I.J. Modern Education and Computer Science, **2025, 3, 45-58**
Published Online on June 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2025.03.04

AI vs. Human Writing: Developing a Novel Method for Text Authenticity Detection in Education

Vijay H. Kalmani*
Department of Computer Science and Engineering, Rajarambapu Institute of Technology, Rajaramnagar, Affiliated to Shivaji University, Kolhapur, Maharashtra – 415414, India
Email: vijaykalmani@gmail.com
ORCID iD: https://orcid.org/0000-0003-0738-3211
*Corresponding Author

Amol C. Adamuthe
Department of Information Technology, Kasegaon Education Society's, Rajarambapu Institute of Technology, Shivaji University, Sakharale, Maharashtra – 415414, India
Email: amol.admuthe@gmail.com
ORCID iD: https://orcid.org/0000-0003-2483-7051

Arati Premnath Gondil
Department of Information Technology, Rajarambapu Institute of Technology, Rajaramnagar, Affiliated to Shivaji University, Kolhapur, Maharashtra – 415414, India
Email: aratigondil2003@gmail.com
ORCID iD: https://orcid.org/0009-0005-1311-5055

Vaishnavi Prashant Patil
Department of Information Technology, Rajarambapu Institute of Technology, Rajaramnagar, Affiliated to Shivaji University, Kolhapur, Maharashtra – 415414, India
Email: vaipp8830@gmail.com
ORCID iD: https://orcid.org/0009-0008-0978-4311

Riya Amar Kore
Department of Information Technology, Rajarambapu Institute of Technology, Rajaramnagar, Affiliated to Shivaji University, Kolhapur, Maharashtra – 415414, India
Email: riyakore4@gmail.com
ORCID iD: https://orcid.org/0009-0005-3055-8931

Vaishnavi Mahadev Metkari
Department of Information Technology, Rajarambapu Institute of Technology, Rajaramnagar, Affiliated to Shivaji University, Kolhapur, Maharashtra – 415414, India
Email: vaishnavimetkari1012@gmail.com
ORCID iD: https://orcid.org/0009-0003-8149-2130

Received: 05 December, 2024; Revised: 11 February, 2025; Accepted: 18 March, 2025; Published: 08 June, 2025

Abstract: Rapid progress in generative artificial intelligence (AI) technologies has brought forth stupendous challenges in differentiating AI-written text from human text. The Naturalness Score, a composite measure that considers lexical diversity, syntactic complexity, sentiment variability, and grammatical faults, is a new idea that emerged from this study. The Naturalness score is part of a larger machine learning framework, although it does have an individual classifier called the Naturalness-Based Logistic Regression Classifier or NLRC. The NLRC model was analyzed against a large, diverse corpus of nearly 45,000 text samples, most of which were student essays, articles, and web-scraped content. The proposed model outperformed all existing baseline models with an accuracy of 96.41%, precision of 0.98, recall of 0.95, and F1 score of 0.96. The high areas under the receiver operating characteristic curve (AUC=1.00) and precision-recall curve (AUC-PR) also indicate the effectiveness of the model in differentiating AI generated from human-written text.

The proposed approach offers several advantages including increased detection accuracy, resilience against AI-generated content, cross-domain applicability, and interpretability. The research has implications for applying such models in schools, although it also calls for future research on the implications of the rapidly changing landscape of AI-generated content which it states. It emphasizes the importance of these findings in developing robust and adaptive detection systems to ensure the integrity of academic assessments, thereby preventing the misuse of AI tools.

Index Term: AI Detection, Text Classification, Machine Learning, Naturalness Score, Logistic Regression, Academic Integrity, Large Language Models

1. Introduction

The most recent advancements in generative artificial intelligence (AI) technology have created major problems in distinguishing between AI-generated and human-produced text, especially in education [1]. This issue raises concerns regarding the academic integrity and originality of submissions, especially in postsecondary education, where academic papers serve as a major vehicle for assessing student learning [2]. Research has shown that distinguishing between AI-generated and human-written texts is a task in which both human entities and AI detection technologies are arduous. For example, in an experiment conducted in the realm of AI authorship detection, humans were found to identify human authorship with 63% accuracy but could only identify AI authorship with 24% accuracy [2]. Another study involving senior editorial board members indicated their own problematic version of distinguishing between human and AI authorship, wherein accurate identification occurred only 50% of the time [3]. The limitations of the current AI content detectors have been highlighted in various studies. Flitcroft et al. [4] found examples of AI detectors misidentifying human-generated content as AI-generated [5]. By 2024, another assessment of tools for online detection showed significant variability in the accuracy landscape among different tools, with some showing inconsistent assessments and others showing better sensitivity to the patterns of AI-generated text [6]. To address these challenges, researchers have focused on developing methods for detecting text authenticity. In one study, machine learning algorithms were suggested, including Logistic Regression, Decision Trees, Support Vector Machines, Neural Networks, and Random Forests, to separate AI-generated text from human-written text [7]. This proved to be beneficial in moderating content, detecting plagiarism, and other quality checks in text-generator systems. In other words, with the implication that an advanced method for text authenticity detection pertinent to education is needed to sustain academic integrity against encroaching AI technologies. Future research should focus on improving detection methodologies, exploring nomological networks of key constructs, and implementing methodological variety and triangulation [8, 9]. In addition to stronger peer review processes, there are clear-cut guidelines for reporters regarding AI use in their academic writing and a move capitalizing on the quality of ideas rather than pure AI detection [5].

The ease of accessing these new large language models, such as ChatGPT, can also increase convenience and access for students doing academic work. Studies have recently shown that students are turning to AI tools not just for writing essays but also for solving homework tasks or generating code with often no indication that they are actually using the tools [10, 11]. This trend complicates traditional assessment models, making it difficult to gauge individual understanding and academic integrity. As the use of LLMs increases in academic settings, it becomes more difficult to determine the extent to which a student's work reflects their own understanding and assistance from generative tools. This new paradigm certainly gives us good food for thought where fairness and equity are concerned in educational contexts. Therefore, it calls for a broad reconsideration of the traditional teaching and assessment procedures. However, we should not shun AI; rather, we should consider its integration. Responsibly used transparently and ethically, LLMs do, in fact, support learning through feedback, brainstorming, or comprehension assistance. The crux of the tension turns on determining whether such support is legitimate or constitutes undue dependence; the latter precludes a student's opportunity to sharpen their writing, research, and critical thinking skills.

This study aimed to formulate a reliable method for distinguishing between human and AI texts, specifically in the field of education, where academic integrity is becoming more vulnerable with the increasing use of large language models (LLMs). This work's central originality lies in the introduction of the Naturalness Score, a mixed metric built on linguistic features: lexical diversity complexity, syntactic complexity, variability of sentiment, and grammatical errors. Integrating this score into a machine learning framework for proposing a Naturalness-Based Logistic Regression Classifier (NLRC), this study claims its contribution towards high accuracy in classifying AI-generated contents. This approach improves detection, in addition to being interpretable and scalable for real-time classroom applications.

