International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.699

Detecting AI-Generated Essays with Lexical Stylometric Features using Machine Learning

[1]Tahmida Haque Sumbula, [2]Suriani Sulaiman*, [3]Adlina Ariffin
[1,2] Department of Computer Science, Kulliyyah of ICT, International Islamic University Malaysia, Gombak, Malaysia.
[3] Department of English Language and Literature, Kulliyyah of IRKHS, International Islamic University Malaysia, Gombak, Malaysia.

Corresponding Author: ssuriani@iium.edu.my
(Received: 30th March 2026; Accepted: 7th May, 2026; Published on-line: 30th July, 2026)

Abstract- The increasing use of generative artificial intelligence (AI) in academic writing has intensified concerns about authorship authenticity. This study investigates whether lexical stylometric features alone can distinguish AI-generated essays from human-written essays. A balanced dataset of 200 undergraduate essays (100 AI-generated and 100 human-written) was analysed using vocabulary richness measures, lexical density, word usage distributions, and word-level n-grams. Logistic Regression, Decision Tree, and Support Vector Machine classifiers were applied to evaluate performance. The results indicated that lexical features provide strong discriminatory capability across the three models with accuracy ranging between 85 – 90% for AI-generated essays and 90-100% for human-written essays, suggesting that effective AI authorship detection can be achieved without complex feature engineering or advanced stylometric features. However, the findings should be interpreted as evidence for a lightweight and interpretable baseline rather than a fully generalizable detector, since robustness against unseen prompts, unseen AI models, paraphrasing, and broader writing domains were not tested. The study contributes to an interpretable lexical-feature framework for AI authorship detection in educational contexts and identifies the need for future validation using different categories of stylometric features, as well as larger, multi-modal, and cross-domain datasets.

Keywords- machine learning, lexical features, stylometric, AI-generated essays, academic writing, ChatGPT

I. Introduction

Generative AI systems have significantly reshaped academic writing, creating new challenges for maintaining authorship integrity in higher education. As AI-generated essays become increasingly sophisticated, institutions require reliable and practical detection approaches.

Many existing detection studies rely on deep learning models and complex feature sets, which may limit interpretability and scalability. Stylometric research, however, suggests that lexical patterns such as vocabulary diversity and word usage tendencies capture meaningful stylistic signals. Yet, the effectiveness of purely lexical features for AI text detection remains insufficiently examined.

This study does not claim that lexical features are sufficient for all AI-detection settings. Instead, it evaluates whether a restricted lexical feature set can provide a transparent and computationally lightweight baseline for distinguishing AI-generated and human-written undergraduate essays under controlled conditions.

II. Related Work

Recent work on AI-generated text detection has explored a range of methods from statistical models to deep learning, highlighting both opportunities and challenges in distinguishing machine-generated content from human writing. A systematic review by [1] surveyed existing approaches, datasets, and evaluation metrics for AI text detection, noting that feature-based and machine learning methods remain important alongside neural models. Another study by [2] performed linguistics analyses as well as examined the characteristics, differences and gaps between AI-generated and human-written contents by comparing various responses collected from both human experts and ChatGPT. The authors then investigated the factors influencing detection systems using machine learning and deep learning models.

A total of eleven (11) machine learning and deep learning models were assessed by [3] using a combined collection of 500 essays and programming assignments by Computer Science students to measure the effectiveness of each model in detecting and distinguishing between AI-generated and human-written texts. Their results showed that Random Forest (RF) performed best on essays (93%), followed by Logistic Regression (LR) and Support Vector Machine (SVM) with accuracies of 92.5% and 91.5% respectively.

Krawczyk et al. [4] investigated AI-generated essay classification using numerical text representations such as

TF-IDF, demonstrating that LR and similar classifiers achieved competitive results on essay datasets when lexical and frequency features were employed. Their findings underscored the value of traditional classifiers for interpretable text classification.

Masih et al. [5] compared traditional machine learning models, including SVM and LR, with transformer-based architectures for differentiating human and AI texts, confirming that classical models remained strong baselines and contributed to explainable predictions in AI detection tasks.

