International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I10P121

Volume 72 Issue 10, 216-224, October 2024

Original Article

Analysis of Student Output on the Use of ChatGPT: A Predictive Model Approach

Jovelin M. Lapates[1], Mark Daniel G. Dacer[2], Derren N. Gaylo[3]

[1,2]*College of Technologies, Bukidnon State University, Malaybalay City, Philippines.*
[3]*College of Education, Bukidnon State University, Malaybalay City, Philippines*.

[1]*Corresponding Author : j.lapates@buksu.edu.ph*

Received: 23 May 2024 Revised: 27 September 2024 Accepted: 08 October 2024 Published: 25 October 2024

Abstract *- Artificial Intelligence (AI) has significantly transformed various aspects of education, with AI-powered language models like ChatGPT gaining popularity due to their unique features and advantages. This study aims to analyze student outputs and develop a predictive model to assess whether essay-type answers, Dropbox submissions, and machine problems were generated using ChatGPT, employing machine learning algorithms such as Naive Bayes (NB), Random Forest (RF), and K-Nearest Neighbors (KNN). Student outputs are evaluated using six AI detection tools: Contentatscale, Crossplag, GPTZero, KazanSEO, Sapling, and ZeroGPT. The results are predicted by NB, RF, and KNN, which were chosen for their strong performance in text classification, robustness, and ability to manage non-linear data. The analysis examines performance metrics, including Recall, Precision-Recall Curve (PRC) Area, and Class Accuracy, to provide insights into the predictive capabilities of these models. The findings reveal that NB outperformed the other algorithms, achieving the highest correctly classified instances at 23.19% and a Kappa statistic of 0.1072, indicating slight agreement in classification accuracy, while RF and KNN recorded 14.49% and 15.94%, respectively. Additionally, NB demonstrated the highest true positive rate of 0.232 and PRC area of 0.466, while KNN achieved the best PRC area at 0.566, reflecting varied performance across models. Generally, while Naive Bayes showed superior accuracy and predictive ability, each model has unique strengths that can be leveraged to analyze student outputs and evaluate the use of tools like ChatGPT in educational settings*.

Keywords *- ChatGPT, KNN, Random Forest, Naïve Bayes, AI detector, Students' output.*

1. Introduction

The rapid advancements in technology over the past few decades have significantly changed how people live, work, and learn. Information Technology (IT) has been at the forefront of these transformations and, as a result, has become a critical part of education. In recent years, the integration of Artificial Intelligence (AI) in the education sector has gained significant attention. One of the most remarkable innovations of AI is the development of large language models like GPT-3.5, now GPT 4.0, which has revolutionized natural language processing and text generation. ChatGPT stands as a substantial language model structured upon the Generative Pre-trained Transformer (GPT) architecture. It embodies the form of generative Artificial Intelligence (AI) capable of crafting content grounded in acquired knowledge. Chatbot technology has seen significant progress, with ChatGPT emerging as a prominent AI language model, marking a significant milestone in this evolution [1]. This contrasts with conventional AI, which predominantly focuses on data analysis and inference [2]. ChatGPT, serving as a sophisticated AI-driven resource, has been utilized by students, educators, and academic establishments to enrich learning encounters [3], streamline instructional methods [4], and cultivate academic advancement [5]. The impact of ChatGPT on education spans various dimensions. It serves as a foundation for personalized learning approaches, fosters engaging and immersive educational settings, and advocates for equal access to education. ChatGPT's evolution within the educational sphere signifies a shift from being a traditional teaching aid to an intelligent collaborator, ultimately to an active participant in the learning journey. This transformation represents a significant change in its role and functionality within education [6]. The rapid advancement of AI is transforming the labor market that education aims to serve, sparking concerns about the content and methods of teaching future generations [7]. These concerns emphasize the need for education to equip future citizens with essential skills and competencies to survive in the rapidly evolving society [7]. While ChatGPT has experienced a meteoric ascent and has captivated the attention of both students and educators in the educational space, irresponsible usage of the technology has been a reason for worry. According to [8], utilizing ChatGPT for educational purposes carries a number of dangers associated with education, such as plagiarism, harmful and

biased content, equity and access, the veracity of the AI-generated information, and an excessive reliance on the tool for assessment. Although there are many reasons to be excited and concerned about ChatGPT, educators are starting to think about how it might affect learning.

