ISSN:3108-253X
Special Issue : February,2026

RAAVI

An International Journal of Multidisciplinary Research

A Stylometric and Machine Learning Approach to Detecting AI-Generated Text

Ms. Pandav Nikita Harish
Assistant Professor in Computer Science Vishwasrao Ransing College
Kalamb–Walchandnagar Tal. Indapur 413114
Affiliated To Savitribai Phule Pune University, Pune 413102, MS (India) Dist. Pune, Maharashtra, India
Email :**nikitapandav063@gmail.com**

Abstract

Currently, AI tools can generate human-like text (Brown et al., 2020). This creates problems in educational writing, online content, and misinformation (Solaiman et al., 2019). It is difficult to identify whether a text is written by a human or generated by AI (Ippolito et al., 2020). In this study, a simple machine learning approach was used to detect AI-generated text. I built my own dataset of 200 paragraphs—100 written by humans and 100 generated by AI. Stylometric features such as sentence length, word length, vocabulary richness, punctuation count, and word count were extracted from the text (Koppel et al., 2011). Logistic Regression, Random Forest, and Support Vector Machine models were trained and tested (Manning et al., 2008). Logistic Regression achieved perfect accuracy on the test set, whereas Random Forest and SVM both reached 97.5% accuracy. What makes this research unique is that I collected and labeled all the data myself, which means the findings are totally original. The results show that machine learning can successfully distinguish AI-generated text from human-written content with high precision (Ippolito et al., 2020). This study proves that even simple statistical features can help detect AI text (Aggarwal and Zhai, 2012).

Keywords

AI-generated text detection, Stylometric analysis, Machine Learning, Text classification, Logistic Regression, Support Vector Machine, Random Forest, Academic integrity.

Introduction

Current improvements in artificial intelligence have led to the development of powerful text-generation systems capable of producing human-like writing (Brown et al., 2020). These

Relish Publications,
Amrapali Apartment, Canal Road, Baramati (MS), India.
Pin code- 413102

Website: www.relishpublications.com
Email: raavipublications@gmail.com
WhatsApp: 9665841996

models generate grammatically correct and contextually meaningful content, making manual identification gradually more difficult (Jawahar et al., 2019).

Many students and content creators use AI tools to generate assignments and articles. Therefore, it is important to detect whether a paragraph is written by a human or AI (Solaiman et al., 2019).The aim of this research was straightforward: to build a detection system using stylometric analysis and traditional machine learning algorithms and see how well it performs. Stylometry appealed to me because it focuses on measurable patterns in writing style such as how long sentences tend to be, how varied the vocabulary is, and how punctuation is used (Koppel et al., 2011). These patterns occur naturally when humans write, and I suspected they might look different when machines generate text (Uchida, 2021) By using a dataset I collected myself, I could ensure that the results were truly my own and that other researchers could reproduce them if they wanted to.

Literature Review

The detection of AI-generated text become an important research topic due to the fast development of large language models (Brown et al., 2020). Many researchers studied different techniques to identify whether a text is written by a human or AI (Ippolito et al., 2020).

Stylometry and Authorship

Stylometry is a method that uses to analyse a person's writing style. Researchers used this to identify authors, even if they tried to stay unknown (Koppel et al., 2011). Everyone has their own unique writing habits. Some people prefer short sentences, while others like long ones. Some repeat words often, while others use a large vocabulary. Everyone write differently(Koppel et al., 2011).

More recent studies applied these techniques to automated writing. Because AI generates text using patterns rather than thought, its writing looks mathematically different from human writing(Uchida, 2021).AI systems produce more uniform sentence structures and more predictable word choices (Ippolito et al., 2020). This means the same statistical methods that identify authors and also identify whether text came from a human or a machine.

Relish Publications,
Amrapali Apartment, Canal
Road, Baramati (MS), India.
Pin code- 413102

Website: www.relishpublications.com

Email: raavipublications@gmail.com

WhatsApp: 9665841996

RAAVI
An International Journal of Multidisciplinary Research
ISSN:3108-253X
Special Issue : February,2026

Approaches to AI Text Detection

As AI writing tools have become more common, researchers have developed various methods for detecting their outputs (Ippolito et al., 2020). One popular method uses large neural networks trained on huge datasets of human and AI text (Brown et al.,2020). These systems can learn complex patterns,but they require significant computational resources and large amounts of training data.

Another approach, called Giant Language Model Test Room, was developed by Gehrmann and colleagues (2019). Some researchers have taken a simpler approach using traditional machine learning algorithms with handcrafted features (Aggarwal and Zhai, 2012).

What Makes AI Writing Different

Some characteristics tend to distinguish AI-generated text from human writing. AI systems often produce more structurally unchanging texts, with sentences of similar lengths and predictable paragraph organization (Uchida, 2021). The vocabulary tends to be more formal and less various than used naturally (Solaiman et al., 2019). AI also tends to use certain function words and punctuation patterns differently from humans (Koppel et al., 2011).

