Hybrid Lexical and Forensic Linguistic Features for AI-Generated Text Detection Using Machine Learning

Amira N. Abd Al-Jaleel[1] and Husam J. Mohammed[2]
[1]*Department of Computer Science, College of Computer Sciences and Information Technology, University of Anbar, Al-Jamea Str. 15, 31001 Ramadi, Iraq*
[2]*Department of Artificial Intelligence, College of Computer Sciences and Information Technology, University of Anbar, Al-Jamea Str. 15, 31001 Ramadi, Iraq*
{ami24c1005, hussamjasim}@uoanbar.edu.iq

Keywords: LLM, Forensics, Cybersecurity, Linguistic Features, Hybrid Framework, Feature Fusion.

Abstract: The rapid and widespread use of large language models (LLMs) has ushered in the rapid increase in the quantity of AI-produced text everywhere on the Internet, which has necessitated significant concerns about the field of digital forensics, cybersecurity, and academic integrity. The correct distinction between text produced by human authors and text produced by artificial intelligence can be described as a difficult task, which is largely conditioned by the growing fluency of the generative models as well as by the little generalizability of the existing detection tools. The model in question uses a combination of TF-IDF in unigrams and bigrams with 21 linguistically interpretable features. The features are created to match the text and capture it in the stylistic, syntactic, semantic, and discourse coherence levels. They were conducted on the large-scale Human vs. LLM dataset that contains 788,000 text samples of various sources of humans and several advanced and modern language models generated texts. It was tested on the performance of several classification algorithms, such as logistic regression, linear support vector machine, random forest, and LightGBM, but not limited to these products. The experimental outcomes proved that retaining natural linguistic patterns and applying criminal linguistic characteristics played an important role in enhancing detection performance. The LightGBM classifier achieved the best results, with an accuracy rate of 93.4% and an F1 score of 0.93 % for human texts and 0.94 % for AI-generated texts. These findings, moving away from hidden deep learning models, help forensic experts because the new method makes digital investigations more reliable and makes AI more responsible. This transparent tool ensures that textual evidence is ready for court; consequently, study contributes to better safety in the digital world.

1 INTRODUCTION

Recent technological advancements have improved online anonymity, which has historically been used for illegal purposes. Cybercrime has become more prevalent and diverse, making it difficult for law enforcement to identify. In response to this growing threat, forensic linguistics has evolved as an important tool for combating cybercrime. It verifies authorship and creates sociolinguistic profiles for investigators to steer their research.[1].

However, Proliferation of digital text poses challenges in ensuring its authenticity and provenance. the ability to discriminate Helpful and misleading textual content generated by humans and AI has become an urgent social and technical concern [2]. Even if you don't, the repercussions are severe. Advanced contract fraud and plagiarism pose a threat to assessments in educational institutions, compromising the value of education. The wide spread of using AI bots, false information, and propaganda created by AI has the potential to shift public opinion and gradually weaken public confidence [3].

Although these advancements, the current detection techniques frequently face issues with interpretability, generalization and immune to stylistic change [3]-[5]. Deeper stylistic, syntactic, and semantic patterns that define the way human write may be neglected by many models that depend only on TF-IDF (Term Frequency-Inverse Document Frequency) or embedded-based features result in disparity of the forensic linguistic characteristics. This gap motivates the development of hybrid

approaches that combine traditional textual representations with handcrafted forensic linguistic features. accordingly, this has motivated to propose intelligible models in forensic environment [6]. Existing approaches are often limited though they promise. Most detectors lack generalizability and only work particularly well on the text of the particular LLMs (Large Language Models) they were trained on and not against newer or unseen models [7]. Moreover, they are prone to low-technology adversarial tricks, such as light paraphrasing or prompt-engineering that can remain unnoticed. It is an active research need to develop an all-inclusive and high-quality machine learning technology that combines a user-friendly feature engineering with advanced and carefully assessed in the generalization aspects.[3], [4].

The main goal of this research, is to design, apply and critically assess a machine learning model that integrates TF-IDF features with 21 forensic linguistic features capturing stylistic, syntactic, semantic, and writing style characteristics. Though AI demonstrates the ability to produce factual reliable content, a critical missing ingredient remains in the quality and depth of such content to cross the bridge of this critical gap by coming up with explainable machine learning models that are especially tailored to digital forensic investigations. The analysis of the experimental results and analysis are presented in Section 4. Section 5 addresses the discussion of the findings, limitations, and general implications. Lastly, Section 6 gives a conclusion and future work directions of the paper.

2 RELATED WORKS

As LLMs become more advanced, researchers have developed methods to differentiate between human written and artificially generated text. many approaches, including feature-based machine learning and forensic linguistic analysis, target multiple aspects of the complicated topic.

Preliminary feature-based machine learning approaches showed promising scores in AI-generated.

Text detection, using Random Forest classifiers with F1-scores of 0.99 % for English emails, focusing on typographic errors and stylometric features, sentence length statistics, though they noted limitations in handling spell-corrected texts and performance Variations across languages [8].

In addition, [9] researchers maintained various machine learning models, where they achieve 88% accuracy with Random Forest and only TF-IDF features. However, they clarify that these features alone may be not enough for complex text types and might lead to interpretability challenges and dataset bias in forensic applications.

Following that time the field has growing to better understanding of the fundamental "fingerprints" of artificially generated text. Sharanya as well as Nawarathna [10] find out that conversion techniques such as BERT very efficient compared to traditional models yet they retain POS (Parts of Speech) features that make them highly attractive for abstracts of research.

Further proof by Ardeshirifar [11] who achieved 94% F1-score using hand-crafted features, provide a level of linguistic insight and explainability that deep learning black box don't offer.