Several studies have also focused on stylometric and linguistic feature-based detection approaches [2], [6] – [13]. Agarwal and Sharma [6] examined the effectiveness of stylometric features such as lexical richness, sentence structure, and word usage patterns for authorship discrimination using machine learning algorithms. Their findings showed that stylometric indicators can successfully capture writing style differences and support accurate classification.

Similarly, [7] proposed an ensemble learning framework for authorship identification, combining multiple classifiers and stylometric features to improve classification robustness. Their results demonstrated that ensemble methods can enhance detection performance by leveraging complementary strengths of different models.

Recent research has also explored the integration of linguistic features with pretrained models. Yadagiri et al. [8] investigated the detection of AI generated text using linguistic features alongside pretrained models, highlighting that linguistic characteristics such as syntax and stylistic patterns remain important signals for distinguishing machine generated content.

Desaire et al. [9] selected 20 linguistic features that includes paragraph complexity, punctuation mark, diversity in sentence length and popular words or numbers to prove that supervised classification methods such as Naïve Bayes and XGBoost can successfully discriminate text generated by ChatGPT from human scientists in academic writing for up to 99% accuracy.

In addition, [10] conducted a comparative study of machine learning algorithms for AI generated text detection, reporting that Logistic Regression and Support Vector Machine achieved strong performance due to their ability to effectively separate lexical feature spaces.

While a few studies focused on detecting AI-generated contents using machine learning and deep learning techniques only in academic writing due to ethical concerns [4], [9], [12], [14], another emerging direction involves improving model interpretability and explainability in authorship analysis. Ramnath et al. [11] introduced the CAVE framework, which provided controllable explanations for authorship verification models, enabling better understanding of the linguistic evidence used by classification systems.

Prior research showed three major directions in AI-generated text detection. First, review-based studies have emphasized that AI-text detection remained methodologically difficult because performance varied across datasets, writing domains, generation models, and evaluation settings [1], [14], [15]. Second, transformer-based and deep learning approaches have demonstrated strong performance by capturing contextual and semantic patterns, but these models were often less transparent and may require greater computational resources [5], [8]. Third, stylometric and classical machine learning approaches have shown that interpretable linguistic features, such as lexical richness, syntactic patterns, punctuation behaviour, and word-frequency distributions, can provide useful authorship signals [6], [9], [10]. For example, Desaire et al. [9] showed that a small set of carefully selected linguistic features can produce high classification performance in a targeted academic-writing setting. However, such findings are usually domain-specific and require cautious interpretation because high in-domain accuracy may reflect dataset regularities rather than robust generalization. Therefore, the gap addressed in this study is not whether lexical features outperform semantic or transformer-based methods, but whether purely lexical stylometric features can serve as an interpretable baseline for AI versus human essay classification in a controlled undergraduate writing context.

III. Methodology

This research implemented an experimental workflow to evaluate whether lexical stylometric features can effectively distinguish AI-generated essays from human-written essays using interpretable machine learning classifiers. The methodology consisted of three major phases; data collection, feature extraction, and model training and evaluation.

A. Data Collection

The data collection process was conducted in a controlled academic setting to ensure consistency between human-written and AI-generated essays. A total of 200 undergraduate essays were collected, consisting of 100 human-written essays and 100 AI-generated essays. The dataset size was selected as a balanced pilot corpus suitable for controlled feasibility testing, not for population-level generalization. Prior AI-detection studies have also used controlled document sets to evaluate detection behaviour, while emphasizing that limited datasets require cautious interpretation and transparent reporting, [15], [16]. Therefore, the present dataset is treated as an exploratory

benchmark to ensure balanced class distribution and variation in writing styles and topics, enabling robust authorship discrimination analysis.

Human-written essays were composed offline under monitored conditions to reduce the possibility of external assistance or online tool usage. Participants were undergraduate students from the ICT and English departments, allowing variation in writing background while maintaining a comparable academic level. AI-generated essays were produced using GPT-4 with the same essay question topic to maintain topic consistency, which is:

> *"ChatGPT is an innovation that has changed the education landscape. Some people perceive this new technology positively while others take it unfavourably. What are your opinions on the use of ChatGPT in academic writing?"*

Each essay contained approximately 500 words and was saved in plain text (.txt) format for uniform preprocessing.

