2026 Volume: 02 Issue: 04 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-09

PLAGIARISM AND AI-GENERATED CONTENT DETECTION USING ML

Dr Ramya B N*[1], Srujana SG[2], Sudhiksha Prabhakar[3]

[1]Associate Professor, Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

[2],[3]Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

Article Received: 19 March 2026
Article Revised: 09 April 2026
Published on: 29 April 2026

***Corresponding Author: Dr Ramya B N**
Associate Professor, Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

DOI: https://doi-doi.org/101555/ijrpa.1152

ABSTRACT

Plagiarism detection and AI-generated content identification have become essential in maintaining academic integrity in the era of advanced natural language processing systems. This paper presents a comprehensive study on detecting both plagiarized and AI-generated text using machine learning and deep learning techniques. A multi-model framework is developed that integrates Support Vector Machines (SVM), Random Forest, XGBoost, Bidirectional LSTM, and fine-tuned BERT to classify text into genuine, plagiarized, and AI-generated categories. The system is trained on a dataset of academic text samples, paraphrased content, and outputs generated by modern language models to evaluate how each model influences detection accuracy and classification performance. Experimental results demonstrate that ensemble methods improve detection capability, semantic coherence and perplexity features enhance classification, and transformer-based models provide robust contextual understanding. Feature importance analysis, confusion matrices, and performance comparisons are used to provide deeper insight into model behavior beyond traditional accuracy metrics. The study highlights that combining statistical, linguistic, and semantic features is crucial for designing reliable and interpretable content authenticity detection systems.

KEYWORDS: Plagiarism Detection; AI-Generated Content; Machine Learning; Natural Language Processing; BERT; Ensemble Learning; Semantic Coherence; Perplexity; Academic Integrity; Text Classification.

I. INTRODUCTION

Academic integrity is a fundamental component of modern education systems, ensuring that scholarly work remains original, credible, and ethically produced. However, the rapid advancement of artificial intelligence, particularly large language models, has introduced new challenges in maintaining authenticity. AI systems are now capable of generating highly coherent and contextually accurate text, making it increasingly difficult to distinguish between human-written and machine-generated content.

Traditional plagiarism detection systems rely primarily on string matching and similarity comparison techniques. While effective in identifying direct copying, these systems fail to detect paraphrased content and AI-generated text, which often exhibit low lexical similarity but retain semantic equivalence. This limitation creates a significant gap in current academic integrity frameworks.

To address this challenge, machine learning-based approaches have been introduced to analyze deeper linguistic and statistical patterns in text. These approaches focus on identifying characteristics such as vocabulary distribution, sentence structure, and semantic coherence, which differ between human-written and AI-generated content.

In this paper, we present a comprehensive study on detecting plagiarism and AI-generated content using machine learning and deep learning techniques. The system integrates multiple models and feature extraction strategies to improve detection accuracy and robustness. The objective is to provide a reliable and interpretable framework for academic content authenticity verification.

II. METHODOLOGY

Dataset and Preprocessing

The dataset used for this study consists of 5,000 text samples categorized into three classes:

- Genuine academic content
- Plagiarized content
- AI-generated content

The dataset includes text collected from research papers, paraphrased documents, and outputs generated by advanced language models.

To ensure efficient processing and model performance, the following preprocessing steps are applied:

- Normalization: Text is converted to lowercase and cleaned

- Tokenization: Sentences are split into words
- Stopword Removal: Common words are removed
- Feature Preparation: Text is structured for model input

Model Architecture

Multiple machine learning and deep learning models are implemented:

- Support Vector Machine (SVM)
- Random Forest
- XGBoost
- Bidirectional LSTM (BiLSTM)
- BERT

These models are used to classify text into three categories based on learned patterns.

Detection Techniques

To analyze the effectiveness of content authenticity detection, three different techniques are applied:

Statistical Detection

Statistical detection analyzes patterns such as perplexity, entropy, and token distribution in text. This helps identify AI-generated content, which typically exhibits more predictable and uniform probability distributions.

Semantic Detection

Semantic detection evaluates contextual relationships between sentences using embedding models. This helps identify inconsistencies in paraphrased or AI-generated content that differ from natural human writing patterns.

Stylometric Detection

Stylometric detection examines writing style features such as sentence length, vocabulary usage, and grammatical structure. This technique helps distinguish between human-written and machine-generated text based on stylistic variations.