By now, ChatGPT has significantly altered how various communities, including Software Engineering (SE), view the potential and capabilities of AI technologies [9]. It is crucial to critically examine technology when incorporating new AI tools into the classroom or altering existing ones based on AI tools in order to ascertain both the intended and unforeseen benefits and repercussions of the technology [10]. The integration of ChatGPT in IT education can potentially transform the learning experience for learners, educators, and researchers. However, it also poses some challenges that need to be addressed. Some of the significant challenges of IT education in the use of ChatGPT are quality of generated content, bias and inaccuracy, technical issues, ethics and privacy, and human interaction in which integration of ChatGPT in IT education may also lead to reduced human interaction, which is critical for effective learning. Learners may become over-reliant on the ChatGPT system and fail to engage with educators or peers, which could limit their learning experience.

Research on the use of ChatGPT in education reveals its potential benefits and challenges. It has been found to enhance student engagement, facilitate personalized feedback, and enable innovative teaching methods, thereby increasing accessibility to learning materials. However, concerns regarding academic integrity, including risks of plagiarism and decreased critical thinking skills due to over-reliance on AI-generated content, are significant. The literature emphasizes the necessity for educators' need to develop strategies that mitigate these risks to ensure the responsible use of AI in education. Studies by [14] and [15] further highlight the transformative potential of ChatGPT while addressing challenges in maintaining the authenticity and quality of student outputs [16].

This study aims to analyze student output using ChatGPT and develop a predictive model that accurately assesses student output. Specifically, proponents are analyzing students' outputs from essay-type questions, Dropbox submissions, and machine problems to determine if they were generated using ChatGPT, using machine learning algorithms for the analysis. Additionally, the proponents are developing a predictive model that predicts student performance on a standardized test. The findings of this study provide insights into the potential use of predictive models in improving IT education. The study contributes to the growing literature on AI and education, offering practical recommendations for educators, policymakers, and researchers. Furthermore, it is essential to note that the specific benefits depend on the design, implementation, and effectiveness of the predictive model approach and the study's findings. Additionally, ethical considerations and responsible use of AI are being considered to ensure a positive impact on end-users.

2. Conceptual Framework

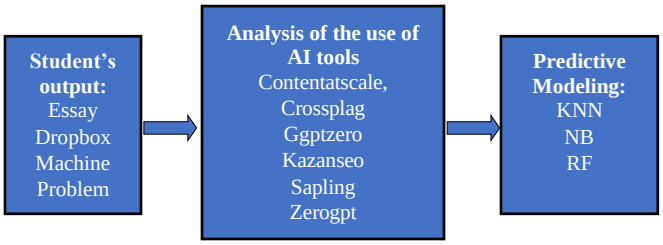

Fig. 1 Conceptual framework of the study

Figure 1 illustrates a flowchart showing the process of analyzing student output (essays, Dropbox submissions, and machine problems) using various AI tools such as Contentatscale, Crossplag, Gptzero, Kazanseo, Sapling, and Zerogpt. This analysis is an intermediary step before applying predictive modeling techniques like KNN, NB, and RF. The flow shows an evaluation of the student's work through AI tools to inform predictive modeling decisions. AI language models, like ChatGPT, have the potential to improve learning, cooperation, and student engagement in the educational setting [11]. The capacity of these models to provide asynchronous communication, which frees students from time limitations to participate in discussions and conversations, is one of their most noteworthy advantages [12]. ChatGPT demonstrates proficiency in forming student groups conducive to collaboration on projects and assignments, nurturing teamwork and problem-solving abilities [13]. ChatGPT holds the potential to play a crucial role in crafting personalized assessments for students. Within the conceptual framework of this study, the application of various AI tools to analyze student output forms a crucial component. These tools encompass algorithms [17] such as KNN [18], which was established on "learning by analogy", that is, by comparing a given test example with training sets that were alike. Both continuous and discrete data were handled by NB [19]. With respect to the quantity of predictors and data points, it was quite scalable. It may be used to create predictions in real-time and is quick. It is insensitive to unimportant details. Easy-to-understand forecasts were generated by RF [20]. Large datasets could be handled by it effectively. Compared to the decision tree algorithm, the RN algorithm predicts outcomes more accurately.

3. Materials and Methods

The study employed data mining techniques. Data mining, as a research design, involves extracting valuable data from student output as datasets to inform decision-making, employing techniques like classification. The data mining aspect of the work relied on students' output, which was collected from major IT subjects from different levels in different types of exams such as essay-type, dropbox submissions, and machine problems. Data mining was employed to derive significant knowledge concerning a particular dataset and to generate essential relationships between variables stored in the dataset. This section provides a brief overview of all the classification algorithms employed in the study.