A study by, Fariello et al [12] provided thorough analysis of difficulties and developments in machine generated text, depending on multiple structural paradigms, the machine detection field has advanced significantly bringing out limitations in feature design and generalization across various kinds of contents.

Utilizing the same Kaggle Human vs. LLM corpus as this study by Repaka, Bondugula, and Adibhatla [13], they concentrated on the effectiveness of computation and distributed training, achieving an 80% accuracy using LSTM and raw embeddings. despite their approach increased computing efficiency, it lacks in maintain good classification accuracy on large datasets without the use of advanced feature engineering.

Godghase et al. [14] further investigate this scale using over 750,000 sentences to show how feature analysis remains reliable discriminator by training the models with both feature analysis and embedding. the feature extraction phase, they focused on a set of linguistic characteristics represented via word choice and the structure of the sentence. they achieved classification accuracy up to 96%.

Later literature focused on the challenges in AI detection tools. Elkhatat et al [15]. evaluated commercially available detection systems and found variability in their ability to correctly classify Al-generated versus human-written text.

The results of research have also revealed the false positive occurrence especially in the cases of paraphrasing or stylistic manipulation of human text. This highlights the need to come up with stronger methods. A modern text detector based on AI is built on the Traditional Authorship Attribution and Stylometry. The basis of these approaches is that writers create unconsciously.

Uniform and quantifiable habits in their style of writing. Khera et al. showed that the distinction between the AI-generated and human-written scientific texts could be improved with the help of the contextualized embeddings, but at the cost of a high computational load [16].

Determining Al generated works is usually based on the existence of measurable stylistic indicators. As can be observed, methods like statistical analysis and perplexity are based on the underlying structure of the language models to be used in order to identify weak artificial signatures of generated texts. [17]. Text detectors based on machine learning are based on supervised learning models which are trained.

On massive datasets of pairs of labelled texts, some of these are written by humans and others by artificial intelligence. The foundations of these classifiers can be either conventional models which use stylometric features or deep neural networks which are trained on discriminative patterns of raw text as outlined above by Maddugoda [18].

Such a strong detection is needed due to the increased risks of AI misuse. Ranade et al. [19] demonstrated the potential for fraudulent Cyber Threat Intelligence (CTI) to compromise security databases. Furthermore, studies by Kumarage et al. [20] on Twitter timelines and Zeng et al. [21] on hybrid human-AI essays emphasize that as AI becomes more sophisticated, detection must rely on the subconscious "stylometric signals" inherent in human writing.

In spite of the current active research, there is still a number of serious limitations in the existing paradigms of detection, it struggles with cross-domain generalization, computational efficiency, and explainability. Building on these findings, this study leverages a hybrid approach by combine TF-IDF representations with 21 handcrafted forensic features capturing stylistic, syntactic, semantic, and writing style characteristics to address the key limitations of cross-domain generalization and computational efficiency seen in prior studies enabling accurate detection while remaining computationally efficient.

3 PROPOSED METHODOLOGIES

The proposed methodology of this research as shown in Figure 1. starts with the publicly available Human vs Chatbot dataset from Kaggle [22] a dataset of text generated by human and a text generated by Generative AI (Gen AI). The text go through a set of preprocessing steps that will be explained later in the next sections. After that, a set of features extracted, in this study two types of features will be extracted including a handcrafted forensic-linguistic features and TF-IDF features. Next, these features are both fused to make the main source of features that the machine learning model will train on. Then the combined features are split into train and test sets. So that after training the model could be tested on the unseen data. A detail about all these steps will be explain in the next sections.

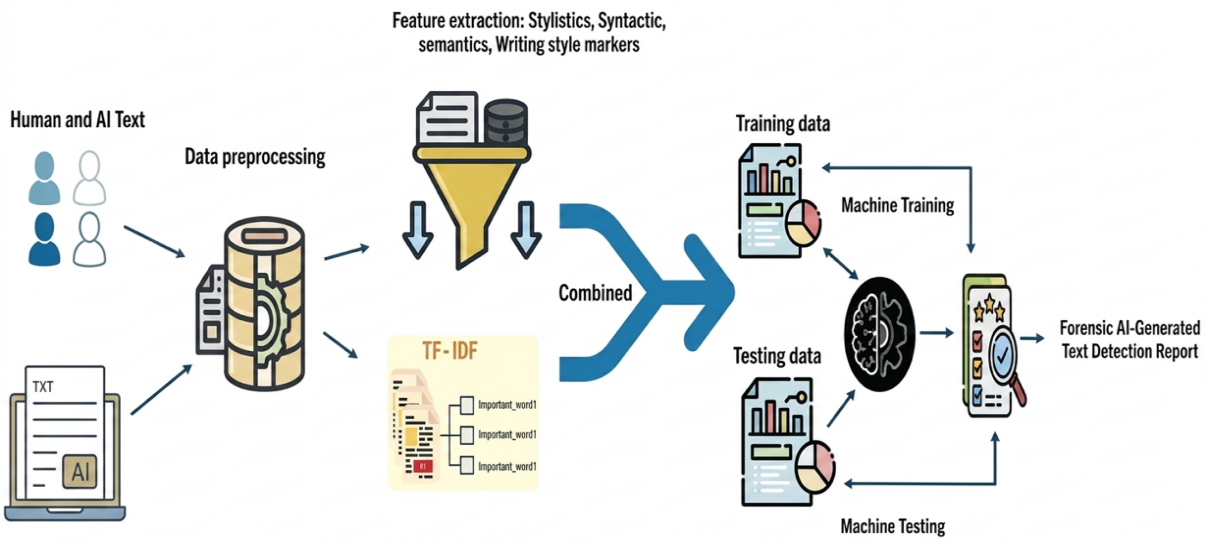

Figure 1: Methodology of the proposed model.