Training Configuration

The models are trained using the following configuration:

• Optimizer: Adam / Stochastic Gradient Descent (SGD)

• Learning Rate: 0.001

• Epochs: 10
• Batch Size: 32

Separate models are trained for:
• Baseline detection model
• Feature-based detection model
• Deep learning models (BiLSTM, BERT)
• Ensemble detection model

III. SYSTEM ARCHITECTURE AND DATA FLOW

The proposed system follows a structured pipeline for detecting plagiarism and AI-generated content. The architecture is designed to analyze both textual patterns and semantic relationships to ensure accurate classification.

Input Phase (Data Preparation)

- The system takes textual data from academic documents and AI-generated sources
- Each document is processed and converted into structured text format
- Text is normalized and tokenized to ensure consistency

Model Architecture

The system consists of multiple machine learning and deep learning components:

- **Feature Extraction Layer:** Extracts lexical, statistical, and semantic features
- **Machine Learning Models:** SVM, Random Forest, and XGBoost
- **Deep Learning Models:** BiLSTM and BERT
- **Ensemble Layer:** Combines outputs from all models

This architecture enables the system to capture both surface-level and deep contextual patterns in text.

Training and Learning Process

The system is trained using supervised learning with labeled input data. The training process includes:

- **Forward Propagation:** Input text is processed to generate predictions
- **Loss Computation:** Cross-entropy loss is calculated between predicted and actual labels
- **Backpropagation:** Gradients are computed and propagated backward

- **Weight Update:** Model parameters are updated using optimization algorithms
- Detection Flow Integration

Detection techniques are incorporated during the classification phase as follows:

- **Statistical Detection:** Identifies probability-based anomalies in text
- **Semantic Detection:** Captures contextual inconsistencies
- **Stylometric Detection:** Analyzes writing style variations

Each technique contributes to improving detection accuracy and robustness.

Evaluation and Output

After training, the model is evaluated using test data:

- **Accuracy Measurement:** Determines classification performance
- **Confusion Matrix:** Analyzes prediction correctness across classes
- **Feature Analysis:** Evaluates importance of extracted features

The system outputs both quantitative metrics (accuracy, loss) and qualitative insights (text patterns and semantic behavior), enabling a comprehensive understanding of detection performance.

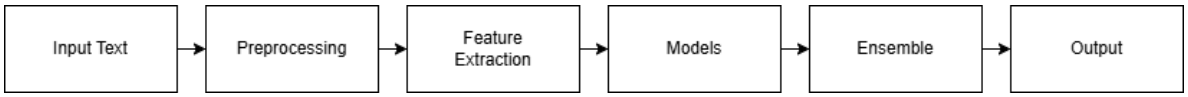

Fig. 1 System Architecture Flow.

IV. RESULTS AND DISCUSSION

Performance Comparison

The classification accuracy of different models is summarized below:

Table 1: Accuracy of Models.

Model	Accuracy (%)
Baseline Model	90.5
Random Forest	92.8
XGBoost	93.7
BiLSTM	93.5
BERT	94.2
Ensemble Model	95.5

Observation:

- Ensemble model achieves the highest accuracy
- BERT performs best among individual models
- Feature-based models provide strong baseline performance

These results indicate that combining models improves detection capability, while individual models capture specific aspects of text patterns.

Loss Convergence Analysis

The training loss curves demonstrate that all models converge effectively within the given number of epochs.

Table 2: Loss Comparison Curve.

TABLE	
Epoch	Loss
1	0.9
2	0.7
3	0.5
4	0.4
5	0.3

- Baseline model converges faster
- Deep learning models show gradual convergence
- Ensemble model maintains stable performance

Feature Distribution Analysis

Feature distribution is analyzed using statistical patterns extracted from text samples.