3.1. Classification Algorithms

There are multiple crucial steps in the procedure. Firstly, prepping the data is crucial. This includes handling missing data by adding or removing it, encoding categorical features with LabelEncoder or OneHotEncoder if needed, and standardizing or normalizing features for distance-metric-based models such as KNN and NB. In order to find and choose pertinent aspects for the model, feature selection is then carried out. The next step involves training different models and fine-tuning their hyperparameters. Examples of these models include KNN, NB, and RF. Finally, the models are evaluated using performance indicators like as precision, recall, and Precision-Recall Curve.

3.1.1. Naive Bayesian

NB classifiers are a class of classification techniques that are produced using the Bayes Theorem. It is composed of multiple algorithms that operate according to the same principle: every pair of features that needs to be classified stands alone. NB applies the Bayes rule in this way:

$$P(y|X) = \frac{P(\square)X|y)P(y)}{P(X)} \quad (1)$$

X is an n-dimensional dependent feature vector, and Y is a class variable.

$$X = (x1, x2, x3, \ldots, xn) \quad (2)$$

In this situation, the class variable (y) can have only two possible outcomes: yes or no. In certain cases, the categorization could be multivariate. We, therefore, need to identify the class Y that has the highest likelihood.

$$y = p(y) \prod ii{=}1\ nn\ pp(XX|yy) \quad (3)$$

NB [21] is a suitable tool for analyzing student output on ChatGPT due to its effectiveness in text classification tasks, computational efficiency, and ability to handle high-dimensional data, such as the wide range of vocabulary used in student responses. It performs well on small datasets, offers interpretability by calculating the likelihood of certain words correlating with specific categories, and is easy to implement without extensive fine-tuning.

3.1.2. K-Nearest Neighbor

Regression and classification issues can be resolved with KNNs, which are supervised machine learning algorithms. New data can be swiftly sorted into precise categories using the KNN model. The KNN algorithm estimates the values of any new data points by using "feature similarity," which measures the distances between a query and each example in the data. It then identifies the K examples that are most similar to the query, chooses the label with the highest frequency (for classification), or averages the labels (for regression). In order to learn, KNN compares a given test tuple with similar training tuples. Every training pair is stored in an n-dimensional pattern space. When a KNN classifier is given an unknown tuple, it searches the pattern space for the k-training tuples that are closest to it. The k "nearest neighbors" of the unknown tuple are these k-training tuples.

The KNN model parameters are as follows:

n_neighbors: The number of neighbors to consider with (default=5).

Weights: The distance (near neighbors have more influence) or uniform (all neighbors are weighted equally).

p: Power parameter for Minkowski distance where:

p=2 is Euclidean,

p=1 is Manhattan.

KNN is helpful in analyzing student output on ChatGPT due to its simplicity, non-parametric nature, and flexibility in handling numerical and textual data. It effectively identifies patterns and clusters [22] among students based on their interactions with ChatGPT, helping to reveal similarities in behavior or learning outcomes [23].

3.1.3. Random Forest

RF algorithm [24] is a versatile and powerful machine-learning technique commonly used for classification and regression tasks. In order to function, it builds several decision trees during the training stage. A random feature selection and a fraction of the training data are used to construct each tree in the forest. During prediction, the algorithm aggregates the predictions of each tree to arrive at a final output. This aggregation, often through averaging for regression tasks or voting for classification tasks, helps improve the overall accuracy and robustness of the model. Additionally, RF is resilient to overfitting, can handle large datasets efficiently, and provides insights into feature importance, making it widely favored for various applications in data science and predictive analytics.

The RF model parameters include:

n_estimators: is the number of trees in the forest.

max_depth: the maximum depth of the tree (control overfitting).

min_samples_split: The minimum number of samples required to split an internal node.

min_samples_leaf: The minimum number of samples required to be at a leaf node.

Random Forest is an effective method for analyzing student output on ChatGPT due to its capability to handle numerical and categorical data, robustness against overfitting, and high accuracy in predictions and classifications. It can identify essential features influencing student outcomes, making it versatile for both classification and regression tasks [25].

3.2. Datasets

Student output from essay-type exams, group activities, and machine problem assessments undergo analysis using six AI tools, namely, Contentatscale, Crossplag, Gptzero, Kazanseo, Sapling, and ZeroGpt. The dataset comprises 18 essays, 12 group activities, and 39 machine problems. Three different data mining predictive models, KNN, NB, and RF, were used and compared in this study. KNN classified data points based on the majority class of their nearest neighbors,

NB utilized probabilistic principles assuming feature independence, and RF constructed an ensemble of decision trees to improve predictive accuracy. These models were integral to the data mining process, each with unique strengths and applications. KNN's simplicity and effectiveness made it suitable for various tasks, NB excelled in text classification with its probabilistic approach, and RF ensemble nature enhanced accuracy and mitigated overfitting. This study used a data mining approach to discover likely hidden valuable and unknown patterns from a collection of data. Through the aid of WEKA, data were analyzed, preprocessed, and summarized to identify relationships. The statistical analysis plan for the three data mining predictive models—KNN, NB, and RF [17]—involves initial data preprocessing, such as feature scaling and handling missing values, followed by model training and evaluation using cross-validation to ensure generalizability. The analysis included performance metrics such as accuracy, precision, recall, and F1 score to assess the models' predictive capabilities, focusing on selecting the most appropriate algorithm based on the dataset characteristics and research objectives.