3.1 Dataset

Table 1 outlines the main characteristics of the dataset used in this study. It consists of five features: the text content, the prompt ID, the length of each text, the word count, and the source label whether human-written or Al-generated authors facilitating automated attribution using machine learning algorithms by capturing stylistic, syntactic, and semantic signals across the dataset instances. this is the exact goal of forensic linguistic in authorship verification and other tasks. The dataset is made up of the texts generated by humans about 56% of the various types Articles, social media posts, Academic content, and emails.

The remaining 44% AI models These AI-generated samples different models, including GPT-3.5 is 7%, and 49% for other large-scale models such as Claude, and Gemini-Pro,BLOOM-7B,Falcon-180B, Flan-T5, Goliath-120B, and LLaMA-13B.the techniques built on this corpus can provide useful forensic analytics in situations specially cybercrime investigations, where evidence may appear in emails, chat logs, or articles determining human vs automated origination is crucial for inquiry and adjudication. In addition, the empirical credibility of the study improved by realistic data amounts, that relatively well-proportioned distribution between human and AI samples.

Table 1: Dataset specifications.

Size	Class distribution	Human Resources	AI-Resources
788,000 text documents	56% Human, 44% AI	Articles, social media posts, Academic content, emails	GPT-3.5, GPT-4, Claude, Gemini-Pro, Bloom-7B, Falcon-180B, Flan-T5, Goliath-120B, LLaMA-13B

3.2 Data preprocessing

Before feature extraction step, entire dataset went through preliminary processing to ensure consistency and uniformity of the data.

Initially, every text converted into a common string format.as they are content-specific, hyperlinks and emails were excluded since they are more stylistic and might introduce irrelevant pattern to authorship. emojis were transforms into textual representations before removing them in order to prevent adding malicious markers into data, then text was further normalized by eliminating whitespace and performing lowercasing to preserve lexical distinction even more guarantee consistency.

Methods like lemmatization, stemming, and stop words removal were purposely avoided as the proposed study focus on the forensic linguistics patterns rather the pure semantic content. As these techniques can conceal significant clues that are central to writing style because these forensic signals represent evidence not the noise must be eliminated.

Such distinctions in language application, morphological decisions, and syntactic regularities that usually set human written text apart from Al-generated content could be disguised by limiting words to their base forms or eliminating indicators. The proposed method enhances the model's ability to distinguish between machine-generated and human-written texts in a way that is criminally significant, by preserving the original lexical and structural forms, thereby enabling the identification of subtle stylistic and distributional patterns. No additional filtering was performed based on word frequency, sentence length, or token length. This choice was made to preserve the natural distribution of linguistic structures, such as short phrases, longer constructions, and rare terms, all of which can contribute to distinguishing between Al-generated content and human writing.

3.3 Feature Engineering and Representation

To convert the original text data into numerical representations that can be used by machine learning models, a hybrid feature extraction strategy was adopted. Initially, two main feature representations were created following the preprocessing stage. In order to capture word importance and local contextual relationships, TF-IDF vectors comprising unigrams and bigrams were first extracted. The subsequent stage of the research involved the development of a set of twenty-one linguistic features that had been designed manually. These features encompass stylistic, syntactic, semantic, and discourse-level aspects, including sentence length, lexical richness, punctuation patterns, part-of-speech distribution, sentiment metrics, and the utilization of formal and transitional expressions.

These characteristics assist in differentiating between human-written texts and those generated by artificial intelligence.

The technique has been demonstrated to synthesis written material, whilst simultaneously determining the manner in which it has been written.

This is of critical importance for the effective detection of AI-generated text, which must be both robust and explainable. Linguistic profiling studies have demonstrated that stylistic patterns, such as sentence complexity, lexical diversity, and punctuation usage, exhibit systematic differences between human and machine-generated texts.

These differences provide interpretable signals for classification models. The extracted features include stylistic measures, such as sentence and lexical complexity, lexical diversity, and punctuation distribution patterns.

In addition, the features encompass syntactic and grammatical characteristics, represented through normalized distributions of Part-of-Speech types, including nouns, verbs, adjectives, adverbs, and pronouns. Furthermore, semantic indicators are derived from sentiment analysis, where each text is mapped to a polarity measure in the range [-1,1] and a subjectivity score in the range [0, 1]. These measures embrace emotional orientation and level of objectivity. Also, explicit style cues, e.g. the frequency of transitional expressions (e.g., how however, therefore) are included. The fact that linguistically informed features should be combined with empirical representations of texts has proven to come up with a more sophisticated and discriminative definition of text. This enables models to simultaneously acquire lexical, structural, stylistic and semantic cues.

3.4 Text Representation Using Enhanced TF-IDF

To convert the textual content into numerical vectors, The Term Frequency-Inverse Document Frequency (TF-IDF) vectorization process was employed. TF-IDF is a method of measuring the importance of a word in a specific document, while reducing the influence of frequently occurring words in the entire corpus. The inclusion of unigrams and bigrams was a deliberate strategy to capture both individual word usage and short contextual dependencies between consecutive words. The maximum features of the TF-IDF space were limited to the top 5,000 most significant n-grams, in order to achieve a balance between representational richness and computational efficiency.

Term frequency (TF) is a metric used to ascertain the frequency of a word within a text relative to the total number of words in said text. A higher TF-IDF score is indicative of a greater degree of relevance of the word to the specific text. A higher TF-IDF score is indicative of a greater degree of relevance of the word to the specific text.