B. Feature Extraction

Feature extraction was performed to transform raw essays into structured numerical representations suitable for machine learning. The study focused on lexical stylometric features, including vocabulary richness, lexical density, word usage distributions, part-of-speech frequencies, function-word and content-word frequencies, and word-level n-grams. Although POS frequencies provide syntactic distributional information, the feature set remains primarily lexical and surface-oriented. Semantic coherence, discourse structure, and contextual embeddings were intentionally excluded to preserve interpretability and isolate the contribution of lexical stylometry. Lexical richness was quantified using the Type-Token Ratio (TTR), defined as:

$$TTR = \frac{V}{N}$$

where V represents the number of unique words (types) and Nrepresents the total number of words (tokens) in the essay. The Hapax Legomenon Rate is calculated as,

$$H = \frac{V_1}{N}$$

where V_1denotes the number of words occurring only once in the text.

Word-level n-grams were extracted to capture recurring local word patterns and short phrase structures. TF-IDF representations were computed to measure the relative importance of terms within individual essays compared with the full corpus. To address feature-selection arbitrariness, the top n-gram features were not selected manually. Instead, unigrams and bigrams were ranked using a supervised chi-square feature-selection procedure on the training data only. The value k = 20 was selected to preserve interpretability, reduce sparsity in a 200-document dataset, and maintain a compact feature set consistent with prior interpretable AI-detection work using small linguistic feature sets [9][16]. To reduce the risk of data leakage, feature selection was performed within the training pipeline before evaluation was done on the validation set. TF-IDF is defined as,

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

where:

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}}$$

and

$$DF(t) = \log\left(\frac{D}{df_t}\right)$$

Here, $f_{t,d}$ denotes the frequency of term t in document d, in which D is the total number of documents, and df_t is the number of documents containing term t.

The top 20 unigrams and bigrams ranked by TF-IDF scores were selected to represent salient lexical patterns. All extracted features were standardized before being used for classification task as shown in Table I.

TABLE I

EXTRACTED FEATURES FOR AUTHORSHIP DISCRIMINATION

Features	**Description**
Word count, Type-Token Ratio (TTR), Hapax Legomenon, Avg. Sentence Length	Word-level statistics including vocabulary richness and sentence length
POS Tags	Distribution of part-of-speech tags
Word n-grams	Frequent bigrams and trigrams
Function and Content Words	Frequency of function vs. content words

C. Machine Learning Models

Three (3) machine learning classifiers; Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM) were used to evaluate AI authorship detection performance based solely on lexical and n-gram features. These models were selected due to their

https://doi.org/10.31436/ijpcc.v12i2.699

interpretability, robustness, and suitability for structured numerical feature spaces.

1) Logistic Regression (LR): Logistic Regression is a supervised linear classification algorithm that models the probability of a binary outcome using the logistic (sigmoid) function. Given a feature vector x, LR estimates the probability that an essay belongs to the AI-generated class as:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

where w represents the weight vector and b the bias term. The model learned optimal parameters by minimizing the log-loss (cross-entropy) function. Logistic Regression was used as a baseline model to assess the linear separability of lexical features and evaluated their discriminative power in distinguishing human-written from AI-generated essays.

2) Decision Tree (DT): Decision Tree is a non-linear classification model that recursively partitions the feature space into subsets based on decision rules. At each node, the model selects a feature that maximizes class separation using impurity measures such as Gini Index,

$$Gini = 1 - \sum_{i=1}^{C} p_i^2$$

where p_i is the proportion of samples belonging to class i. The tree structure provides an interpretable hierarchical representation of decision boundaries and feature importance. In this study, DT was implemented to capture potential non-linear relationships within lexical and TF-IDF features.

3) Support Vector Machine (SVM): Support Vector Machine is a margin-based classifier that seeks to find the optimal hyperplane separating two classes while maximizing the margin between them. For linearly separable data, SVM solves the optimization problem,

$$\min_{w,b} \frac{1}{2} \| w \|^2$$

subject to:

$$y_i(w^T x_i + b) \geq 1$$

where x_i represents feature vectors and $y_i \in \{-1,1\}$ are class labels. In this study, a linear kernel was applied, as the lexical and TF-IDF features demonstrated strong linear separability. SVM allows robust classification, particularly in high-dimensional feature spaces.