The accuracy and precision of the three data mining models were tested and validated, which delivered meaningful information about what kind of data mining models were the best fit for students' output data mining analysis.

Table 1. Students' output from essay-type exams subjected to six AI tools

Contentatscale	Crossplag	Gptzero	Kazanseo	Sapling	ZeroGpt
0.0%	0.0%	1.0%	0.2%	8.0%	0.0%
93.0%	0.0%	1.0%	0.4%	5.0%	0.0%
1.0%	80.0%	3.0%	0.3%	19.4%	7.5%
100.0%	100.0%	50.0%	0.2%	24.4%	21.0%
49.0%	18.0%	17.0%	2.8%	21.5%	2.7%
0.0%	0.0%	0.0%	0.1%	0.0%	0.0%
0.0%	0.0%	2.0%	0.2%	0.8%	5.4%
0.0%	0.0%	0.0%	0.3%	4.1%	0.0%
0.0%	0.0%	19.0%	0.2%	0.0%	17.3%
100.0%	90.0%	52.0%	93.1%	94.3%	17.6%
1.0%	18.0%	49.0%	0.4%	14.5%	20.8%
0.0%	0.0%	0.0%	4.8%	22.8%	10.6%
0.0%	0.0%	0.0%	0.4%	14.6%	6.8%
50.0%	100.0%	0.0%	12.0%	24.4%	4.7%
0.0%	0.0%	7.0%	0.3%	27.6%	0.0%
100.0%	100.0%	50.0%	63.5%	90.7%	10.3%
1.0%	0.0%	0.0%	0.2%	0.0%	5.0%
0.0%	100.0%	0.0%	0.2%	0.0%	0.0%

Table 2. Students' output from group activity subjected to six AI tools

Contentatscale	Crossplag	Gptzero	Kazanseo	Sapling	ZeroGpt
0.0%	86.0%	34.0%	33.3%	39.0%	27.8%
0.0%	80.0%	51.0%	0.9%	46.2%	10.7%
20.0%	100.0%	50.0%	17.9%	66.2%	36.8%
0.0%	0.0%	5.0%	0.2%	29.0%	5.2%
20.0%	0.0%	55.0%	1.8%	42.9%	27.7%
25.0%	100.0%	50.0%	25.2%	39.8%	6.2%
0.0%	0.0%	19.0%	2.2%	39.0%	3.1%
0.0%	82.0%	0.0%	51.2%	1.3%	0.0%
0.0%	78.0%	0.0%	1.2%	0.0%	0.0%
0.0%	0.0%	0.0%	41.1%	0.8%	0.0%
0.0%	0.0%	16.0%	1.7%	0.0%	0.0%
0.0%	0.0%	0.0%	2.0%	0.0%	0.0%

Table 3. Students' output from machine problems subjected to six AI tools

Contentatscale	Crossplag	Gptzero	Kazanseo	Sapling	ZeroGpt
0.0%	100.0%	1.0%	97.2%	0.0%	4.2%
0.0%	100.0%	1.0%	98.6%	0.0%	2.0%
0.0%	100.0%	0.0%	99.3%	70.7%	9.6%