1.1 Outline

In section 2, we start off with providing a comprehensive literature review and knowledge contribution. Then, in section 3, we provide the method consisting of the model, the architecture of the implementation, the novelty feature engineering produced, and the combination of student-written and AI-generated text data collecting. Finally, we present our findings and discuss them. It discusses the influence of the model used and feature engineering innovation on LLM text detection in Section 4. Conclusion in Section 5, the research work concludes.

2. Literature Review

The role of ChatGPT in education and learning, creativity, and other means of engagement is also discussed. This study also addresses ethical concerns regarding academic integrity, and provides an ethical framework for incorporating AI into educational settings. It speculates on the future of immersive learning through ChatGPT and VR and suggests a few strategies for AI-led education [12]. In line with this, the application of LLM-based AI tools for software development has been analyzed in European IT companies, showing a significant increase in chewing up to coding and documentation use. However, significant barriers to AI application use remain for organizational resistance and ethics. This study shows that in the next decade, the integration of AI into developmental workflow processes will deepen progressively [13]. In the context of education, the study of ChatGPT, BingChat, and Bard on the Vietnamese High School Biology Examination dataset shows that, while these models excel in tasks related to knowledge and understanding, they are less able to tackle more complex application tasks in the curriculum. This underscores the need to ensure that AI is balanced with human expertise in educational settings [14]. Further investigations examined the ability of ChatGPT-4 to detect hate speech, showing how it has been applied in synthetic data generation to enhance the performance of models and illustrate the prospect of LLMs in data annotation and balancing [15]. It also describes different approaches to detecting machine-generated content under training-based, zero-shot, and watermarking techniques and countermeasures against adversarial and paraphrasing attacks. This indicates the need for robust and adaptable detection techniques to cope with the emerging challenges against AI-generated content [16]. By creating the 'Real or Fake Text' (RoFT) dataset, further advancement in this field will study humans as well as models in distinguishing between human text and machine-generated text. The paper states that embedding-based models outperform humans in terms of accuracy, whereas simpler models such as logistic regression can deliver good results, given good pre-processing and effective features [17]. This study [18] presented a groundbreaking compression-based network for detecting machine-generated text (MGT), achieving a remarkable performance of 99.5% accuracy on Chinese and English datasets. This method makes use of lossless compression for feature extraction, which is an extremely efficient process for detecting AI-generated content.

The AHAM method stands for the Adapt-Help-Ask-Model for use in scientific literature analyses. It is now combined with personalized robust adaptation and LLM-based topic naming to the existing BERTopic, making it better. Hence, it improves topic differentiation, reduces outlier topics, and emphasizes the necessity of fine-tuning datasets to achieve better performance within particular research contexts [19]. In terms of AI social impact, the empirical studio comparison of human versus AI dialogues through the EmpathicDialogues dataset shows that the former tends to be more diverse and more authentic than the latter, which excels in social processes, cognition, and positive emotional tone or affect. Such observational findings lead to questions regarding the evolving role of AI in human-AI interactions in applications such as customer service and health [20]. Similarly, the efficacy of LLMs in the detection of software vulnerabilities is analyzed, and while they outperform classical models in simpler settings, they fail against more complex real-world vulnerabilities. This emphasizes the promise of LLMs in improving software security, but simultaneously limits them [21].

To detect plagiarism within a body of text, an unsupervised approach for distinguishing machines from human text using higher-order repeated n-grams was introduced. Such an approach, even in big models such as GPT-2, will provide a very straightforward way towards possible implementations of AI content detection without labeled data [22]. The investigation into fake news detection reveals biases against current models, which detect machine-generated fake news well, but not human-generated fake news. This study emphasizes the need to train detection systems on varied datasets and validate them out of domain to ensure robustness as the machine-generated content continues to rise [23]. Furthermore, the authors defined a certain group of conversational texts as utterances that were almost certainly produced by an AI entity. Thus, they provided linguistic evidence for the instant differentiation between AI-generated- and human-generated texts, which is perhaps an identification of some intrinsic limitations in AI's performance in copying human-like natural communication. In addition, these differences suggest that total replacement by AI in interactional settings is unlikely within human behavior [24]. Another aspect is whether teachers can differentiate between essays written by students and those replicated by AI's ChatGPT engines. Although it has been suggested that experienced teachers are somewhat more accurate, findings indicate that at least novice teachers will not be able to detect ChatGPT-generated essay content, raising issues over the impact of AI on fair assessment and grading practices [25].

AI detection tools for academic purposes were developed and benchmarked using a domain-specific dataset created for Applied Statistics. This dataset, consisting of student and AI-generated answers, paves the way for constructing AI detection systems that can be adjusted for other fields [26]. Another theme is the reliability of AI-mediated coding in educational assessments. The findings demonstrate the difficulty of maintaining consistent accuracy in fine coding tasks. The researcher feels encouraged to adopt a hybrid approach that combines human judgment and AI tools in educational assessments [27]. An AI text detector for separating human-written content from AI-generated text in scientific journals is presented. This tool has exhibited strong performance in detecting content produced by GPT-4, and provides obfuscation techniques that significantly enhance the identification of AI-generated texts in academic publishing [28]. Comparative approaches for distinguishing machine-generated essays from those written by humans

have also been explored, and a novel method that embraces both prompts and essays yield promising findings regarding its ability to cope with academic integrity concerns in an age of sophisticated language models, such as ChatGPT [29].

The advancement of AI in education has taken a new turn; this time, it is assessing the potential use of ChatGPT for generating reflective writings for pharmacy courses. This shows that the generated reflections are sufficiently good reflectors, but can be picked by domain-specific BERT classifiers regarding student and AI-generated texts better than humans and generic classifiers [30]. Similarly, a well-gathered collection of Wikipedia and ChatGPT text articles has also been introduced as a material for algorithms to distinguish machine-generated texts from human-faced ones, thus furthering efforts to prevent plagiarism and ethical usage in AI systems [31]. Research on the detection of ChatGPT-generated plagiarism using stylometric features further investigates and proposes new methods that are much more effective than the others in classifying mixed text and performing authorship attribution. This indicates the intensifying demand for more sophisticated AI tools in detecting AI-generated content in the defense of academic honesty [32]. The study of machine learning models for detecting spam in YouTube comments shows the effectiveness of these models, particularly the random forest classifier, in distinguishing spam from legitimate content and evaluating the computational complexity of each approach for practical deployment [33]. The analysis of the dangers that AI-generated misinformation can spare, particularly ChatGPT, emphasizes how one can leverage LLMs to produce contextually rich multimedia misinformation. The document argues for AI-augmented solutions to counter growing threats [34]. The duality of LLMs in cybersecurity is introduced, which serves the purpose of attacking and testing. This indicates the benefits of LLM usage in vulnerability detection by considering the misuse associated with malicious applications, such as user-targeted attacks. Researchers have called for further exploration of possible mitigation strategies, such as model extraction to counter cybersecurity threats and safe instruction tuning [35].