Each model was trained on 80% of the dataset and tested on the remaining 20%, with hyperparameters tuned using 5-fold cross-validation to optimize accuracy. The performance of each classifier was evaluated using accuracy, precision, recall, and F1-score, providing a clear comparison of their effectiveness on lexical-only features. Table II summarizes the model training procedure including data splitting strategy applied to the machine learning models while Fig. 1 illustrates the proposed methodology for the detection of AI versus human authorship.

TABLE II

SUMMARY OF EXPERIMENTAL PROCEDURE

Step	Procedure Step	Description
1	Dataset Splitting	Dataset divided into training (80%) and testing (20%) sets with stratified sampling.
2	Feature Extraction	Lexical features were extracted from all essays using frequency counts and TF-IDF.
3	Feature Preprocessing	No stemming or stop-word removals applied to maintain originality of essays.
4	Model Training	ML models trained using 5-fold cross-validation for Logistic Regression, Decision Tree, and Support Vector Machine.
5	Model Evaluation	ML models evaluated on test set using accuracy, precision, recall, and F1-score for human-written and AI-generated essays.

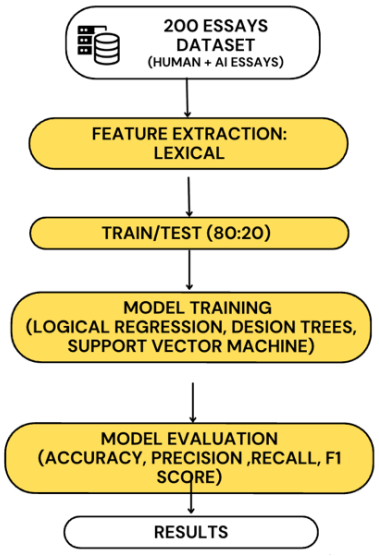

Fig. 1. Proposed methodology for lexical-based AI vs human authorship detection

https://doi.org/10.31436/ijpcc.v12i2.699

D. Experimental Setup

Table III summarizes the technical configuration and hyperparameter values applied to each of the machine learning model.

TABLE III

TECHNICAL CONFIGURATION AND HYPERPARAMETER VALUES FOR MODEL TRAINING

Component	Parameter Setting	Description
General Settings (for all models)	5-fold CV, StandardScaler, random_state = 42	Splits data into 5 folds to tune hyperparameters reliably, scaling features for normalization when required, random_state ensures reproducibility of splits
Logistic Regression	C = 1.0; max_iter = 1000	Default regularization strength (C), max_iter: max iterations to converge
Decision Tree	max_depth = None; min_samples_split = 2, min_samples_leaf = 1,	Max_depth: tree depth limit (None = unlimited), min_samples_split: min samples to split a node, min_samples_leaf: min samples at leaf,
SVM	Kernel = linear; C = 1.0; gamma = scale, probability = True	Kernel: linear decision boundary; default regularization strength (C), gamma: influence of single points (scale = default), probability: enables probability estimates

IV. RESULTS

Fig. 2 compares the top 15 Part-of-Speech (POS) distributions in human-written and AI-generated essays. Both classes showed frequent use of nouns and verbs, reflecting the academic essay format. AI-generated essays however, showed relatively higher frequencies of nouns, verbs, adjectives, punctuation, and adpositions, suggesting more compressed informational phrasing and more uniform academic sentence construction.

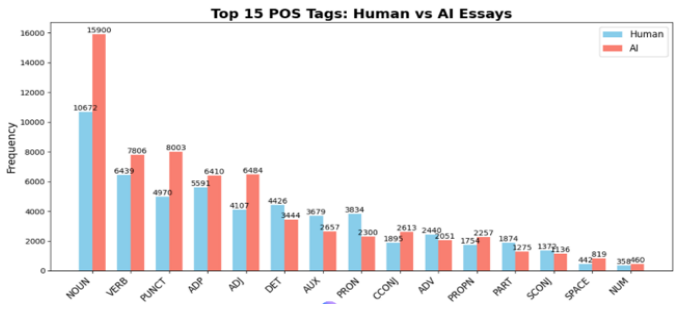

Fig. 2. Summary of top 15 Part-of-Speech (POS) tags

As illustrated in Fig. 2, human-written essays in contrast showed relatively greater use of determiners, pronouns, auxiliary verbs, adverbs, and particles, suggesting more variable and conversational sentence patterns. These differences support the use of surface-level stylometric features, but they should not be interpreted as universal markers of AI authorship. POS distributions are likely to vary according to essay topic, prompt wording, model version, and student writing proficiency. Fig. 3 demonstrates the Receiver Operating Characteristic (ROC) curves of both types of essays.