3.5 Features Fusion (Combined)

Effectively distinguishing between human-authored and AI-generated text requires representations that captures both what is written and how it is written. Meaning that lexical information alone is not enough to recognize deeper stylistic and structural properties. As many research demonstrated that writing style, syntactic organization, play a crucial role in detecting the source of a text, particularly when dealing with highly fluent machine-generated language [1], [6]. Asa result of these findings, this study adopts a feature fusion strategy that combines lexical, linguistic, and stylometric information into a unified representation.

Specifically, TF-IDF features were fused with a set of twenty-one handcrafted forensic-linguistic and stylometric features to capture higher-level textual properties that are not reflected by word frequency alone.

While TF-IDF encodes statistical word usage and local contextual patterns, the handcrafted features model complementary aspects of textual behavior, including lexical diversity, syntactic complexity, punctuation usage patterns, sentence length statistics, and character-level distributions.

These characteristics have been proven in researches to be particularly successful in identifying stylistic consistency, diversity, and discourse-level clues that differ between human authored and AI-generated text [7], [23].

3.6 Machine Learning Models

This study uses four machine learning methods to evaluates the efficacy of the suggested hybrid feature representation. These models were created to capture a variety of the learning paradigms to ensure that there would be a strong comparison between linear, margin-based, ensemble-based and gradient boosting to be able to demonstrate deeply the interaction of various modeling concepts to linguistic and stylistic factors in the process of the separation of writing done by AI and text done by humans.

Logistic Regression is reliable and understandable baseline model.in addition to its suitability for high-dimensional sparse representations such as TF-IDF vectors. Its incorporation allows for evaluating the linear distinguishable of human and Al-generated text

when lexical and handcrafted linguistic features are integrated.

Also Because of its ability to deal high-dimensional feature spaces and its efficiency in optimizing class separation through margin optimization. Support vector machine (SVM)are especially useful for Al-generated text identification because they can capture small stylistic differences encoded in language representations and are resilient to feature sparsity. Because of this, Linear SVM is a great option for identifying subtle distinctions in writing styles produced by humans and by machines.

While Random Forest As an ensemble of decision trees its capability of capturing subtle interactions between handcrafted forensic-linguistic features and lexical patterns that linear models may fail to exploit and also to model non-linear relationships between features and class labels. In addition, Random Forest is also inherently resistant to overfitting and provides a feature importance signal that can be used to reinforce the classification decision. LightGBM (light Gradient Boosting Machine) is a high-performance gradient boosting framework that accommodates high dimensional and wide data. This is because it can be trained to learn complex decision boundaries by sequential boosting, which renders it appropriate to the case at hand that is a mix of sparse TF-IDF features and dense linguistic indicators. The large potential of generalization and high computational efficiency is attributed to LightGBM, which is one of the reasons why it is a good option to achieve high classification accuracy on large text corpora and be scaled.

3.7 Evaluation Metrics

The performance of all classification models was evaluated using standard metrics for binary classification. Accuracy was used to measure the overall proportion of correctly classified samples, while precision assessed the proportion of predicted positive instances that were correctly identified. Recall measured the model's ability to identify all actual positive samples. In addition, the F1-score was used to provide a balanced evaluation by combining precision and recall into a single performance measure, making it particularly suitable for datasets where both false positives and false negatives are important.

4 EXPERIMENTAL RESULTS AND CLASSIFICATION RESULTS

This section presents the classification performance obtained through three experiments. The results were provided using common evaluation metrics, such as accuracy, precision, recall, and Fl-score, in order to assess the model's efficiency in discriminating between the human generated text and the artificially generated one.

4.1 Dataset Description and Importance of Data Splitting

HumanvsChatbot dataset from Kaggle [22], which its publicly available was used in this study. It contains (788,000) entries text instances. And consists of five features: the text content, the source label (human-written or AI-generated), the prompt ID, the length of each text, and the word count.

The process of splitting the data is a very crucial step in machine learning.to make sure model is tested on un seen, unbiased data, in order to reflect its performance in real world scenarios.

This is important where the ability to generalize beyond the training data is important in order to make the classification task succeed. the data were divided into training and testing subsets. 20% of the samples were reserved for testing, while the rest 80% of data used for model training. Each instance was annotated with a binary class label indicating its origin, as human-authored texts were assigned label 1 and AI-generated texts were assigned label 2. These labels were used exclusively for assessing the model's ability to distinguish between human and Al-generated content.

4.2 Experimental configuration

The three experiments were conducted in order to assess the role of preprocessing and feature engineering on Al-generated text detection and classification task explained.

4.2.1 Experiment 1

Lexical-Only Baseline Representation: First experiment run with basic preprocessing steps applied including Lemmatization, Punctuation marks, stop words removal, URLs, email addresses, whitespace, and noise removal. Only TF-IDF vectors with a maximum of 5,000important unigrams and bigrams feature were extracted, without the incorporating of the 21 handcrafted forensic linguistic features. This setup served as a baseline to evaluate the predictive power of lexical representations alone. Preliminary results showed moderate performance throughout the models proving limitations in relying only on TF-IDF feature vectors specially if the goal is to detect deeper characteristics important in distinguishing between AI-generated text and human-authored text is not enough.

Table 2 represents a summary of the performance under the Lexical-Only Baseline Representation experiment. Were the highest accuracy of 0.88, is attained by the LightGBM classifier, then Logistic Regression and Linear SVM providing same results. Finally, Random Forest showed weaker recall for human-authored texts, indicating limitations when relying solely on lexical representation.