0.0%	100.0%	3.0%	95.6%	0.0%	9.2%
0.0%	100.0%	2.0%	99.3%	2.8%	23.0%
0.0%	100.0%	3.0%	97.3%	1.9%	0.0%
0.0%	100.0%	2.0%	96.2%	0.0%	7.1%
0.0%	100.0%	12.0%	98.4%	100.0%	0.0%
0.0%	100.0%	3.0%	60.6%	0.0%	42.9%
0.0%	100.0%	1.0%	98.0%	1.8%	1.5%
0.0%	100.0%	2.0%	92.5%	0.0%	12.8%
0.0%	100.0%	7.0%	52.4%	0.0%	31.9%
0.0%	100.0%	3.0%	86.0%	0.0%	31.6%
0.0%	100.0%	1.0%	99.0%	1.3%	21.8%
0.0%	100.0%	3.0%	94.0%	0.5%	20.7%
0.0%	100.0%	2.0%	99.2%	100.0%	9.4%
0.0%	100.0%	7.0%	99.8%	5.1%	28.3%
0.0%	100.0%	2.0%	47.6%	26.8%	19.3%
0.0%	100.0%	2.0%	99.0%	22.7%	32.2%
0.0%	100.0%	2.0%	90.5%	0.0%	29.6%
0.0%	100.0%	1.0%	99.0%	100.0%	22.7%
0.0%	100.0%	2.0%	99.0%	98.4%	8.3%
0.0%	100.0%	2.0%	90.2%	0.4%	19.9%
0.0%	100.0%	3.0%	98.2%	19.6%	32.3%
0.0%	100.0%	3.0%	39.3%	0.1%	17.2%
0%	100.0%	1%	16%	0%	30%
0%	100.0%	5%	1%	0%	24%
0%	100.0%	1%	73%	6%	0%
0%	100.0%	0%	19%	1%	0%
0%	100.0%	1%	3%	0%	8%
0%	100.0%	10%	55%	0%	0%
0%	100.0%	1%	34%	0%	35%
0%	100.0%	1%	61%	0%	25%
0%	100.0%	1%	49%	0%	0%
0%	100.0%	3%	99%	0%	5%
0%	100.0%	6%	21%	0%	5%
0%	100.0%	6%	3%	17%	0%
0%	100.0%	2%	28%	0%	3%
0%	100.0%	2%	98%	4%	59%

3.3. Performance Metrics

Metrics such as False Positive Rate (FP Rate), Recall, and Precision-Recall Curve (PRC) Area [26] are crucial for analyzing student output on the use of ChatGPT, as they provide insights into model performance and the accuracy of classifications. TP Rate and Recall help assess how effectively the model identifies relevant positive outputs, while FP Rate highlights potential misclassifications that could lead to misleading conclusions about student satisfaction. These metrics are precious in handling imbalanced datasets, focusing on the performance of the positive class, which is essential for informed decision-making and continuous improvement in the educational applications of ChatGPT. The formulas are as follows:

False Positive Rate

$$FP\ Rate = \frac{FP}{FP+TN} \tag{4}$$

Recall

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

Precision-Recall Curve

$$PRC\ Area = \int_0^1 Precision\ (r)dr \tag{6}$$

These metrics provide essential insights into the performance and reliability of models analyzing student outputs on ChatGPT, enabling educators to understand interactions better and enhance learning experiences.

4. Results and Discussion

The presentation of the study's findings is followed by discussions that cover observations, model testing, and evaluations that were done during the investigation.

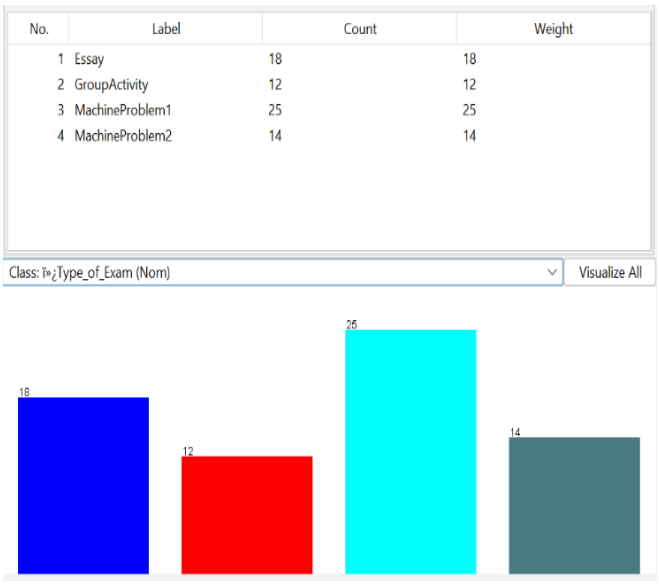

Fig. 2 Frequency of distribution

4.1. Data Visualization

The data processing phase involves the examination of seven attributes related to the type of exam and utilizing six distinct AI tools: Contentatscale, Crossplag, Gptzero, Kazanseo, Sapling, and Zerogpt. The dataset comprised 69 instances categorized as nominal type, with 52 distinct attribute values observed. Among these values, 49 were identified as unique, accounting for 71% of the dataset. Visual representations, in the form of bar graphs, as shown in Figure 2, were generated to visualize the frequency distribution of attribute values, providing insights into the dataset's composition and characteristics.