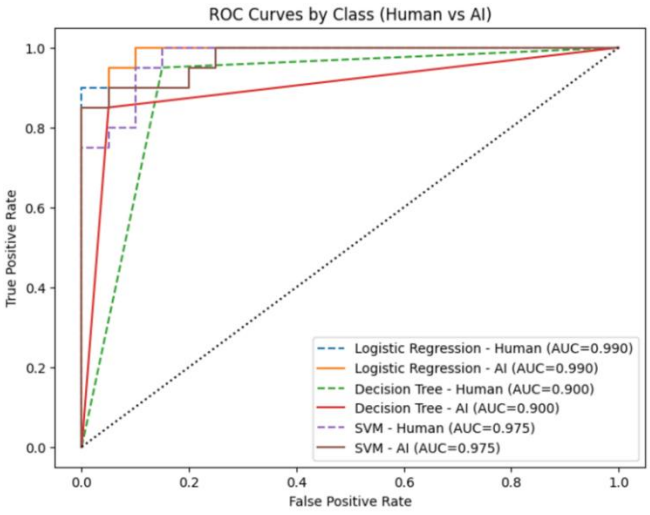

Fig. 3. ROC Curves of Human and AI Essays

The ROC curves in Fig. 3 demonstrate the comparison between the three models; Logistic Regression, SVM, and Decision Tree for classifying Human vs AI essays. The x-axis shows the false positive rate and the y-axis shows the true positive rate, with the diagonal dotted line representing random guessing. All models performed well with curves close to the top-left corner, indicating strong classification ability.

Logistic Regression performed best with an AUC of 0.990, followed by SVM with an AUC of 0.975, while the Decision Tree performed comparatively lower with an AUC of 0.900. The stronger performance of the linear models suggested that the selected lexical and TF-IDF features formed a relatively linearly separable feature space in this dataset. Therefore, the reported performance reflected the effectiveness of these models within the specific experimental setting and should not be generalized beyond the conditions of this study.

The confusion matrices presented in Fig. 4, Fig. 5 and Fig. 6 compare the performance of Logistic Regression, Decision Tree, and SVM in classifying Human vs AI essays. As illustrated in Fig. 4, the Logistic Regression classifier

https://doi.org/10.31436/ijpcc.v12i2.699

correctly classified all 20 Human essays, with 2 AI essays misclassified as Human, showing strong overall accuracy. For the Decision Tree classifier, 18 Human essays and 18 AI essays were correctly identified, misclassifying 2 instances in each class, indicating a slightly more balanced performance as shown in Fig. 5. The SVM model produced almost identical results as Logistic Regression, correctly classifying 20 Human and 17 AI essays with 3 AI essays misclassified as Human as demonstrated in Fig. 6.