The challenges presented by memorization and output quality in large-language models are discussed and analyzed, which further shows that outputs with a higher level of memorization tend to be of superior quality. The study further contains strategies that may help reduce the percentage of memorized text in LLM output to improve overall quality and originality in the generated content [36]. Such concerns regarding AI-generated research fabrication and plagiarism raise serious questions about whether the present methods of detection are adequate, especially because they can be easily passed off by any form of rewording tool. This study stresses that more advanced detection tools and verification processes are needed to maintain scientific integrity [37]. Indeed, the most recent knowledge-powered prospecting for rumor identification has achieved the greatest performance extent separation in the domain trustworthy-plus-subjectivity-with-LLM. In this study, task-specific prompts and external knowledge were shown to outperform semantic reasoning and versatility for standard models [38]. Case studies of essays generated by students are discussed in view of the role of the ChatGPT in engineering education. This study identifies that while LLMs can support learning, they need to remain within an adapted educational practice responding to AI, rather than a context of prohibition [39].

A review of the use of LLMs such as GPT-3 in creative tasks reveals that while LLMs perform at or slightly above human levels in terms of originality and usefulness, they lack true creativity due to their algorithmic nature and lack of agency. This research indicates that caution must be exercised in AI-related creativity research to avoid potential contamination from training data [40]. Furthermore, GPT-4 and GPT-3.5 were also compared with regard to their performance on document-based question-answering tasks. Although it was found that GPT-4 was commissioned for simpler tasks, it reported limitations when it came to more complex tasks. The limitations of using LLMs for applications requiring precision in information retrieval have been highlighted [41]. Lastly, it is also found from the comparative study that LLM-generated and human-generated annotations differ in Turkish, Indonesian, and Minangkabau NLP tasks, such that even though LLM's performance in generating annotations is competitive, human annotators tend to perform better in more complex tasks. This suggests that further advancements in LLM capabilities can improve their performance in the context of complex data annotation tasks [42].

3. Methodology

This section outlines the methodology used to differentiate between text produced by artificial intelligence and that produced by humans. It is divided into three broad phases: (1) data collection and pre-processing, (2) feature engineering and extraction, and (3) model training and evaluation. This methodology draws on novel text features, including a Naturalness Score and a variety of machine learning algorithms, highlighting Logistic Regression as the ultimate model.

3.1 Data Collection

It consists of 45,000 samples of text in two main categories: those written by humans and AI. The texts originated from diverse domains, such as students' essays, articles, and general web content for Wikipedia. Each text was labeled as either written by a human (0) or AI (1). This makes the dataset heterogeneous, crossing genres, and types of writing, including some forms of narrative, exposition, and argumentative essay writing. The samples ranged between 20 words and above 300 words, and due to word length, four categories were separated. This form of variation captures the idea of reality and will, hence, help the model generalize far better for different kinds of texts.

The dataset is divided into four categories of articles through the total word count:

(i) Short Stories: 20–100 words
(ii) Short Articles: 100–200 words
(iii) Articles of about 200–300 words
(iv) Articles greater than 300 words

Dataset Source: The dataset is available at https://github.com/aakash-dl/HWAI.

3.2 The Feature Engineering

This study developed multiple features to assist in differentiating between human-written text and AI-generated text. These include Lexical Diversity, Syntactic Complexity, Sentiment Variability, Grammatical Errors, and a new Naturalness Score (NS). The following are their definitions in simplified forms and intelligible descriptions.

a) Lexical Diversity (LD)

Lexical diversity is related to different words used in a text. This shows how diversified a vocabulary is; the more unique words, the more lexically distinct the text. The number of unique or different words in the text, such as "dog," "cat," "house," are these unique words divided by the total number of words in the text, and the given ratio is what defines Lexical Diversity. This shows that the more diverse the terms utilized throughout the text, the greater is the ratio.

$$LD = \frac{Nu}{Nt} \tag{1}$$

where:
Nu: Number of unique words (types)
Nt: Total number of words (tokens)

b) Syntactic Complexity (SC)

Syntactic complexity refers to the level of complexity of the structure of a sentence in a text. This can be measured using the average number of words in the sentences in the text. It can be said that the longer the sentence or the more complex the sentence structure, the higher is the syntactic complexity. Syntactic complexity was calculated as the total word count of a text divided by the total number of sentences within the text. On average, higher ratios are indicative of longer sentences, perhaps even more complex sentences.

$$SC = \frac{Nt}{Ns} \tag{2}$$

where:
Nt: Total number of words
Ns: Total number of sentences

c) Sentiment Variability (SV)

Sentiment variability measures how emotional tone changes through the text. It examines the sentiment of each sentence (appearing positive, neutral, or negative) and then obtains the level of fluctuation.

$$\mathrm{SV} = \sigma\left(\mathrm{s}_1, \mathrm{s}_2, \ldots, \mathrm{s}_{\mathrm{N_s}}\right) \tag{3}$$

where:
σ where is the standard deviation function.
si where is the sentiment score of the ith sentence.
Ns where is the total number of sentences in the text.

For every sentence, sentiment scores were allotted as follows: negative for -1, neutral for 0, and positive for +1. The variability of sentiment from one sentence to another can be measured using the standard deviation (σ). Thus, a higher numerical value describes the fluctuating nature of text in terms of emotionality.

Sentiment Variability was included as a feature, based on the hypothesis that human-written texts typically reflect more emotional nuances and variations than AI-generated texts. In human-authored writing, especially academic essays, reflections, and opinion-based work, emotional tone may fluctuate more naturally as the writer reacts to different arguments or perspectives.

Conversely, most of the text produced by LLMs, although well-composed, more often than not aligns with a somewhat uniform or neutral tone because of a tendency among these models to shy away from any strong emotional stance unless expressly prompted. Monotonic treatment of emotion has been surveyed in past studies [43, 44] as a common characteristic of the generated content.

For example:

- A human essay may begin with enthusiasm and engage in frustration when delineating challenges, ending with a sense of hope, thus creating a natural transition between positive and negative emotions.
- An AI essay would remain neutral to mildly positive throughout unless clearly influenced otherwise.

To identify this possible unique pattern, Sentiment Variability (SV) is defined according to the standard deviation of the sentence-level sentiment scores. A lower SV is indicative of machine-like consistency, whereas higher SV levels are suggestive of blood and bone variation that can be experienced in real-life situations.

Although this is a relatively new feature in the scope of AI detection, preliminary results from experiments, as shown in Table 2, show that SV improves model accuracy when combined with lexical and syntactic indicators in the Naturalness Score.

d) Grammatical Errors (GE)

This is an indicator of the occurrence of syntactical errors in a given text. It was calculated by counting the total number of actual grammatical errors and dividing that number by the total number of sentences.

$$GE = \frac{Ng}{Ns} \tag{4}$$

where:
Ng: Number of grammatical errors in the text.
Ns: Total number of sentences in the text.