4.2.2 Experiment 2

Normalized Linguistic Feature Representation: Second experiment executed by combining 21 forensic linguistic features with TF-IDF vectors were again limited to 5,000 features preserving same preprocessing steps in experiment 1.

Table 2: TF-IDF-Based lexical baseline.

Model	Accuracy	Precision (Human /AI)	Recall (Human /AI)	F1-score (Human /AI)
Logistic Regression	0.84%	0.82/ 0.86	0.82/ 0.85	0.82/ 0.86
Linear SVM	0.84%	0.82/ 0.86	0.82/ 0.86	0.82/ 0.86
Random Forest	0.82%	0.89/ 0.79	0.68/ 0.94	0.77/ 0.85
LightGBM	0.88%	0.88/ 0.88	0.85/ 0.91	0.86/ 0.90

The aim is to reduce noise and standardize text while leveraging stylistic, syntactic, semantic, and writing-style markers the results demonstrate a notable enhancement in model performance compared to the TF-IDF-only baseline, verifying the advantages of the linguistic feature in the classification and detection task.

Experiment 2 results summarized in Table 3, demonstrated a noticeable improvement across evaluation metrics when compared to the lexical-only baseline. As highest accuracy, precision and recall values for both classes were achieved due to the fusion of linguistic normalization and handcrafted features, particularly for Al-generated text noting that LightGBM surpassed all models.

Table 3: Normalized linguistic feature representation.

Model	Accuracy	Precision (Human /AI)	Recall (Human /AI)	F1-score (Human /AI)
Logistic Regression	0.80%	0.77/ 0.82	0.77/ 0.82	0.77/ 0.82
Linear SVM	0.84%	0.82/ 0.86	0.82/ 0.86	0.82/ 0.86
Random Forest	0.86%	0.91/ 0.83	0.75/ 0.94	0.82/ 0.88
LightGBM	0.93%	0.91/ 0.94	0.93/ 0.93	0.92/ 0.93

4.2.3 Experiment 3

Hybrid Feature Fusion with Minimal Preprocessing: The experiment begins with a light preprocessing stage, in which stop words and punctuation are not removed, and neither lemmatization nor stemming is performed.TF-IDF feature vectors (with 5,000 features) were used, combined with 21 manually designed criminal linguistic features. This setup outperformed the previously mentioned experiments, achieving the highest overall performance, particularly in ensemble models such as LightGBM and Random Forest. The results confirm a strong ability to distinguish between human-written and AI-generated texts by preserving natural writing patterns while incorporating forensic linguistic features.

As shown in Table 4, the third experiment achieved the best overall performance among all settings, as the feature space evolved from purely lexical features to hybrid criminal indicators. In the "Lexical Features Only" experiment, the models relied on text that had been cleaned and processed using back-of-the-book alignment, resulting in a base accuracy of 0.88 for the LightGBM model. while, the third experiment showed a significant improvement excluding back-of-the-line by maintaining the punctuation and stop words, and incorporating manually designed forensic features, raised the

LightGBM model's accuracy to 0.93 and the F1 score to 0.94 in detecting AI-generated texts.

Table 4: Hybrid feature fusion with minimal preprocessing.

Model	Accuracy	Precision (Human /AI)	Recall (Human /AI)	F1-score (Human /AI)
Logistic Regression	0.81%	0.79/ 0.83	0.79/ 0.83	0.79/ 0.83
Linear SVM	0.86%	0.84/ 0.87	0.84/ 0.87	0.84/ 0.87
Random Forest	0.87%	0.90/ 0.85	0.79/ 0.93	0.84/ 0.89
LightGBM	0.93%	0.92/ 0.95	0.94/ 0.93	0.93/ 0.94

Figure 2 shows the Receiver Operating Characteristic (ROC)curves of the validation generalization test for the LightGBM classifier The curve demonstrates a reliably high true positive rate across different false positive thresholds with the most consistent performance (AUC=0.99) This indicates a strong discriminative capability of the model in distinguishing between human-authored and Al-generated texts.

The precision-recall curve as Figure 3. presents the average precision (AP) of LightGBM model where it is approximately 0.99 percentage, which means that the curve has a high precision level at high recall rates. The validation of this statement is that this model can offer a highly satisfactory balance between accuracy and recall, especially in a case where the classes distributions are somewhat skewed, which is necessary to successful identification of AI-generated content.

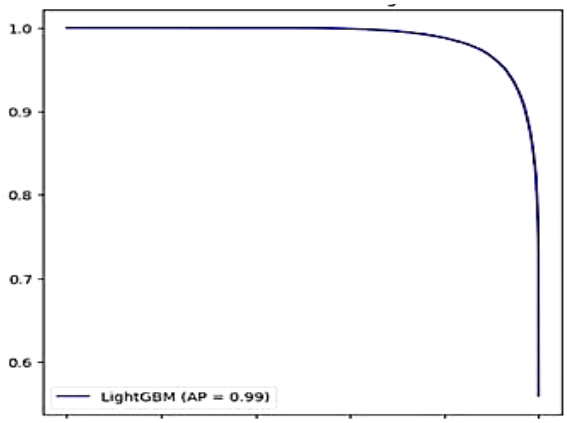

Figure 2: ROC curve for LightGBM.