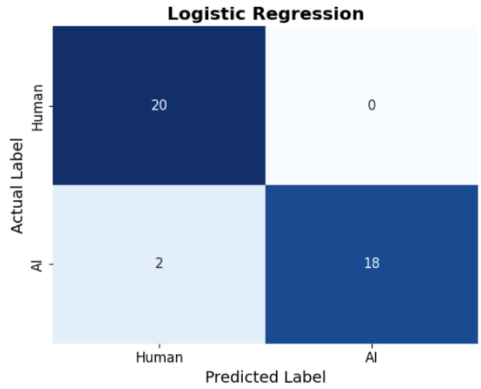

Fig. 4. Confusion Matrix for Human vs AI Text Classification for LR Model

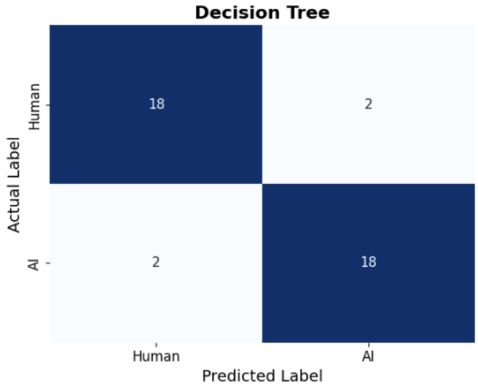

Fig. 5. Confusion Matrix for Human vs AI Text Classification for DT Model

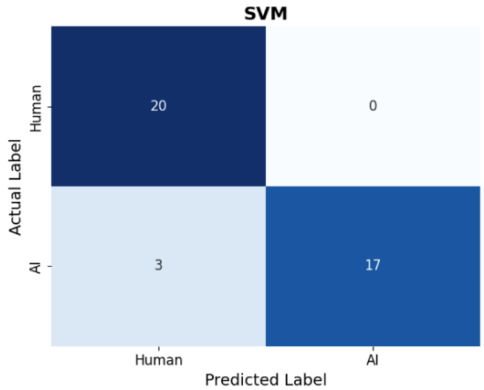

Fig. 6. Confusion Matrix for Human vs AI Text Classification for SVM Model

The results showed that all three models performed well, with Logistic Regression and SVM showing slightly higher correct Human predictions, while Decision Tree demonstrated more balanced classification across both classes. Human-written essays were generally classified more accurately than AI-generated essays, as most errors occurred when AI-generated essays were misclassified as Human. Table IV summarized the overall class-wise performance of Logistic Regression, Decision Tree, and SVM in classifying Human versus AI-generated essays.

TABLE IV
PERFORMANCE OF ML CLASSIFIERS

Model	Class	Accuracy	Precision	Recall	F1-Score
Logistic Regression	Human	1	0.909	1	0.952
	AI	0.9	1	0.9	0.947
Decision Tree	Human	0.9	0.9	0.9	0.9
	AI	0.9	0.9	0.9	0.9
Support Vector Machine	Human	1	0.87	1	0.93
	AI	0.85	1	0.85	0.919

As demonstrated in Table IV, Logistic Regression achieved the best classification performance with excellent F1-scores (0.952 for Human and 0.947 for AI) in which it correctly identified all Human essays (100% recall). The model also achieved a perfect precision for the AI class indicating that every sample predicted as AI essay was actually AI.

Following Logistic Regression, the Support Vector Machine (SVM) classifier also demonstrated strong performance, achieving 100% accuracy and recall for Human essay detection while having an F1-score of 0.930. However, SVM achieved an acurracy and recall of only 0.850 for AI-generated essays, suggesting that it failed to identify 15% of the actual AI instances, resulting in the lowest AI accuracy among the three models.

In contrast, the Decision Tree classifier achieved comparatively lower but consistent performance across both classes with 90% detection accuracy for both Human and AI essays, and 0.900 for all other measures. While stable and balanced, it did not achieve the same level of predictive accuracy as Logistic Regression and Support Vector Machine. These overall results support the usefulness of lexical stylometric features as an interpretable baseline for AI-generated essay detection.

Fig. 7 illustrates the performance comparison of the three classifiers. Linear models (Logistic Regression and SVM) were found to outperform non-linear tree-based approaches for lexical and TF-IDF-based authorship detection in this dataset as depicted in Fig. 7. This indicates

that the extracted features provide strong linear separability between human-written and AI-generated essays. Nevertheless, the high performance may partly reflect dataset simplicity, controlled prompt design, balanced class distribution, and topic consistency between Human and AI-generated essays.