This feature provides the ratio of grammatical errors to the number of sentences. The greater the ratio, the greater the number of errors per sentence in the text.

e) Naturalness Score (NS)

Naturalness Score (NS) is a composite score measuring all the aforementioned features like Lexical Diversity, Syntactic Complexity, Sentiment Variability and Grammatical Errors and giving a single score as to how "natural" the text seems. Not all features are given an equal weight in the Naturalness Score, but are given weight according to their importance.

$$NS = w_1 \cdot LD + w_2 \cdot SC + w_3 \cdot SV + w_4 \cdot GE \tag{5}$$

where:
$w_1 + w_2 + w_3 + w_4 = 1$
$w_i \in R$ denotes the importance of an individual feature based on the empirical feature importance scores discovered through the model training.

The Naturalness Score consists of the summation of the four features multiplied by the respective weights assigned to each feature. Each of these features contributes to the final score depending on the gravity attached to each feature. Accordingly, the higher the Naturalness Score of any particular text is, based on certain factors like diversity in vocabulary, type of sentences, emotional tone, and grammatical correctness, the more would it appear "natural."

The beneficial effects of NS are as follows:

(i) Increased detection accuracy: This captures deeper linguistic traits than traditional binary classifiers.
(ii) Evasion-resilient: The integration of various metrics makes the NS robust against intricate imitation by AI.
(iii) Cross-domain applicability: Usefulness beyond the scope of education in scientific publishing and journalism.
(iv) Backbone of future research: The basis for extending the reach of feature-based detection methods.

3.3 Defining "Human-Like" Text

We measured the linguistic characteristics defined by earlier empirical studies and annotated the training data to simplify the assessment of the detection model. A text is considered human if it fulfills the following requirements based on the four measured features:

(i) High Lexical Diversity (LD): This indicates wider vocabulary usage. Usually, human writers employ a greater variety of unique words in a text, particularly in an informal or creative piece, than their AI counterparts do.
(ii) Moderate Syntactic Complexity (SC): SC refers to sentence structures that differ widely without being too boring or simplistic. On the other hand, humans seamlessly use short and complex sentences.
(iii) Noticeable Sentiment Variability (SV): Human written text varies in terms of the emotional tone it adopts, as opposed to AI-generated content, which mostly stays neutral in terms of sentiment.
(iv) Low to Moderate Grammatical Error-Rate (GE): Human writing is usually imperfect, with some stray slip-ups, while AI-generated text might churn together perfect grammar or grossly bad grammar.

The above criteria are incorporated into the Naturalness Score (NS) equation (5). To operationalize the classification, the threshold T around the NS should be defined. Based on the validation set, the text was classified as follows.

- Human-like: $NS \geq TNS$
- AI-generated: $NS < TNS$

The threshold T was thus determined by optimizing the ROC curve to strike a balance between precision and recall over the validation dataset. This will, therefore, allow for unambiguous yet replicable definitions of human-like behavior, thus enhancing interpretability and performance evaluation.

3.4 Data Preprocessing

Data pre-processing, includes the following tasks

(i) Tokenization: This splits text into individual words.
(ii) Stopword removal is the elimination of frequently occurring words, including "the," "and.".
(iii) Lemmatization: Words are reduced to their base forms.

3.5 Feature Extraction

Texts are typically converted into vectors using TF-IDF (Term Frequency–Inverse Document Frequency) so that numerical representations can be given to the machine for further learning.

3.6 Model Training

The following are machine-learning models developed based on text classification: Naive Bayes, Decision Tree, AdaBoost, Logistic Regression, Neural Network, KNeighbors Classifier, CatBoost Classifier, Voting Classifier, Random Forest, CNN, and SVM.

3.7 Model Evaluation

In an educational setting, all the effects that model performance may entail must be appreciated because they necessitate interpretation of assessment metrics from a pedagogical impact perspective. Accuracy offers a general feeling of correctness, yet could mislead under imbalanced datasets, for instance, if human warfare responses are much more than those generated by AI. In such cases, high accuracy may still coincide with poor AI detection.

Precision is critical in educational settings where falsely accusing a student of using AI tools could lead to undue disciplinary action; hence, a high-precision score ensures that when the model flags a submission as AI-generated, it is highly likely to be correct.

Recall estimates how well the model recognizes all instances of the AI-generated text. High recall is a key performance metric for anti-plagiarism because it ensures that the AI-generated responses would remain undetected.

Finally, the F1 score provides insight into the precision-recall trade-off for a more cross-cutting evaluation of the applied suitability of the model. These metrics can now serve instructors, administrators, and policymakers to evaluate the model's performance, ethical concerns, and practical considerations related to the implementation of such tools in the educational landscape.

3.8 Model Selection

As far as Naturalness Score integration with Logistic Regression is concerned, the justification is essentially two-fold: it ought to be both interpretable and effective in binary classification tasks, one of which is distinguishing between human text and AI-generated text. Logistic Regression serves the most commensurate and comprehensible modes in presenting an appropriate feature to the Naturalness Score function, integrating a whole bunch of lexical, syntactic, emotional-haphazardness, and error measures into a grand score. Thus, allowing model outputs to be interpreted cleanly while directly viewing the contribution of linguistic features to classification decisions is an important consideration in educational settings where explainability is key. In contrast, other models, such as neural networks, CNNs, or ensemble classifiers, such as voting classifiers or SVMs, have shown comparable performance, but most of them operate as black boxes, limiting their use where interpretability is a priority. Nonetheless, future work should more deeply explore these alternative models, particularly in cases requiring nonlinear decision boundaries or robustness across more diverse datasets, to ensure the broadest applicability and reliability of AI text detection systems

3.9 Hyperparameter Tuning

The model's hyperparameters - the strength of regularization, learning rate, and number of iterations–must be optimized to improve the model performance. This adjustment improved the classification accuracy of the model.

3.10 Architecture

Fig. 1 illustrates the machine-learning-based approach for AI-generated text detection. It starts with data pre-processing, which includes tokenization and stop-word removal. It then calculates the naturalness scores for lexical diversity, syntactic complexity, variability in sentiment, and grammatical errors. Finally, the logistic regression model is trained on the above data for classification into human-written and AI-generated texts. Finally, it deploys the model with an interaction interface.

