paraphrasing's potential to obscure AI-detection.

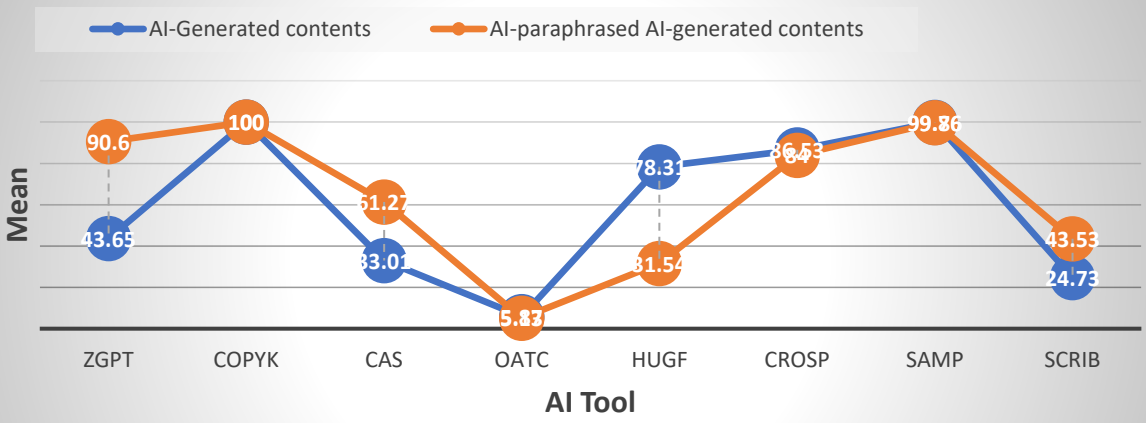

Fig. 2: A comparison of AI-paraphrased AI-generated contents results with data from Martins et al. (2024)

From Fig. 2, the performance of AI-detection tools on AI-generated content, as presented in the data from Martins et al. (2024), shows varying levels of effectiveness in identifying original AI-written texts. The data were used to compare with the paraphrased versions of the same content. The data demonstrates that some tools are highly effective

Cite this article as:
Wonu, N., Martins, L.I., & Victor-Edema, U.A.(2024). Utilization of AI detection application software packages for quality research integration and content assessment in Mathematics Education. *FNAS Journal of Computing and Applications, 2*(1), 124-133.

across both categories, while others show substantial variation when detecting paraphrased AI content. For AI-generated content, COPYK, SAMP, and CROSP show standout performance, with COPYK and SAMP achieving perfect or near-perfect detection rates (100% and 99.86%, respectively), while CROSP follows closely with 86.53% accuracy. Tools like HUGF (78.31%) and ZGPT (43.65%) perform moderately well, while others like CAS (33.01%) and SCRIB (24.73%) struggle with AI-generated content detection. OATC performs particularly poorly, with an accuracy of just 5.87%.When it comes to paraphrased AI-generated content, some tools show consistent or improved performance. COPYK maintains a perfect 100% detection rate, showing that it is highly robust against paraphrasing. Similarly, SAMP retains its high accuracy, barely dropping from 99.86% to 99.76%. ZGPT significantly improves when dealing with paraphrased AI content, increasing from 43.65% to 90.6%. CROSP remains steady, with only a slight decrease from 86.53% to 84%. On the other hand, certain tools show reduced accuracy with paraphrased AI content. HUGF drops significantly from 78.31% to 31.54%, and CAS shows a modest increase from 33.01% to 61.27%. SCRIB also improves from 24.73% to 43.53%, but still struggles with paraphrased content. OATC, however, remains ineffective across both scenarios, with detection rates dropping slightly from 5.87% to 5.13%. While tools like COPYK, SAMP, and CROSP demonstrate high resilience to paraphrased AI content, others like OATC and HUGF exhibit vulnerabilities, suggesting that paraphrasing AI-generated content can either enhance or diminish a tool's detection accuracy depending on the tool's algorithm.

Discussion

The descriptive statistics in Table 1 reveal considerable variation in the performance of AI-detection tools when identifying human-paraphrased human-generated content. Among the tools tested, OpenAI Text Classifier (OATC) had the highest mean rating (93.80 ± 11.82), indicating it was the most accurate at detecting this type of content. This was closely followed by Huggingface (HUGF), which had a mean rating of 88.62 ± 25.22, suggesting it also performed relatively well. In contrast, Sampling (SAMP) had the lowest mean rating (0.03 ± 0.08), suggesting that it was ineffective for this task. The results of the one-way ANOVA show a highly significant difference between the tools in their mean ratings ($F_{7, 112} = 16.72$, $p = 0.00$). This suggests that the AI-detection tools vary significantly in their ability to detect human-paraphrased human-generated content, supporting the rejection of the null hypothesis that all tools perform equally. The wide range of mean ratings highlights that some tools are much better suited to this task than others. This finding is in disagreement with Martins et al. (2024) which established that on human-generated content, 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.

Similarly, Table 2 presents descriptive statistics for the performance of AI-detection tools in identifying AI-paraphrased AI-generated content. Here, Copyleak (COPYK) achieved a perfect mean score of 100.00 ± 0.00, indicating that it consistently identified AI-paraphrased content without error. Sampling (SAMP) also performed near-perfectly, with a mean rating of 99.76 ± 0.58. In contrast, OpenAI Text Classifier (OATC) had the lowest mean rating (5.13 ± 11.49), indicating that it struggled significantly with this type of detection. The one-way ANOVA results again indicate a significant difference between the AI-detection tools ($F_{7, 112} = 25.27$, $p = 0.00$), meaning that not all tools perform equally well in identifying AI-paraphrased AI-generated content. As with the previous case, this result leads to the rejection of the null hypothesis, underscoring the differing capabilities of the tools for this specific task. This finding is in agreement with Martins et al. (2024) which established that 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.