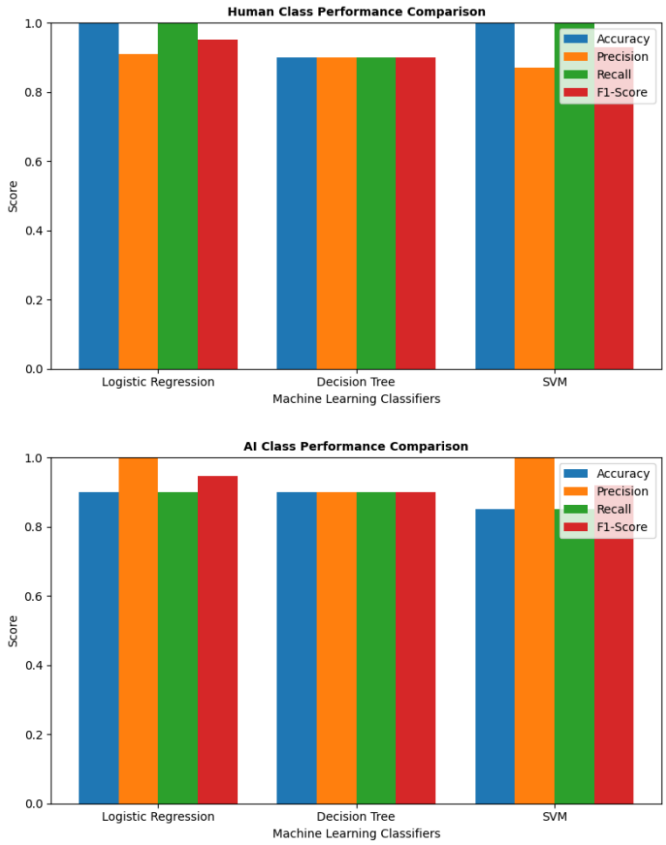

Fig. 7. Performance comparison of classification models for Human vs AI

V. Limitation

Although the proposed lexical stylometric framework achieved strong classification performance, several limitations must be acknowledged. First, the dataset contains only 200 essays, making it suitable for pilot-level feasibility testing but insufficient for broad generalization. Second, the AI-generated essays were produced under controlled prompt conditions, which may have made the two classes easier to separate than in real academic submissions. Third, the study deliberately excluded semantic coherence features, contextual embeddings, discourse-level representations, and transformer-based models. As a result, the findings demonstrate the value of lexical features as an interpretable baseline, not as a replacement for semantic or contextual approaches. Fourth, the reported accuracy may be influenced by the balanced dataset, similar essay length, and controlled topic distribution. Finally, the model was not tested against paraphrased AI-generated essays, edited AI-generated essays, multiple LLM families, or essays from other academic disciplines. These limitations indicate that the framework should be used for research and screening support only, not as sole evidence in academic misconduct decisions.

VI. Conclusion

This study evaluated the extent to which lexical stylometric features can distinguish AI-generated from human-written undergraduate essays using interpretable machine learning classifiers. The results showed that lexical richness, word usage distributions, POS frequencies, and TF-IDF-based n-gram features can provide strong in-domain discriminatory performance, with Logistic Regression achieving the highest F1-score of 0.952 in detecting human-written essays and 0.947 in detecting AI-generated essays within a controlled test setting. SVM performed equally strongly, while Decision Tree showed comparatively lower but consistent performance.

The findings support the feasibility of lexical stylometric features as a lightweight and interpretable baseline for AI authorship analysis in educational contexts. However, the study does not claim that lexical features alone are sufficient for robust real-world AI-based text detection. The limited dataset size, controlled prompt design, and absence of cross-model or cross-domain validation require cautious interpretation. Future research should expand the dataset, include multiple generative AI models, evaluate paraphrased and human-edited AI outputs, and compare lexical-only models with semantic, contextual, and transformer-based approaches.

Acknowledgment

This research is funded under the Fundamental Research Grant Scheme (FRGS/1/2023/SSI09/UIAM/02/1) sponsored by the Ministry of Higher Education (MOHE), Malaysia. The authors would like to express their sincere gratitude and appreciation to the constructive comments by the IJPCC reviewers, which significantly contributed to improving the quality of this manuscript.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

Author(s) Contribution Statement

Tahmida contributed to the conceptualization of the study, data collection, methodology, feature extraction, model implementation, analysis of results, and preparation of the

initial manuscript draft. Suriani contributed to the supervision of the research, research design, methodological validation, interpretation of findings and critical revision of the manuscript. Adlina contributed to the funding acquisition, supervision of linguistic and academic-writing aspects of the study and interpretation of essay-related features. All authors reviewed and approved the final version of the manuscript.