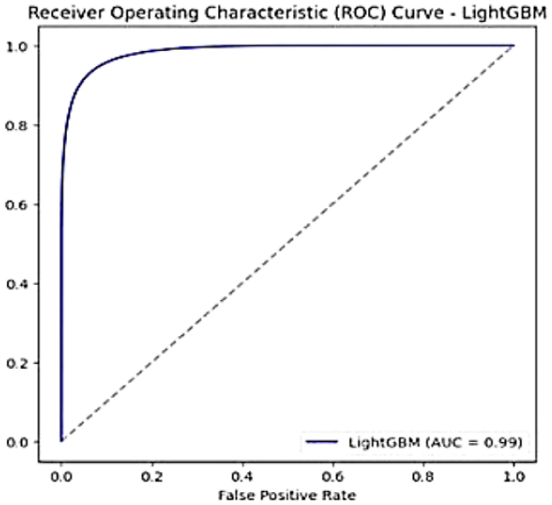

Figure 3: Precision-r ecall curve for LightGBM.

The confusion matrix of the LightGBM model validates the classification accuracy as illustrated in Figure 4 which reveals that the majority of human-authored and Al-generated samples are correctly classified.

A relatively small number of misclassifications can be noticed, as some incorrectly labeled human texts as Al-generated and vice versa. These errors may be attributed to overlapping stylistic characteristics between highly polished human writing and fluent AI-generated text, explaining the inherent difficulty of the task.

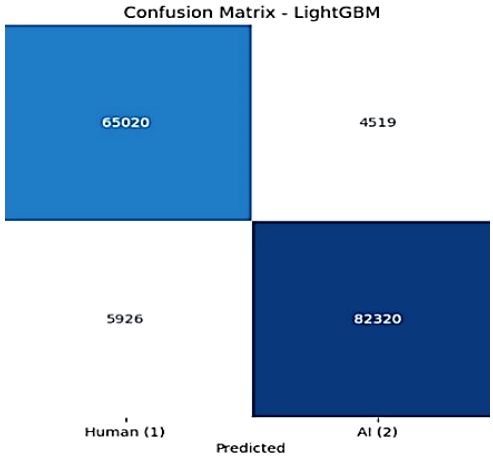

Figure 4: The confusion matrix of LightGBM.

Table 5: Statistical summary of linguistic features.

Sample ID	Avg. Sentence Length	Avg.Word Length	Lexical Richness (TTR)	Punctuation Density	Num. of Questions	Sentiment Polarity	Sentiment Subjectivity
705752	23.08	5.39	0.44	0.017	0	0.137	0.563
222201	14.84	4.91	0.43	0.034	2	0.133	0.392
270777	18.39	4.90	0.57	0.021	1	0.009	0.507
194271	11.28	4.78	0.42	0.034	0	0.108	0.414
78217	4.36	3.52	0.26	0.135	16	0.379	0.507

4.3 Feature Importance and Error Analysis

The accuracy of performance of the proposed model was high on a hidden data feature. One needs to comprehend rationality of decisions made by the model in order to ensure the authentic AI identification. Specific linguistic can be studied as it is demonstrated in the analysis. with some of these indicators being offered in Table 5, we will be able to perform a complete analysis of the scores to expose the distinct stylistic finger taps, left by various authors. First of all, it can be observed that the lexical richness (type token ratio) has been proposed as a significant discriminator.

For instance, sample 270777 exhibits a high richness score of 0.57, reflecting varied word choices typical of human expression. Conversely, lower scores (e.g.0.26 in sample 78217) indicate common, repeated writing patterns in Al-generated content. A substantial degree of variation in average sentence length is evident in relation to structural complexity, with sentences ranging from those of a sophisticated nature, exhibiting an average length of 23.07 words, to those of a simpler nature, with an average length of 11.27 words. The distinction between human authorship and statistical consistency in AI text is of significant importance. However, an examination of the punctuation behavior reveals an extremely elevated punctuation ratio of 0.135 and a question count of 16. Such 'bursty' utilization of punctuation is a compelling indication that conventional TF-IDF methodologies would likely categories as noise and eliminate. Finally, the issue of subjectivity and sentiment is addressed. The observations indicate subjectivity scores as high as (0.56). This metric is crucial for detection, given that language models (LLMs) are typically reinforced to remain neutral (closer to 0.0), whereas human writing often contains the inherent bias and personal perspective reflected in these higher scores.

5 CONCLUSIONS

The study demonstrated a robust approach that integrates the forensic linguistic analysis and traditional lexical. Representations are usable actively to get more capabilities to recognize Al-generated text. The framework that has been proposed integrates the functions of TFIDF with 21 linguistic indicators that are created manually and they encompass all the stylistic, syntactic and semantic characteristics and the results reveals that machine learning classifiers, especially LightGBM which is trained on engineered stylometric and statistical features, can be able to discriminate between human and AI-generated text with validation of F 1 of 0.93% using human texts and 0.94% using AI-generated texts and 93.4% percent accuracy. Despite the promising results, still there are several limitations determined. current feature-based detectors are not as strong as needed to be reliable when used as standalone in patrolling new AI models, but they are effective in well-known distributions. Dataset focused on particular genres which may affect generalizability of the results to other languages. To work on this in the future, it is highly suggested to expand dataset to cover diverse linguistic patterns implement deep learning models such as transformers to further improve effectiveness. develop model-agnostic frameworks, create large-scale adversarial benchmark datasets, and investigate real-time adaptive learning mechanisms to create more robust detection systems that will stand up to the constant generative AI evolution. also Working on multi-class classification systems is to be considered to not only identify AI-generated text, but to also determine the model that generated it, which would be useful in the context of forensics.

In conclusion the proposed model provides interpretable solution to digital forensic and cybersecurity application that has a possibility to increase its generalization ability and resilience to adversarial attacks in the future.

REFERENCES

[1] R. Sousa-Silva, "Fighting Cyber-malice: A Forensic Linguistics Approach to Detecting AI-generated Malicious Texts," [Online]. Available: https://www.weforum.org/agenda/2024/01/.