The results from both tables demonstrate significant differences in the performance of AI-detection tools, depending on the type of content being analyzed (human-paraphrased or AI-paraphrased). Some tools, like Copyleak and Sampling, excel at detecting AI-paraphrased content, while others like OpenAI Text Classifier are more effective for detecting human-paraphrased content. These findings underscore the importance of selecting the right tool based on the specific type of content to be analyzed for accurate detection results.

The comparison of AI-detection tools from Martins et al. (2024) as indicated in Figures 1 and 2, reveals how paraphrasing, whether applied to human-generated or AI-generated content, significantly affects the performance of these tools. For human-generated content, most tools initially perform well, with COPYK, ZGPT, CAS, and others

Cite this article as:
Wonu, N., Martins, L.I., & Victor-Edema, U.A.(2024). Utilization of AI detection application software packages for quality research integration and content assessment in Mathematics Education. *FNAS Journal of Computing and Applications, 2*(1), 124-133.

achieving high accuracy rates, some over 80%. However, when the content is paraphrased, many tools, such as COPYK and ZGPT, experience dramatic drops in accuracy, signaling their vulnerability to paraphrasing techniques. On the other hand, a few tools, like OATC, HUGF, and SCRIB, manage to retain higher accuracy, proving more resilient against paraphrased human content. For AI-generated content, the results differ. Tools like COPYK, SAMP, and CROSP excel in detecting both AI-generated and paraphrased AI content, maintaining high or perfect detection rates. Interestingly, ZGPT improves significantly when detecting paraphrased AI content. However, other tools like HUGF and CAS struggle with paraphrased AI content, showing significant declines or only modest improvements in accuracy. OATC consistently performs poorly in both categories, reflecting its weakness in detecting AI-related content. Generally, paraphrasing tends to reduce the detection accuracy for human-generated content, but AI-generated content can sometimes be better identified after paraphrasing. Tools like COPYK, SAMP, and CROSP are more reliable across different content types, while others display significant variability, especially when content is altered through paraphrasing.

Conclusion

The study highlights the significant variation in the accuracy of AI-detection tools when identifying different types of paraphrased content. For human-paraphrased human-generated content, tools like OATC and HUGF performed best, while SAMP was ineffective. In contrast, for AI-paraphrased AI-generated content, COPYK and SAMP excelled, with near-perfect detection rates, while OATC struggled. The one-way ANOVA results confirm that the differences between the tools are statistically significant for both types of content. The study revealed that manually paraphrasing human-generated content can impair the effectiveness of some AI detection tools. However, for AI-generated content, using AI to paraphrase does not significantly impact the performance of AI detection tools in identifying the content as initially generated by AI. No single tool performs uniformly well across all types of paraphrased content, emphasizing the need to choose AI-detection tools based on the specific content type being analyzed. The study underscores the importance of tool selection to enhance detection accuracy in both human and AI-paraphrased content scenarios.

Recommendations

This study made the following recommendations based on the findings

1. Each of the tools should be upgraded regularly to be able to detect content written with more sophisticated versions of AI tools.
2. The tools should be developed in such a way that maintains consistency in detection accuracy.
3. The capacity of tools should be enhanced in such a way that users, especially those academics can rely on their effectiveness.

References

Abbadia, J. (2023). Exploring the Role of AI in Academic Research. Mind The Graph. https://mindthegraph.com

Ahmad, K., Junaid Q., Ala, A., Waleed, I., Ammar E., Driss, B., &Moussa A., (2020). Data-Driven artificial intelligence in education: A comprehensive review [Preprint]. https//doi:10.35542/osf.io/zvu2n.

Appleby,C.(2023). Best AI detection tools to Catch Cheating and Plagiarism. https//www.bestcolleges.com/news/best-ai-detection-tools-cheating-plagiarism/

Catherine, A., Gao, C.A., Howard, F.M, Markov N.S, & Ramesh, S. (2022). Comparing scientific abstracts generated by ChatGPTto original abstracts using an artificial intelligence out detector, plagiarism detector, and blinded human reviewers. https://doi.org/10.1101/2022.12.23.521610.

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).

Hasker, S. (2023). How AI is Enhancing, not Threatening, the Future of Professionals. https://techcrunch.com/2023/10/31/how-ai-is-enhancing-not-threatening-the-future-of-professionals/

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. *Faculty of Natural and Applied Sciences Journal of Computing and Applications*, *1*(1), 10-16. https://www.fnasjournals.com

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

Uzun, L. (2023). ChatGPT and Academic Integrity Concerns: Detecting Artificial Intelligence Generated Content. Language Education &Technology. *LET 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, LU,.&Xiaozhong,,(2023). AI vs. Human – Differentiation Analysis of Scientific Content Generation.arXiv:2301.10416v1. https://doi.org/10.48550/arXiv.2301.10416

Cite this article as:
Wonu, N., Martins, L.I., & Victor-Edema, U.A.(2024). Utilization of AI detection application software packages for quality research integration and content assessment in Mathematics Education. *FNAS Journal of Computing and Applications*, *2*(1), 124-133.