Data Availability Statement

The data supporting the findings of this study consists of human-written and AI-generated undergraduate essays collected from students at the International Islamic University of Malaysia and generated for research purposes. The data are not publicly available due to the copyrighted materials owned by the funding organization which includes student writing samples that are subject to privacy and ethical restrictions. However, anonymized data or relevant experimental details may be made available upon reasonable request, subject to institutional approval and ethical restrictions.

Ethics Statement

This study involved the analysis of undergraduate essay texts for authorship classification purposes. Human-written essays were collected under controlled academic conditions and were anonymized before analysis to protect participant identity and confidentiality. No personally identifiable information was used in the feature extraction or model training process. The study was conducted in accordance with institutional research ethics guidelines.

References

[1] T. Kehkashan, R. A. Riaz, A. S. Al Shamayleh, A. Akhunzada, N. Ali, M. Hamza, and F. Akbar, "AI generated text detection: A comprehensive review of methods, datasets, and applications," *Computer Science Review*, vol. 58, 100793, 2025.

[2] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection," Cornell University, 2023. Preprint at arXiv. arXiv:2301.07597v1

[3] H. Alamleh, A. A. S. AlQahtani and A. ElSaid, "Distinguishing human-written and ChatGPT-generated text using machine learning," *Systems and Information Engineering Design Symposium (SIEDS)*, Charlottesville, VA, USA, pp. 154-158, 2023, doi: 10.1109/SIEDS58326.2023.10137767.

[4] N. Krawczyk, B. Probierz and J. Kozak, "Towards AI-generated essay classification using numerical text representation," *Applied Sciences*, vol. 14(21), 9795, 2024.

[5] A. Masih, B. Afzal, S. Firdoos, J. Mahmood, A. Ali, M. S. Abdulnabi and D. M. Balungu, "Classifying human vs. AI text with machine learning and explainable transformer models," *Scientific Reports*, vol. 15, 43310, 2025.

[6] A. Agarwal and P. Sharma, "A comparative study of stylometric features for authorship discrimination using machine learning," *Proceedings of the 2023 International Conference on Learning Representations (ICLR 2023)*, pp. 1–12, 2023.

[7] A. Abbasi, A. R. Javed, F. Iqbal, Z. Jalil, T. R. Gadekallu, and N. Kryvinska, "Authorship identification using ensemble learning," *Scientific Reports*, vol. 14 (1), 6789, 2024.

[8] A. Yadagiri, L. Shree, S. Parween, A. Raj, S. Maurya, and P. Pakray, "Detecting AI-generated text with pre-trained models using linguistic features," *Proceedings of the 21st International Conference on Natural Language Processing (ICON), pp.188-196*, AU-KBC Research Centre, Chennai, India, NLP Association of India (NLPAI), 2024.

[9] H. Desaire, A. E. Chua, M. Isom, R. Jarosova, D. Hua, "Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools," *Cell Reports Physical Science*, vol. 4 (6), 101426, 2023.

[10] L. J. A. Chang, "Detecting AI generated text: A Comparative study of machine learning algorithms," in *Proceedings of the 29th World Multi Conference on Systemics, Cybernetics and Informatics (WMSCI 2025)*, pp. 294–299, 2025.

[11] S. Ramnath, K. Pandey, E. Boschee, and X. Ren, "CAVE: Controllable authorship verification explanations," in *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025 – Long Papers)*, pp. 8939–8961, Association for Computational Linguistics, 2025.

[12] J. Zu, M. Fauss, and C. Li, "Effects of generation model on detecting AI-generated essays in a writing test," *Educational Testing Service*, Princeton, NJ, 2024.

[13] J. Kubrusly, E. D. dos Santos, and L. S. Pelegrino, "Detecting AI-generated texts using machine learning models," *Communications in Statistics: Case Studies, Data Analysis and Applications*, vol. 11 (4), pp. 495-512, 2025.

[14] C. Chaka, "Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review," *Journal of Applied Learning & Teaching*, vol. 7(1), pp. 115–126, 2024.

[15] D. Weber-Wulff et al., "Testing of detection tools for AI-generated text," *International Journal for Educational Integrity*, vol. 19, article 26, 2023.

[16] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLOS ONE*, vol. 14, no. 11, e0224365, 2019.