4.2. Analysis of AI Tools

The AI tools utilized in the analysis included a) Contentatscale, b) Crossplag, c) Gptzero, d) Kazanseo, e) Sapling, and f) Zerogpt. These six AI tools stand out as some of the best in their respective domains. Their combined capabilities offer comprehensive solutions for various tasks in AI. Each tool played a vital role in processing and interpreting the data, contributing to a comprehensive understanding of the dataset's attributes and patterns. Identifying AI-generated text can be challenging due to specific indicators: repetition of words and phrases, limited depth, and a propensity for inaccuracies or outdated information [13]. AI outputs often lack the nuanced understanding and creativity of human-generated content, resulting in a more robotic and generic tone [27].

The results obtained from running the dataset through WEKA using Contentatscale, Crossplag, Gptzero, Kazanseo, Sapling, and Zerogpt reveal varying performances across different metrics. Contentatscale achieved the highest score in attribute 1 with 43, indicating its proficiency in analyzing that attribute. Crossplag excelled in attribute 3 with a score of 45, showcasing its strength in detecting similarities and discrepancies within the dataset.

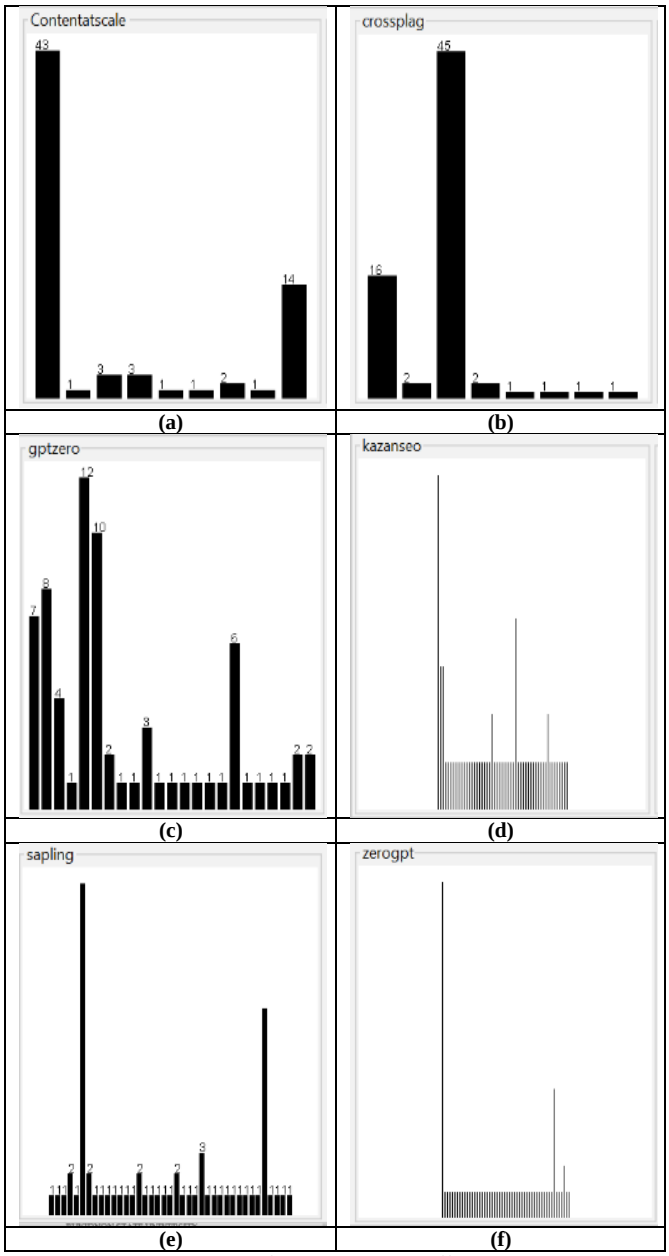

Fig. 3 AI tools performance across different matrix

Table 4. Processing time

Predictive Modeling Algorithm	Time
Naïve Bayesian	0.01 seconds
Random Forest	0.11 seconds
KNN	0.01 seconds

Gptzero demonstrated versatility by obtaining relatively high scores across multiple attributes, particularly in attributes 5 and 6. Kazanseo displayed consistent performance across several attributes, with notable scores in attributes 1, 2, and 6. Sapling exhibited strong performance in attribute 6, indicating its effectiveness in enhancing and refining the dataset. ZeroGpt achieved the highest score in attribute 1 with 45, showcasing its capability in analyzing and interpreting data.