[2] A. Kayabas, A. E. Topcu, Y. I. Alzoubi, and M. Yildiz, "A Deep Learning Approach to Classify AI-Generated and Human-Written Texts," Applied Sciences (Switzerland), vol. 15, no. 10, May 2025, [Online]. Available: https://doi.org/10.3390/app15105541.

[3] B. Boutadjine, F. Harrag, and K. Shaalan, "Human vs. Machine: A Comparative Study on the Detection of AI-Generated Content," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 24, no. 2, Feb. 2025, [Online]. Available: https://doi.org/10.1145/3708889.

[4] L. Chang, "Detecting AI-Generated Text: A Comparative Study of Machine Learning Algorithms," J. Syst. Cybern. Inf., vol. 23, pp. 36-41, Dec. 2025, [Online]. Available: https://doi.org/10.54808/JSCI.23.06.36.

[5] G. P. Georgiou, "Differentiating between Human-Written and AI-Generated Texts Using Linguistic Features Automatically Extracted from an Online Computational Tool."

[6] S. Mitrović, D. Andreoletti, and O. Ayoub, "ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text," Jan. 2023, [Online]. Available: http://arxiv.org/abs/2301.13852.

[7] C. Opara, "Distinguishing AI-Generated and Human-Written Text Through Psycholinguistic Analysis," May 2025, [Online]. Available: http://arxiv.org/abs/2505.01800.

[8] K. Gryka, K. Gradon, M. Kozlowski, M. Kutyla, and A. Janicki, "Detection of AI-Generated Emails - A Case Study," in ACM International Conference Proceeding Series, Association for Computing Machinery, Jul. 2024, [Online]. Available: https://doi.org/10.1145/3664476.3670465.

[9] M. S. Ibrahim, J. A. Eleiwy, H. M. Muhi-Aldeen, Y. Al-Yasiri, and A. A. Nafea, "Human to Chatbot Text Classification Using Multi-Source AI Chatbots and Machine Learning Models," Journal of Intelligent Systems and Internet of Things, vol. 16, no. 1, pp. 152-165, 2025, [Online]. Available: https://doi.org/10.54216/JISIoT.160113.

[10] S. Parimanoharan and R. D. Nawarathna, "Assessing Classical Machine Learning and Transformer-based Approaches for Detecting AI-Generated Research Text," [Online]. Available: https://orcid.org/0000-0000-0000-0000.

[11] R. Ardeshirifar, "Comparing Hand-Crafted and Deep Learning Approaches for Detecting AI-Generated Text: Performance, Generalization, and Linguistic Insights," AI and Ethics, vol. 5, no. 4, pp. 4197-4209, Aug. 2025, [Online]. Available: https://doi.org/10.1007/s43681-025-00699-4.

[12] S. Fariello, G. Fenza, F. Forte, M. Gallo, and M. Marotta, "Distinguishing Human From Machine: A Review of Advances and Challenges in AI-Generated Text Detection," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 6-18, 2025, [Online]. Available: https://doi.org/10.9781/ijimai.2024.12.002.

[13] K. Teja Repaka, M. Anirudh Bondugula, and S. Sashaank Adibhatla, "Benchmarking Distributed Machine Learning Systems with Large Language Models on Human vs. LLM Text Corpus."

[14] G. A. Godghase, R. Agrawal, T. Obili, and M. Stamp, "Distinguishing Chatbot from Human," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.04647.

[15] A. M. Elkhatat, K. Elsaid, and S. Almeer, "Evaluating the Efficacy of AI Content Detection Tools in Differentiating Between Human and AI-Generated Text," International Journal for Educational Integrity, vol. 19, no. 1, Dec. 2023, [Online]. Available: https://doi.org/10.1007/s40979-023-00140-5.

[16] R. Khera, A. F. Pedroso, V. K. Keloth, H. Xu, G. S. Silva, and L. H. Schwamm, "Scientific Writing in the Era of Large Language Models: A Computational Analysis of AI-Versus Human-Created Content," Stroke, vol. 56, no. 10, pp. 3078-3083, Oct. 2025, [Online]. Available: https://doi.org/10.1161/STROKEAHA.125.051913.

[17] D. Valiaiev, "Detection of Machine-Generated Text: Literature Survey," 2023, [Online]. Available: https://doi.org/0000001.0000001.

[18] C. Maddugoda, "A Comprehensive Review: Detection Techniques for Human-Generated and AI-Generated Texts."

[19] P. Ranade, A. Piplai, S. Mittal, A. Joshi, and T. Finin, "Generating Fake Cyber Threat Intelligence Using Transformer-Based Models," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2102.04351.

[20] T. Kumarage, J. Garland, A. Bhattacharjee, K. Trapeznikov, S. Ruston, and H. Liu, "Stylometric Detection of AI-Generated Text in Twitter Timelines," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.03697.

[21] Z. Zeng, L. Sha, Y. Li, K. Yang, G. Gašević, and G. Chen, "Towards Automatic Boundary Detection for Human-AI Collaborative Hybrid Essay in Education," 2024, [Online]. Available: https://www.kaggle.com/c/asap-aes.

[22] Z. Grinberg, "Human vs. LLM Text Corpus," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/starblasters8/human-vs-llm-text-corpus.

[23] G. P. Georgiou, "Differentiating Between Human-Written and AI-Generated Texts Using Automatically Extracted Linguistic Features," Information (Switzerland), vol. 16, no. 11, Nov. 2025, [Online]. Available: https://doi.org/10.3390/info16110979.