Researchers have found that machine-generated text shows less burstiness than human writing (Uchida, 2021). Burstiness refers to the tendency of humans to use words in clusters, writing a word several times in quick succession before moving on. Human writing shows greater variation in predictability.

Machine Learning for Text Classifications

Machine learning has been successfully applied to many text classification problems, from spam detection to sentiment analysis (Manning et al., 2008). Traditional algorithms such as Logistic Regression, Support Vector Machines and random forest have proven effective for a wide range of tasks (Manning et al., 2008). The quality of the features used for training is what makes these algorithms work well (Aggarwal and Zhai, 2012).

Feature engineering is critical for text classification. The features must capture meaningful differences between classes while being computable from the available text (Aggarwal and Zhai, 2012).

Relish Publications,
Amrapali Apartment, Canal
Road, Baramati (MS), India.
Pin code- 413102

Website: www.relishpublications.com

Email: raavipublications@gmail.com

WhatsApp: 9665841996

ISSN:3108-253X
Special Issue : February,2026
RAAVI
An International Journal of Multidisciplinary Research

What This Research Contributes

Previous studies have made valuable contributions; however, gaps remain. Many rely on large datasets and complex architectures that are not accessible to all researchers (Brown et al., 2020).

I wanted to show that you don't need a supercomputer or advanced AI to detect text effectively. I used a simple, straightforward method with my own data to prove that anyone can get great results. (Aggarwal and Zhai, 2012).

Objectives of the Study

The objectives of research are:

1. To analyse the stylometric differences between AI-generated and human-written texts.
2. To extract statistical writing features for classification.
3. To implement and compare multiple machine-learning models.
4. The detection performance was evaluated using standard metrics.

Methodology

This section describes how the research was conducted, from building the dataset to training and evaluation.

Building the Dataset

The dataset is the foundation of this research; therefore, I paid careful attention to creating a balanced, representative, and original dataset. I collected 200 paragraphs in total, with 100 written by humans and 100 generated by AI.On Different topics were used , such as: Climate change, Artificial intelligence, Education, Healthcare, Technology.I stored the final dataset in CSV format with two columns: text containing the paragraph contentand label indicating whether it was "Human" or "AI" generated. This simple format makes it easy to load and process data using standard tools. Table summarizes the key characteristics of the dataset.

Relish Publications,
Amrapali Apartment, Canal
Road, Baramati (MS), India.
Pin code- 413102

Website: www.relishpublications.com

Email: raavipublications@gmail.com

WhatsApp: 9665841996

ISSN:3108-253X
Special Issue : February,2026
RAAVI
An International Journal of Multidisciplinary Research

Dataset Summary

Characteristic	Value
Total Samples	200
Human-Written Samples	100
AI-Generated Samples	100
Average Paragraph Length	100-200 words
Topics Covered	5
Data Format	CSV

Extracting Stylometric Features

Because machine learning models cannot directly understand raw text, the text was converted into numerical features. The following features were extracted:

- Average sentence length
- Average word length
- Vocabulary richness (unique words / total words)
- Number of punctuation marks
- Total word count

These features represent writing style and patterns.I implemented feature extraction using Python's NLTK library, which provides reliable tools for text processing. Each paragraph was processed to calculate all five features, producing a numerical feature matrix ready for machine learning.

Preparing Data for Training

Before training the models, the data were preprocessed and divided into training and testing sets.The data is an 80-20 ratio, with 80% for training and 20% for testing. Stratified sampling ensured that both subsets maintained the same class balance as the original dataset. This resulted in 160 samples for training (80 human, 80 AI) and 40 samples for testing (20 human, 20 AI).

Relish Publications,
Amrapali Apartment, Canal
Road, Baramati (MS), India.
Pin code- 413102

Website: www.relishpublications.com

Email: raavipublications@gmail.com

WhatsApp: 9665841996

ISSN:3108-253X
Special Issue : February,2026
RAAVI
An International Journal of Multidisciplinary Research

Machine Learning Models

Three supervised learning algorithms widely used for classification tasks were implemented.

Logistic regression is a machine learning tool used to categorize data into one of two choices(like Yes/No or True/False).Instead of just guessing, it calculates the probability that an item belongs to a specific category.It is simple, interpretable, and computationally efficient (Hosmer et al., 2013). The model learns a decision boundary that separates human-written from AI-generated text based on the stylometric features.Random Forest is an ensemble method that builds many decision trees during training and outputs the class that is the majority vote of all trees. It handles nonlinear relationships well and is resistant to overfitting (Breiman, 2001).

support vector machine finds the optimal hyperplane that separates samples of different classes. The RBF kernel allows it to handle nonlinear decision restrictions (Cortes and Vapnik, 1995). SVM is mostly effective in high-dimensional spaces, which can be useful for classification.All models were applied using scikit-learn in Python, with the default hyperparameters used for the initial experiments.