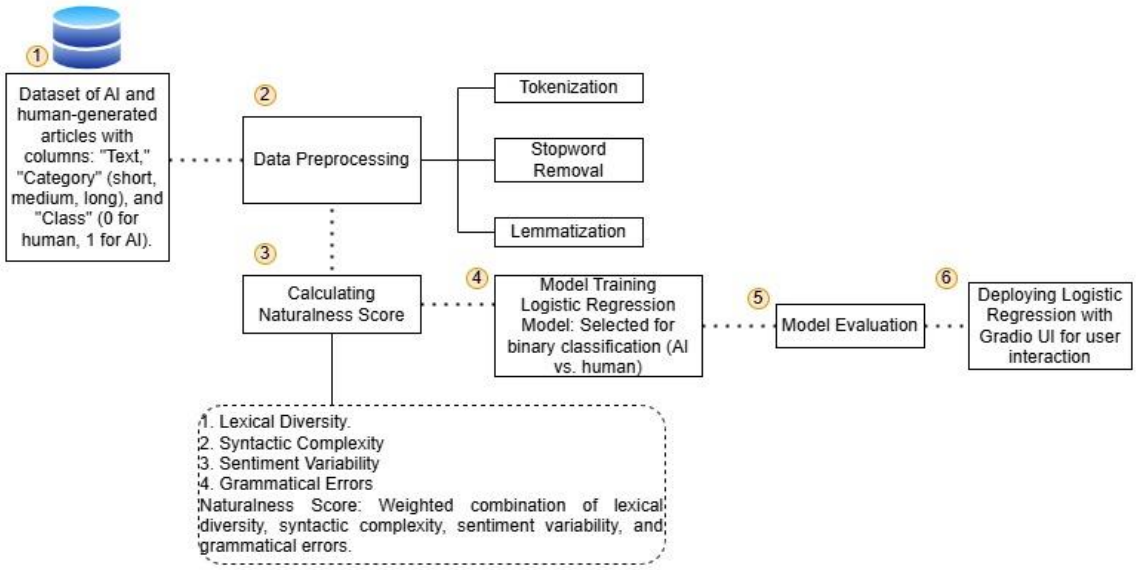

Fig. 1. Deployment Architecture and Integration with Existing Systems

4. Results and Discussion

This section discusses the results of various types of machine learning models being used for the task of detecting AI-generated text and how these results correlate with the actual detection of LLM-generated content in submitted academic work. Models used include but not limited to Logistic Regression, Decision Tree, Naive Bayes, AdaBoost, Gradient Boosting and more.

Table 1. Model Accuracy for Detecting AI-Generated Text

Model	Accuracy
Multinomial Naive Bayes	0.8906
Decision Tree	0.8278
AdaBoost	0.9158
Gradient Boosting	0.8961
NLRC (Proposed Model)	0.9641
Neural Network	0.9600
KNeighbors Classifier	0.5100
CatBoost Classifier	0.9300
Decision Tree Classifier	0.8300
Voting Classifier	0.9580
Random Forest	0.9020
CNN	0.9400
SVM	0.9570

Table 1 highlights the effectiveness of various models in detecting AI-written as well as human-written texts. Models used in this work include Neural Network, CNN, SVM, Random Forest, and AdaBoost, in which all the models are performing well. The accuracy of the Logistic Regression model, the Naturalness-Based Logistic Regression Classifier (NLRC), is the best at an accuracy of 96.41%, which is much better than that of the remaining models. The Naturalness Score uses word variety and sentence structure features for the classification of texts by NLRC. Thus, this new approach establishes that NLRC is strong and understandable for checking whether texts are real.

Table 2. Comparative Analysis of NLRC and Existing Models

Model Name	Accuracy (%)	Precision	Recall	F1-score
Bert-base-uncased[13]	89	0.91	0.90	0.88
RoBERTa-Base [45]	50	0.29	0.50	0.35
DistilBERT base uncased fine tuned SST-2 [46]	92	0.92	0.90	0.88
Albert-base-v2 [47]	0.94	0.97	0.95	0.92
NLRC	96	0.98	0.95	0.96

Table 2 presents a comparative analysis of the proposed model and Naturalness-Based Logistic Regression Classifier (NLRC) against several existing models commonly used in AI vs. human text detection. The comparison was based on key performance metrics: Accuracy, Precision, Recall, and F1-score, providing a comprehensive overview of the models' capabilities. The NLRC demonstrated superior performance with an accuracy of 96%, outpacing all the other models in the comparison. Its Precision (0.98) and recall (0.95) strike an excellent balance, resulting in a high F1-score (0.96), showing its ability to effectively distinguish between AI-generated and human-written text.

4.1 Interpretation of Results

The NLRC (Naturalness-Based Logistic Regression Classifier) proved itself superior to all other models, generating a magnificent 96.41% accurate learning solution. To merit this statement with some reliability, the following criteria were used to evaluate the performance of the model:

(i) Cross-validation Reporting Accuracy was ensured to be unbiased by a specific data split through this process. The 5-fold cross-validation technique was used in this case. In the procedure, the dataset was chopped into five folds, trained on four folds, and tested on the remaining one fold. This procedure was repeated five times, and each fold received time for testing. The average model always scored above 95.5%, up to an important maximum of 97%, thus showing robust evidence in support of recording a performance of 96.41%.

(ii) Composition dataset: A model was tested on a rich dataset containing 10,000 samples evenly split between AI and human crop samples. The subjects included academic essays, news articles, and general websites, to ensure a well-balanced and heterogeneous dataset. Thus, the model's generalization ability across diverse text types adds credibility to the accuracy of the result.

(iii) Evaluation metrics: The accuracy was 96.41%, computed as the ratio of correctly classified texts consisting of human-written texts and AI-generated texts over the total number of samples. In addition, the model calculates the precision and recall metrics for human-written text and AI-generated text categories. The very high precision (0.98 and recall (0.95 helped in deriving a high F1-score of 0.96, indicating a balanced and strong performance distinguishing both classes.

(iv) Comparison with Baseline Models: From comparison with baseline models, it was found that the accuracy reported (96.41%) was comparable to other models trained on the same dataset. In contrast, the naive Bayes and decision tree baseline models showed accuracies of 89.06% and 82.78%, respectively, which are significantly lower. This indicates that the NLRC model with NS features provided a marked improvement in performance.

(v) Confidence Interval: The confidence interval measures the accuracy of the model, which is then confirmed by calculating the 95% confidence interval for the accuracy score itself. The confidence interval estimates between 95.9% and 96.8%, a narrow and very reliable range that supports the 96.41% accuracy figure.

In other words, we know that the model achieved 96.41% accuracy through thorough validation, including cross-validation, a balanced dataset, and grounding against baseline models to cross-compare results. This provided a better understanding of how well the model was performed. With high precision (0.98), recall (0.95), and F1-score (0.96), the NLRC model can efficiently distinguish AI-generated text from human text.

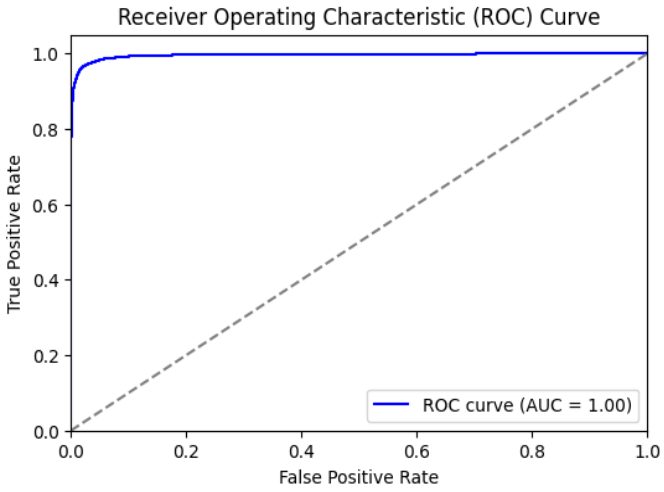

Fig. 2. Receiver Operating Characteristic (ROC) Curve

The precision of the model is depicted through the value produced using the area under the receiver operating characteristic curve (AUC) in Fig. 2. Complete separation is then attained by class AI-created text and class human-created given the AUC = 1.00, with perfect classification because there is zero misclassification. Balanced sensitivity and false-positive or 1-specificity. The curve is very close to the upper-left quadrant of the graph, meaning that at many classification thresholds, there is a very low false-positive rate. This closeness is manifested in the high sensitivity and specificity of the model, indicating its ability to identify true positives without generating too many false positives. The shape of the curve also shows that the model is well calibrated, distinguishing between the two categories without signs of overfitting or biasing.