APPENDIX

Table A.1 Mathematical formulation of preprocessing.

Stage	Mathematical transformation	Forensic purpose
1. Noise Removal	`$\Phi 1(T) = T - \{r \in R \text{ r} \subset \text{T}\}$ `	(Where R is set of URL/Email regex patterns).
2. Normalization	$\Phi 2(T) = \{lower(c)\ \forall c \in T\}$	Standardizes character case.
3. Redundancy Handling	$\Phi 3(T) = compress(T,2)$ (Reduces consecutive chars > 2)	Mitigates impact of typos or excessive expressive elongation.
4. Emoji Handling	$\Phi 4(T) = \{demojize(e)\ \forall e \in E \cap T\}$	Converts symbols to text to capture semantic context
5. Punctuation Norm	$\Phi 5(T) = normalize\_punct(T)$	Standardizes punctuation marks for statistical analysis
6.Tokenization	$\Phi 6\ (T) = split\ (T, delimiter="\s")$	Decomposes text into discrete units (tokens) to enable granular feature analysis
7. Whitespace Correction	$\Phi 8(T) = trim(replace(T,"\s+"," "))$	Ensures consistent tokenization boundaries

A. Data Preprocessing Stages and Mathematical Transformations

This section provides mathematical formalization of the preprocessing phase as Each text document T undergoes sequence of transformations {Φ_1, Φ_2, ..., Φ_}. Where each Φi represents specific processing stage and the preprocessing Equation Pipeline as follows:

$$T\_cleaned = \Phi 8(\Phi 7(...\Phi 1(T))).$$

B. Feature Engineering (21 Features) Compatible with Cleaned Data

The following code segment illustrates part of the practical implementation of extracting the 21 features from clean processed text. Which it's an essential phase to ensure the model can effectively differentiating between human generated texts and AI-generated texts.

Appendix Conclusions

The appendix has explored part of the core of computations that have been executed in this research data cleaning, feature extraction. This script maintains high standards of transparency forming reproducible framework upon the final results were achieved.

```
# STEP 5: FEATURE ENGINEERING (21 Featu
res)— COMPATIBLE WITH CLEANED DATA
print("\n" + "="*100)
print("STEP 5: Feature Engineering (21
Features) — Optimized & Compatible")
print("="*100 + "\n")
import numpy as np
import pandas as pd
import string
from collections import Counter
from nltk import sent_tokenize, word_to
kenize, pos_tag
from textblob import TextBlob

# FEATURE EXTRACTION FUNCTION
def extract_all_features(text):
    if not isinstance(text, str):
        text = ""
    # Sentence and word tokenization
    sentences = sent_tokenize(text)
    words = [w for w in word_tokenize(t
ext.lower()) if w.isalpha()]
    features = {}
    # ========================
    # 1. Stylistic Features (11)
    features['num_sentences'] = len(sen
tences)
    features['avg_sentence_length'] = n
p.mean([len(s.split()) for s in sentenc
es]) if sentences else 0
    features['num_words'] = len(words)
    features['avg_word_length'] = np.me
an([len(w) for w in words]) if words el
se 0
    features['num_unique_words'] = len(
set(words))
    features['lexical_richness'] = len(
set(words)) / len(words) if words else
0
    features['num_punctuation'] = sum(1
for c in text if c in string.punctuatio
n)
    features['punctuation_ratio'] = fea
tures['num_punctuation'] / max(len(text
), 1)
    features['num_questions'] = text.co
unt('?')
```

```
    features['num_exclamations'] = text
.count('!')
    features['capital_ratio'] = sum(1 f
or c in text if c.isupper()) / max(len(
    # ========================
    # 2. Syntactic Features (5)
    try:
        pos_tags = pos_tag(words)
        tag_counts = Counter(tag for _,
tag in pos_tags)
        total = max(len(pos_tags), 1)

        features['noun_ratio'] = sum(ta
g_counts.get(t, 0) for t in ['NN', 'NNS
', 'NNP', 'NNPS']) / total

g_counts.get(t, 0) for t in ['VB', 'VBD
', 'VBG', 'VBN', 'VBP', 'VBZ']) / total
        features['adj_ratio'] = sum(tag
_counts.get(t, 0) for t in ['JJ', 'JJR'
, 'JJS']) / total
        features['adv_ratio'] = sum(tag
_counts.get(t, 0) for t in ['RB', 'RBR'
, 'RBS']) / total
        features['pronoun_ratio'] = sum
(tag_counts.get(t, 0) for t in ['PRP',
'PRP$']) / total
    except:
        features.update({k: 0 for k in
['noun_ratio', 'verb_ratio', 'adj_ratio
', 'adv_ratio', 'pronoun_ratio']})
    # 3. Semantic Features (2)
    # ========================
    try:
        blob = TextBlob(text)
        features['sentiment_polarity']
= blob.sentiment.polarity
        features['sentiment_subjectivit
y'] = blob.sentiment.subjectivity
    except:
        features['sentiment_polarity']
= 0
        features['sentiment_subjectivit
y'] = 0
    # ========================
    # 4. Writing Style Features (3)
    text_lower = text.lower()
    features['num_transitions'] = sum(t
ext_lower.count(w) for w in ['however',
'moreover', 'furthermore', 'therefore',
'thus'])
    features['num_formal_words'] = sum(
text_lower.count(w) for w in ['utilize'
, 'demonstrate', 'indicate', 'significa
nt', 'various'])
    features['first_person_count'] = su
m(text_lower.count(f' {w} ') for w in [
' i ', ' me ', ' my ', ' we ', ' us ',
' our '])
    return features
# ------------------------------------
```