Model Evaluation

The performance was evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score

Implementation

The implementation was performed in Python using:

- Pandas for data handling
- NLTK for text processing
- Scikit-learn for machine learning models

Relish Publications,
Amrapali Apartment, Canal
Road, Baramati (MS), India.
Pin code- 413102

Website: www.relishpublications.com

Email: raavipublications@gmail.com

WhatsApp: 9665841996

ISSN:3108-253X
Special Issue : February,2026
RAAVI
An International Journal of Multidisciplinary Research

The steps included:

- Loading CSV file
- Extracting stylometric features
- Training models
- Testing models
- Calculating accuracy

Experimental Results

Cross-Validation Analysis

Five-fold cross-validation was performed to check the stability of the models. The accuracy values obtained for the different folds were:0.975, 1.000, 1.000, 1.000, and 0.974.

The average cross-validation accuracy was **98.98%**, indicating that the model performed consistently across different data splits and did not depend on a single training–testing combination.

Model Performance Comparison

The overall accuracies of the classifiers are summarized below.

Table 1: Accuracy of Machine Learning Models

Model	**Accuracy**
Logistic Regression	100%
Random Forest	97.5%
Support Vector Machine	97.5%

Among the three models, Logistic Regression achieved the highest accuracy on the test dataset. Random Forest and Support Vector Machine also produced strong and almost similar results.

Relish Publications,
Amrapali Apartment, Canal
Road, Baramati (MS), India.
Pin code- 413102

Website: www.relishpublications.com

Email: raavipublications@gmail.com

WhatsApp: 9665841996

ISSN:3108-253X
Special Issue : February,2026
RAAVI
An International Journal of Multidisciplinary Research

Detailed Evaluation of Random Forest Model

To understand the model's accuracy, I calculated the precision, recall, F1-score for the Random Forest classifier.

Table 2: Classification Report for Random Forest

Class	Precision	Recall	F1-Score	Support
AI	0.95	1.00	0.97	19
Human	1.00	0.95	0.98	21
Overall Accuracy			**97%**	40

The data proves the model successfully told the difference between human and AI content. High precision and recall values suggest that the classifier makes few incorrect predictions.

Result Interpretation

The experimental results confirm that stylometric features contain useful information for distinguishing AI-generated texts from human-written content. Even without deep learning techniques, traditional machine learning algorithms can achieve high precision. The model performed well in testing showing it is stable. But since a small dataset is used it is needed to test it with more varied data to be sure it works just as well in the real world.

Conclusion:

The experimental results showed strong performance across all three algorithms. Logistic Regression achieved perfect classification accuracy on the test dataset, while Random Forest and Support Vector Machine also produced high accuracy values of 97.5%. The cross-validation average accuracy of 98.98% further indicates that the models performed consistently across different data splits.

While these initial results look good, my small sample size means I need broader testing. Using a wider variety of data will help prove that the method works reliably. In the future, I also plan to add deeper language analysis (semantic features) so the system is not easily fooled by more advanced AI models.

Relish Publications,
Amrapali Apartment, Canal
Road, Baramati (MS), India.
Pin code- 413102

Website: www.relishpublications.com

Email: raavipublications@gmail.com

WhatsApp: 9665841996

Overall, this study confirms that an efficient and computationally simple framework can effectively detect AI-generated academic text using stylometric analysis.

References

[1] T. Koppel, J. Schler and S. Argamon,"Authorship Attribution in the Wild,*Language Resources and Evaluation*, vol. 45, no. 1, pp. 83–94, 2011.

[2] A. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, "Automatic Detection of Generated Text is Easiest When Humans are Fooled," *Proceedings of ACL*, 2020.

[3] E. Jawahar, M. Sagot, and D. Seddah,"What Does BERT Learn About the Structure of Language?" *Proceedings of ACL*, 2019.

[4] S. Gehrmann, H. Strobelt, and A. Rush, "GLTR: Statistical Detection and Visualization of Generated Text," *Proceedings of ACL System Demonstrations*, 2019.

[5] I. Goodfellow et al., "Generative Adversarial Nets," *NeurIPS*, 2014.

[6] A.Solaiman et al.,"Release Strategies and the Social Impacts of Language Models," 2019.

[7] S. Uchida, "Statistical Characteristics of Machine-Generated Text," 2021.

[8] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[9] C. C. Aggarwal and C. Zhai, *Mining Text Data*, Springer, 2012.

[10] T. Brown et al., "Language Models are Few-Shot Learners," *NeurIPS*, 2020.

❑❑❑

Relish Publications,
Amrapali Apartment, Canal
Road, Baramati (MS), India.
Pin code- 413102

Website: www.relishpublications.com

Email: raavipublications@gmail.com

WhatsApp: 9665841996