Table 5. Stratified Cros Validation

Estimates	NB	RF	KNN
Correctly Classified Instances	23.19%	14.49%	15.94%
Incorrectly Classified Instances	76.81%	85.51%	84.06%
Kappa statistic	0.1072	0.0148	0.0762
Mean absolute error	0.0324	0.036	0.0349
Root mean squared error	0.1371	0.1378	0.143
Relative absolute error	87.02%	96.60%	93.69%
Root relative squared error	100.46%	100.99%	104.79%
Total Number of Instances	69	69	69

Table 6. Accuracy performance measures

Weighted Avg.	NB	RF	KNN
FP Rate	0.127	0.127	0.074
Recall	0.232	0.145	0.159
PRC Area	0.466	0.262	0.566
Class	0.151	0.106	0.159

The time taken to build the predictive models they were varied significantly across different algorithms. While NB and KNN models were constructed almost instantaneously, RF took slightly longer, approximately 0.11 seconds. These differences in processing times highlight the varying computational demands and complexities associated with each predictive modeling algorithm.

The results of the Stratified Cross Validation illustrate the performance of three predictive modeling algorithms: NB, RF, and KNN. Despite differences in the absolute metrics values, such as correctly classified instances and error rates, all three algorithms show relatively low accuracy rates, with NB achieving the highest at 23.19%. However, the Kappa statistic, which measures agreement beyond chance, indicates only marginal agreement for all algorithms.

Furthermore, the high relative absolute error and root relative squared error percentages suggest considerable deviations from the actual values in the predictions generated by each algorithm. The accuracy performance measures, including FP Rate, Recall, PRC Area, and Class, provide insight into the predictive capabilities of three algorithms: NB, RF, and KNN. While the FP Rate and Recall metrics indicate the algorithms' abilities to classify instances and avoid false positives correctly, the results suggest that all three algorithms struggle to achieve high accuracy rates, with NB performing slightly better than RF and KNN. However, the PRC Area metric reveals substantial variation in precision-recall trade-offs across the algorithms, with NB exhibiting the highest area under the curve.

5. Conclusion

The analysis of student output using ChatGPT through a predictive model approach reveals several key findings. Despite employing various algorithms such as NB, RF, and KNN, the predictive accuracy of the models remains relatively low. While NB exhibited slightly better performance of correctly classified instances, the overall accuracy rates were modest across all algorithms. Additionally, the precision-recall curve analysis highlighted significant variations in the algorithms' precision-recall trade-offs, with NB showing the highest area under the curve.

However, the study emphasizes the need for further refinement and exploration of alternative modeling approaches to enhance predictive accuracy and reliability in assessing student output with ChatGPT. These findings underscore the complexities inherent in analyzing student-generated text data and suggest avenues for future research to improve predictive modeling outcomes in educational settings.

Recommendation

This study offers valuable insights into the use of AI-powered language models like ChatGPT in educational settings by employing machine learning algorithms to detect AI-generated content in student outputs. The comparison of NB, RF, and KNN algorithms, along with the evaluation using six AI detection tools, provides a framework for assessing student submissions.

Given that NB outperformed other models in classification accuracy, it is recommended to further explore and refine this algorithm for broader applications in AI content detection.

However, since each model demonstrated unique strengths, future research should consider combining these models to enhance overall performance in detecting AI-generated content. Additionally, academic policy-making bodies should consider the results of this study when developing guidelines on the use of AI in education.

Ethical Considerations

The researchers sought faculty members handling IT major subjects to obtain voluntary participation in the study concerning students' output. This decision was based on providing sufficient information about the study's purpose, methods, demands, risks, and potential benefits, ensuring that participants fully understood both the research and the implications of their involvement.

The communication process was not merely a formality but aimed to establish a mutual understanding between the researchers and the faculty regarding the data gathered from student outputs. The researchers emphasized that faculty members could share or withhold student output and could withdraw from the study at any time without penalty. Throughout the study, the researchers ensured that confidentiality and anonymity were strictly maintained.

Acknowledgment

The paper is made possible with the generous funding of the Center for Education and Advocacy, Bukidnon State University's Research Unit, BukSU Research Committee, and BukSU Research Ethics Committee for their invaluable insights and guidance in shaping the direction of this study.

References

[1] Partha Pratim Ray, "ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope," *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 121-154, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Woondeog Chang, and Jungkun Park, "A Comparative Study on the Effect of Chatgpt Recommendation and AI Recommender Systems on the Formation of a Consideration Set," *Journal of Retailing and Consumer Services*, vol. 78, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Gunther Eysenbach, "The Role of Chatgpt, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation with ChatGPT and a Call for Papers," *JMIR Medical Education*, vol. 9, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ahmed Tlili et al., "What if the Devil is My Guardian Angel: Chatgpt as a Case Study of Using Chatbots in Education," *Smart Learning Environments*, vol. 10, pp. 1-24, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Yogesh K. Dwivedi et al., "So What if Chatgpt Wrote It?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy," *International Journal of Information Management*, vol. 71, pp. 1-63, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Hao Yu, "The Application and Challenges of ChatGPT in Educational Transformation: New Demands for Teachers' Roles," *Heliyon*, vol. 10, no. 2, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Sushil Kumar Sharma, Shailendra C. Jain Palvia, and Kuldeep Kumar, "Changing the Landscape of Higher Education: From Standardized Learning to Customized Learning," *Journal of Information Technology Case and Application Research*, vol. 19, no. 2, pp. 75-80, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[8] Torrey Trust, Jeromie Whalen, and Chrystalla Mouza, "ChatGPT: Challenges, Opportunities, and Implications for Teacher Education," *Contemporary Issues in Technology and Teacher Education*, vol. 23, no. 1, pp. 1-23, 2023. [Google Scholar] [Publisher Link]