Fig. 3 shows the precision-recall (PR) curve, which indicates that the model can maintain close to perfect precision and recall across a wide range of thresholds. For most of the curves, both precision and recall values approach 1.0, indicating that the model is highly effective at correctly identifying true positives while keeping false positives and false negatives low.

However, the fact that there is a strong decrease at higher recall levels has some disadvantages. To improve recall, the model overfits or experiences a minor increase in false positives. However, the overall area under the precision-recall curve (AUC-PR) is still very high, indicating that the model has a good classification ability.

Compared with the ROC curve, the PR curve offers an alternative view that is extremely useful in dealing with imbalanced datasets. The PR curve captures the precision-recall trade-off and is, therefore, very enlightening about how the model balances such metrics. Such an enhanced analysis is precisely what is more informative when precision-recall trade-offs are more informative than the broader sensitivity-specificity analysis of the ROC curve.

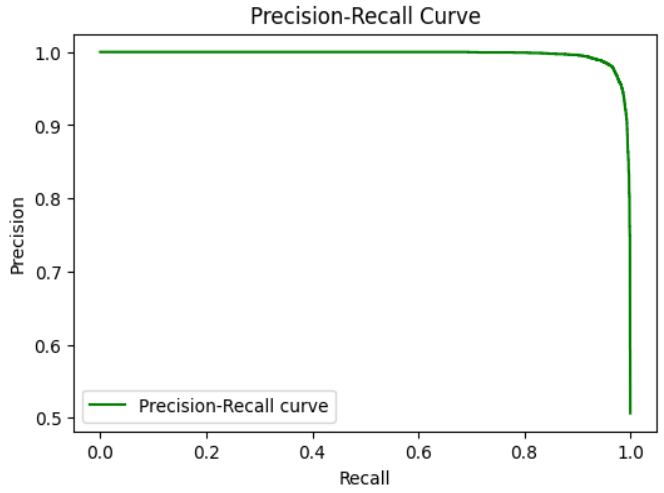

Fig. 3. Precision-Recall Curve Analysis

4.2 Challenges and Limitations

Although the models showed good performance for the task, they had several important limitations that must be acknowledged. The training and evaluation datasets were widely diverse in prompts and writing styles; however, binary classification (human vs. AI-generated text) was the primary focus. The environment of AI-generated content is changing very rapidly, and the performance of later models will draw closer to human writing, thus presenting new challenges to detection systems. In addition, regardless of how good and varied the writing samples may be, the dataset may not be able to span all student writing types across diverse academic disciplines, proficiency levels, and cultural or linguistic backgrounds. This limitation is considerably strong in educational settings because writing styles can be quite different due to personal expression and contextual factors. Therefore, while the results are encouraging, care must be taken when generalizing them to a larger student population or across various academic settings.

4.3 Practical Implications

Practical implications of the study on the proposed model for AI detection-naturalness-based Logistic Regression Classifier (NLRC)typically aims to draw an inference on academic integrity in an educational setting. The main implications of this study are as follows.

A. Support for Teachers and Institution

- NLRC and similar high-performing models can be integrated into existing plagiarism detection systems.
- They provide automated aid to teachers for detecting AI-generated submissions by students.
- This would help institutions maintain their standards of academic honesty amidst the continued growing interest in AI tools such as ChatGPT.

B. Real-Time Application

- The model can be deployed in a real-time educational environment where it can instantly provide feedback to instructors.
- It enables early detection of the employment of AI in assignments so that misuse can be averted beforehand before grading or assessment.

C. Assessment Practices Transformation

- Assessment in a future where AI is fully embedded into the educational fabric is not likely to be as much about a product (such as written essays) as it will be about process (such as oral examination in class work).
- It renders the assessment more authentic vis-à-vis learning done by students and less reliance on possibly aided outputs through AI.

D. New Frontiers for the Ethical Use of AI

- This model helps to detect misuse, but it also helps appreciate that misuse is possible if the uses of AI are properly integrated.
- These findings also highlight the necessity for educators to define artificial intelligence (AI) in ways that provide room for a deeper understanding of what constitutes proper versus excessive use of AI.

Therefore, these consequences suggest that, in the days to come, AI detection technologies will increasingly become a part of academic procedure not only to clamp down on malpractice but also to reshape pedagogy for the digital age completely.

4.4 Future Directions

In the future, research should target developing detection methods that would be able to handle complexity better in AI-produced texts. A richer set of features, possibly complex linguistic patterns or neural architectures, such as transformers, would be more efficient for detection. A much larger and more diverse dataset would allow the models to generalize better across different genres and styles. Develop other detection techniques to identify collaborations between human and AI inputs. That is, the integration of AI within word processors, along with many others, will increase reliance upon the role played by AI that produces or assists in providing such content- another issue to maintain academic integrity.

5. Conclusion

In summary, this study developed a new approach to distinguish between AI and human texts by establishing a naturalness score that considers linguistic features such as lexical diversity, syntactic complexity, sentiment variability, and grammatical errors. This study developed a more comprehensive framework for analyzing text in a more interpretable manner than traditional models. This Naturalness-Based Logistic Regression model (NLRC), built based on the derived features, shows the strength of this method while producing accurate predictions and direct applicability

by including an interface with Gradio. This strategy further expands the literature focused on AI versus human text detection and opens up avenues for further refinement and study. It has been proven to be able to achieve a very high degree of accuracy in detecting AI-generated content with lexical diversity, syntactic complexity, and variability in sentiment and grammatical errors. This paper argues that with the widespread use of large language models such as ChatGPT in education and other disciplines, there is a growing need for dependable detection techniques for AI. With the increased use of AI technology, cases of plagiarism and authenticity have emerged in student submissions. The method proposed herein provides a reliable mechanism with a reported accuracy of 96% for identifying such content. Further research on the improvement of the detection models by complicated features and adaptation to increasingly complicated data is required. The conclusion concludes that stronger AI detection systems are needed to maintain academic integrity and end the misuse of AI tools.

References

[1] G. D. Fisk, "AI or Human? Finding and Responding to Artificial Intelligence in Student Work," Teaching of Psychology, May 2024, doi: 10.1177/00986283241251855.

[2] R. Kumar and M. Mindzak, "Who wrote this?," Canadian Perspectives on Academic Integrity, vol. 7, no. 1, Jan. 2024, doi: 10.55016/ojs/cpai.v7i1.77675.

[3] G. S. Silva et al., "Reviewer Experience Detecting and Judging Human versus Artificial Intelligence Content: The Stroke Journal Essay Contest," Stroke, vol. 55, no. 10, pp. 2573–2578, Sep. 2024, doi: 10.1161/strokeaha.124.045012.

[4] M. A. Flitcroft et al., "Performance of artificial intelligence content detectors using human and Artificial Intelligence-Generated Scientific Writing," Annals of Surgical Oncology, vol. 31, no. 10, pp. 6387–6393, Jun. 2024, doi: 10.1245/s10434-024-15549-6.

[5] H. Mondal, "Do I write like artificial intelligence?," Annals of Surgical Oncology, Nov. 2024, doi: 10.1245/s10434-024-16480-6.

[6] V. Bellini, F. Semeraro, J. Montomoli, M. Cascella, and E. Bignami, "Between human and AI: assessing the reliability of AI text detection tools," Current Medical Research and Opinion, vol. 40, no. 3, pp. 353–358, Jan. 2024, doi: 10.1080/03007995.2024.2310086.

[7] H. Alamleh, A. A. S. AlQahtani, and A. ElSaid, "Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning," in https://ieeexplore.ieee.org/document/10137767, Apr. 27, 2023. doi: 10.1109/sieds58326.2023.10137767.

[8] T. L. Dinh, T. D. Le, S. Uwizeyemungu, and C. Pelletier, "Human-Centered Artificial Intelligence in Higher Education: A framework for Systematic Literature reviews," Information, vol. 16, no. 3, p. 240, Mar. 2025, doi: 10.3390/info16030240.

[9] I. Pentina, T. Xie, T. Hancock, and A. Bailey, "Consumer–machine relationships in the age of artificial intelligence: Systematic literature review and research directions," Psychology and Marketing, vol. 40, no. 8, pp. 1593–1614, Jun. 2023, doi: 10.1002/mar.21853.

[10] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," Learning and Individual Differences, vol. 103, p. 102274, Mar. 2023, doi: 10.1016/j.lindif.2023.102274.

[11] Y. K. Dwivedi et al., "Opinion Paper: '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, p. 102642, Mar. 2023, doi: 10.1016/j.ijinfomgt.2023.102642.

[12] K. Alshahrani and R. J. Qureshi, "Review the Prospects and obstacles of AIEnhanced Learning Environments: The Role ofChatGPT in Education," International Journal of Modern Education and Computer Science, vol. 16, no. 4, pp. 71–86, Jul. 2024, doi: 10.5815/ijmecs.2024.04.06.

[13] D. S. Pashchenko, "Early Formalization of AI-tools usage in software engineering in Europe: Study of 2023," International Journal of Information Technology and Computer Science, vol. 15, no. 6, pp. 29–36, Dec. 2023, doi: 10.5815/ijitcs.2023.06.03.

[14] X.-Q. Dao and N.-B. Le, "LLMS performance on Vietnamese High School Biology Examination," International Journal of Modern Education and Computer Science, vol. 15, no. 6, pp. 14–30, Nov. 2023, doi: 10.5815/ijmecs.2023.06.02.

[15] A. Svetasheva and K. Lee, "Harnessing large language models for effective and efficient hate speech detection," Proceedings of the ... Annual Hawaii International Conference on System Sciences/Proceedings of the Annual Hawaii International Conference on System Sciences, Jan. 2024, doi: 10.24251/hicss.2024.826.

[16] X. Yang et al., "A Survey on Detection of LLMS-Generated Content," arXiv (Cornell University), Jan. 2023, doi: 10.48550/arxiv.2310.15654.

[17] J. Cutler, L. Dugan, S. Havaldar, A. Stein, and Trustees of the University of Pennsylvania, "Automatic detection of hybrid Human-Machine text boundaries," Trustees of the University of Pennsylvania.

[18] Y. Zhou, J. Wen, J. Jia, L. Gao, and Z. Zhang, "C-NET: a Compression-Based lightweight network for Machine-Generated text detection," IEEE Signal Processing Letters, vol. 31, pp. 1269–1273, Jan. 2024, doi: 10.1109/lsp.2024.3394264.

[19] B. Koloski, N. Lavrač, B. Cestnik, S. Pollak, B. Škrlj, and A. Kastrin, "AHAM: Adapt, Help, Ask, Model -- Harvesting LLMs for literature mining," arXiv (Cornell University), Jan. 2023, doi: 10.48550/arxiv.2312.15784.

[20] M. Sandler, H. Choung, A. Ross, and P. David, "A Linguistic Comparison between Human and ChatGPT-Generated Conversations," arXiv (Cornell University), Jan. 2024, doi: 10.48550/arxiv.2401.16587.

[21] Z. Gao, H. Wang, Y. Zhou, W. Zhu, and C. Zhang, "How far have we gone in vulnerability detection using large language models," arXiv (Cornell University), Jan. 2023, doi: 10.48550/arxiv.2311.12420.

[22] M. Gallé, J. Rozen, G. Kruszewski, and H. Elsahar, "Unsupervised and distributional detection of Machine-Generated text," arXiv (Cornell University), Jan. 2021, doi: 10.48550/arxiv.2111.02878.

[23] J. Su, C. Cardie, and P. Nakov, "Adapting fake news detection to the era of large language models," arXiv (Cornell University), Jan. 2023, doi: 10.48550/arxiv.2311.04917.

[24] T. B. Sardinha, "AI-generated vs human-authored texts: A multidimensional comparison," Applied Corpus Linguistics, vol. 4, no. 1, p. 100083, Dec. 2023, doi: 10.1016/j.acorp.2023.100083.