[9] Lasha Labadze, Maya Grigolia, and Lela Machaidze, "Role of AI Chatbots in Education: Systematic Literature Review," *International Journal of Educational Technology in Higher Education*, vol. 20, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ben Williamson et al., Chapter 25: *Critical Perspectives on AI in Education: Political Economy, Discrimination, Commercialization, Governance and Ethics*, Handbook of Artificial Intelligence in Education, Edward Elgar Publishing, pp. 553- 570, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Abderahman Rejeb et al., "Exploring the Impact of ChatGPT on Education: A Web Mining and Machine Learning Approach," *The International Journal of Management Education*, vol. 22, no. 1, pp. 1-14, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[12] Chenglu Li, and Wanli Xing, "Natural Language Generation using Deep Learning to Support MOOC Learners," *International Journal of Artificial Intelligence in Education*, vol. 31, pp. 186-214, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[13] Flori Needle, AI Detection: How to Pinpoint AI Generated Text and Imagery [+ Detection Tools], Hubspot. [Online]. Available: https://blog.hubspot.com/marketing/ai-detection#detect-ai-text

[14] Md. Mostafizer Rahman, and Yutaka Watanobe, "ChatGPT for Education and Research: Opportunities, Threats, and Strategies," *Applied Sciences*, vol. 13, no. 9, pp. 1-21, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Maha Zayoud et al., "Impact of ChatGPT on Education: Challenges and Opportunities," *International Conference of Management and Industrial Engineering*, vol. 11, pp. 75-85, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Marta Montenegro-Rueda et al., "Impact of the Implementation of ChatGPT in Education: A Systematic Review," *Computers*, vol. 12, no. 8, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Vraj Sheth, Urvashi Tripathi, and Ankit Sharma, "A Comparative Analysis of Machine Learning Algorithms for Classification Purpose," *Procedia Computer Science*, vol. 215, pp. 422-431, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Wentian Kang et al., "ChatGPT-based Sentiment Analysis and Risk Prediction in the Bitcoin Market," *Procedia Computer Science*, vol. 242, pp. 211-218, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[19] Or Peretz, Michal Koren, and Oded Koren, "Naive Bayes classifier, An Ensemble Procedure for Recall and Precision Enrichment," *Engineering Applications of Artificial Intelligence*, vol. 136, pp. 1-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Atsushi Mizumoto, Sachiko Yasuda, and Yu Tamura, "Identifying ChatGPT-Generated Texts in EFL Students' Writing: Through Comparative Analysis of Linguistic Fingerprints," *Applied Corpus Linguistics*, vol. 4, no. 3, pp. 1-11, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Harry Zhang, "The Optimality of Naive Bayes," *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, Miami Beach, Florida, USA, 2004. [Google Scholar] [Publisher Link]

[22] N.S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, vol. 46, no. 3, pp. 175-185, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[23] Prajwal Singh, Diving into K-Nearest Neighbors (KNN) with ChatGPT, Medium, 2024. [Online]. Available: https://medium.com/@prajwlsingh/diving-into-k-nearest-neighbors-knn-with-chatgpt-d938b32d03aa

[24] Ernest Yeboah Boateng, Joseph Otoo, and Daniel A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *Journal of Data Analysis and Information Processing*, vol. 8, no. 4, pp. 341-357, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[25] Andy Liaw, and Matthew Wiener, *Classification and Regression by Random Forest*, R news, vol. 2, pp. 18-22, 2002. [Google Scholar] [Publisher Link]

[26] Jesse Davis, and Mark Goadrich, "The Relationship between Precision-Recall and ROC Curves," *Proceedings of the 23rd International Conference on Machine Learning*, New York, United States, pp. 233-240, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[27] Kai Riemer, and Sandra Peter, "Conceptualizing Generative AI as Style Engines: Application Archetypes and Implications," *International Journal of Information Management*, vol. 79, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]