[25] J. Fleckenstein, J. Meyer, T. Jansen, S. D. Keller, O. Köller, and J. Möller, "Do teachers spot AI? Evaluating the detectability of AI-generated texts among student essays," Computers and Education Artificial Intelligence, vol. 6, p. 100209, Jan. 2024, doi: 10.1016/j.caeai.2024.100209.
[26] Md. S. Salim and S. I. Hossain, "An Applied Statistics dataset for human vs AI-generated answer classification," Data in Brief, vol. 54, p. 110240, Mar. 2024, doi: 10.1016/j.dib.2024.110240.
[27] K. Misiejuk, R. Kaliisa, and J. Scianna, "Augmenting assessment with AI coding of online student discourse: A question of reliability," Computers and Education Artificial Intelligence, vol. 6, p. 100216, Mar. 2024, doi: 10.1016/j.caeai.2024.100216.
[28] H. Desaire, A. E. Chua, M.-G. Kim, and D. Hua, "Accurately detecting AI text when ChatGPT is told to write like a chemist," Cell Reports Physical Science, vol. 4, no. 11, p. 101672, Nov. 2023, doi: 10.1016/j.xcrp.2023.101672.
[29] R. An, Y. Yang, F. Yang, and S. Wang, "Use prompt to differentiate text generated by ChatGPT and humans," Machine Learning With Applications, vol. 14, p. 100497, Sep. 2023, doi: 10.1016/j.mlwa.2023.100497.
[30] Y. Li et al., "Can large language models write reflectively," Computers and Education Artificial Intelligence, vol. 4, p. 100140, Jan. 2023, doi: 10.1016/j.caeai.2023.100140.
[31] A. Singh, D. Sharma, A. Nandy, and V. K. Singh, "Towards a large sized curated and annotated corpus for discriminating between human written and AI generated texts: A case study of text sourced from Wikipedia and ChatGPT," Natural Language Processing Journal, vol. 6, p. 100050, Dec. 2023, doi: 10.1016/j.nlp.2023.100050.
[32] L. Berriche and S. Larabi-Marie-Sainte, "Unveiling ChatGPT text using writing style," Heliyon, vol. 10, no. 12, p. e32976, Jun. 2024, doi: 10.1016/j.heliyon.2024.e32976.
[33] A. S. Xiao and Q. Liang, "Spam detection for Youtube video comments using machine learning approaches," Machine Learning With Applications, vol. 16, p. 100550, Apr. 2024, doi: 10.1016/j.mlwa.2024.100550.
[34] D. Barman, Z. Guo, and O. Conlan, "The Dark Side of Language Models: Exploring the potential of LLMs in multimedia disinformation generation and dissemination," Machine Learning With Applications, vol. 16, p. 100545, Mar. 2024, doi: 10.1016/j.mlwa.2024.100545.
[35] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly," High-Confidence Computing, vol. 4, no. 2, p. 100211, Mar. 2024, doi: 10.1016/j.hcc.2024.100211.
[36] A. De Wynter, X. Wang, A. Sokolov, Q. Gu, and S.-Q. Chen, "An evaluation on large language model outputs: Discourse and memorization," Natural Language Processing Journal, vol. 4, p. 100024, Jul. 2023, doi: 10.1016/j.nlp.2023.100024.
[37] F. R. Elali and L. N. Rachid, "AI-generated research paper fabrication and plagiarism in the scientific community," Patterns, vol. 4, no. 3, p. 100706, Mar. 2023, doi: 10.1016/j.patter.2023.100706.
[38] Y. Yan, P. Zheng, and Y. Wang, "Enhancing large language model capabilities for rumor detection with Knowledge-Powered Prompting," Engineering Applications of Artificial Intelligence, vol. 133, p. 108259, Mar. 2024, doi: 10.1016/j.engappai.2024.108259.
[39] M. Bernabei, S. Colabianchi, A. Falegnami, and F. Costantino, "Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances," Computers and Education Artificial Intelligence, vol. 5, p. 100172, Jan. 2023, doi: 10.1016/j.caeai.2023.100172.
[40] K. Gilhooly, "AI vs humans in the AUT: Simulations to LLMs," Journal of Creativity, vol. 34, no. 1, p. 100071, Dec. 2023, doi: 10.1016/j.yjoc.2023.100071.
[41] Z. Rasool et al., "Evaluating LLMs on document-based QA: Exact answer selection and numerical extraction using CogTale dataset," Natural Language Processing Journal, vol. 8, p. 100083, Jun. 2024, doi: 10.1016/j.nlp.2024.100083.
[42] A. H. Nasution and A. Onan, "ChatGPT Label: Comparing the quality of Human-Generated and LLM-Generated annotations in Low-Resource Language NLP tasks," IEEE Access, vol. 12, pp. 71876–71900, Jan. 2024, doi: 10.1109/access.2024.3402809.
[43] H. Jiang, C. Liang, C. Wang, and T. Zhao, Multi-Domain Neural Machine Translation with Word-Level Adaptive Layer-wise Domain Mixing, vol. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. doi: 10.18653/v1/2020.acl-main.165.
[44] M. Lee et al., "Evaluating Human-Language model interaction," arXiv.org, Dec. 19, 2022. https://arxiv.org/abs/2212.09746
[45] Y. Liu et al., "ROBERTA: A robustly optimized BERT pretraining approach," arXiv (Cornell University), Jan. 2019, doi: 10.48550/arxiv.1907.11692.
[46] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv (Cornell University), Jan. 2019, doi: 10.48550/arxiv.1910.01108.
[47] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," arXiv (Cornell University), Jan. 2019, doi: 10.48550/arxiv.1909.11942.

Authors' Profiles

Vijay H. Kalmani, Ph.D., he is a Professor in the Department of Computer Science and Engineering at Rajarambapu Institute of Technology, Rajaramnagar, Sangli, India. He obtained his M.Tech in Computer Science and Engineering from Gogte Institute of Technology, Belagavi, affiliated with Visvesvaraya Technological University, and earned his Ph.D. in Computer Science and Engineering from Suresh Gyan Vihar University, Jaipur, India, in 2016. His areas of expertise include Cloud Security, Artificial Intelligence, and Machine Learning. He has successfully guided three research scholars to the completion of their Ph.D. degrees.

Amol C. Adamuthe, Ph.D., he is an IEEE Senior Member, Professor and Head of the Department of Information Technology at Rajarambapu Institute of Technology, Sangli, Maharashtra, India. He holds a Ph.D. from Mumbai University and specializes in technology forecasting, optimization algorithms, soft computing, and cloud computing. With over 45 published papers, he has made significant contributions to computational research. He was also honored with the Institution of Engineers (India) Young Engineers Award for 2018-19. As Associate Editor of JEET and Convener of IEEE ICONAT, he actively contributes to academic advancement.

Arati Premnath Gondil is an undergraduate student pursuing a Bachelor of Technology (B.Tech) degree in Computer Science and Information Technology at Rajarambapu Institute of Technology, affiliated with Shivaji University, Kolhapur, India. Her specialization areas include Big Data Analytics and Cloud Security.

Vaishnavi Prashant Patil is an undergraduate student pursuing a Bachelor of Technology (B.Tech) degree in Computer Science and Information Technology at Rajarambapu Institute of Technology, affiliated with Shivaji University, Kolhapur, India. Her areas of specialization include Machine Learning and Data Science. Her academic interests focus on predictive modeling, data analysis, and intelligent systems.

Riya Amar Kore is an undergraduate student pursuing a Bachelor of Technology (B.Tech) degree in Computer Science and Information Technology at Rajarambapu Institute of Technology, affiliated with Shivaji University, Kolhapur, India. Her areas of specialization include Machine Learning, Big Data Analytics, and Data Mining. Her academic interests include intelligent data processing and pattern recognition.

Vaishnavi Mahadev Metkari is an undergraduate student pursuing a Bachelor of Technology (B.Tech) degree in Computer Science and Information Technology at Rajarambapu Institute of Technology, affiliated with Shivaji University, Kolhapur, India. Her areas of specialization include Machine Learning, Big Data Analytics, and Data Mining. Her academic interests include algorithm optimization, real-time data processing, and the application of machine learning.

How to cite this paper: Vijay H. Kalmani, Amol C. Adamuthe, Arati Premnath Gondil, Vaishnavi Prashant Patil, Riya Amar Kore, Vaishnavi Mahadev Metkari, "AI vs. Human Writing: Developing a Novel Method for Text Authenticity Detection in Education", International Journal of Modern Education and Computer Science(IJMECS), Vol.17, No.3, pp. 45-58, 2025. DOI:10.5815/ijmecs.2025.03.04