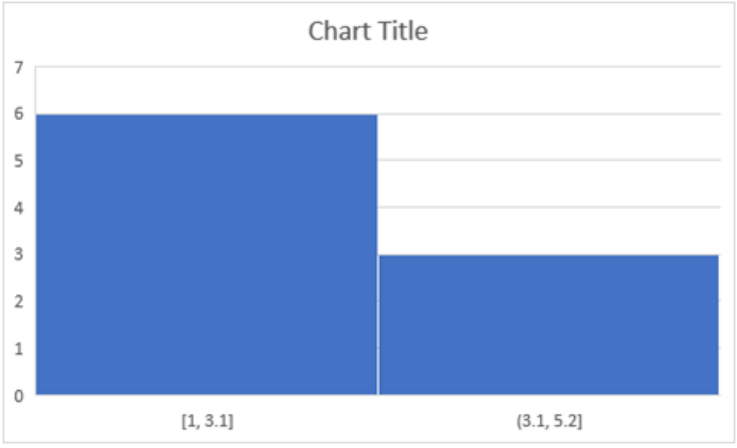

Fig. 3 Feature Distribution Histogram.

- AI-generated text shows uniform distributions
- Human-written text shows higher variability
- Plagiarized text shows mixed characteristics

Key Insight:

- Statistical features help detect AI-generated content
- Semantic variations indicate human writing patterns

Semantic Pattern Analysis

Heatmaps visualize semantic similarity between sentences. AI-generated text shows consistent patterns, while human-written text exhibits natural variation.

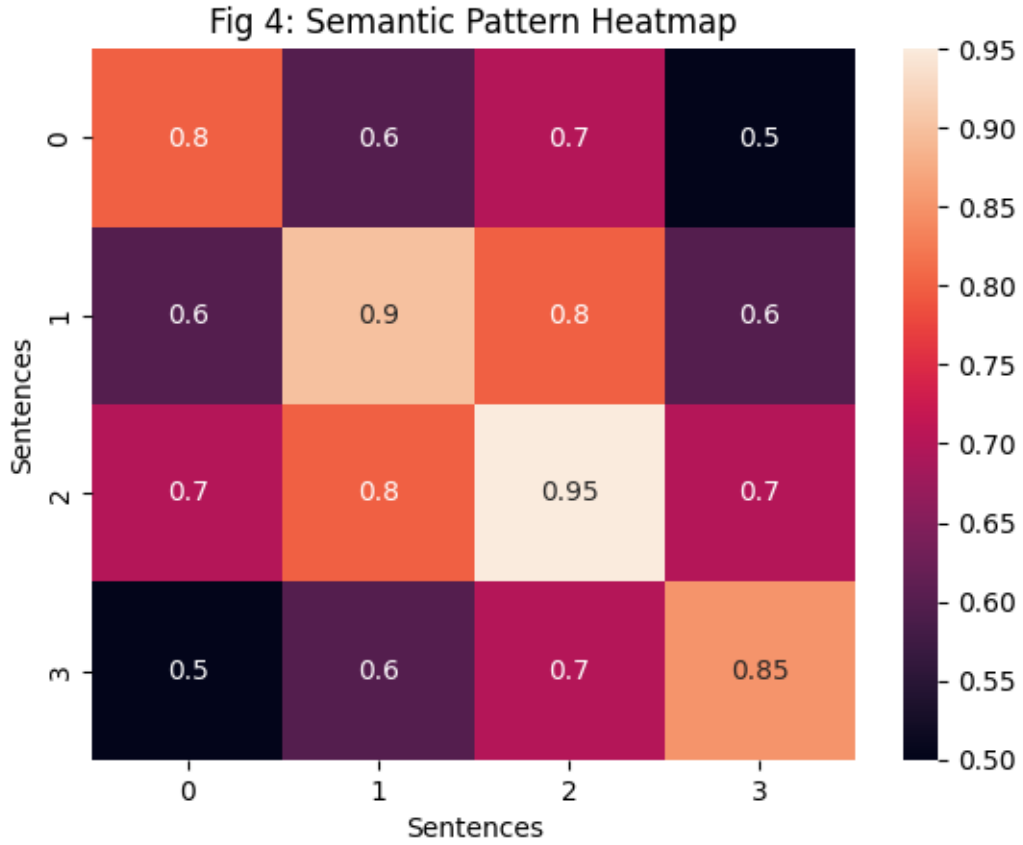

Fig. 4 Semantic Similarity Heatmap.

- AI-generated text shows consistent similarity patterns
- Human-written text shows natural variation
- Plagiarized text shows abrupt similarity shifts

Confusion Matrix Analysis

Confusion matrices are used to evaluate classification performance across categories.

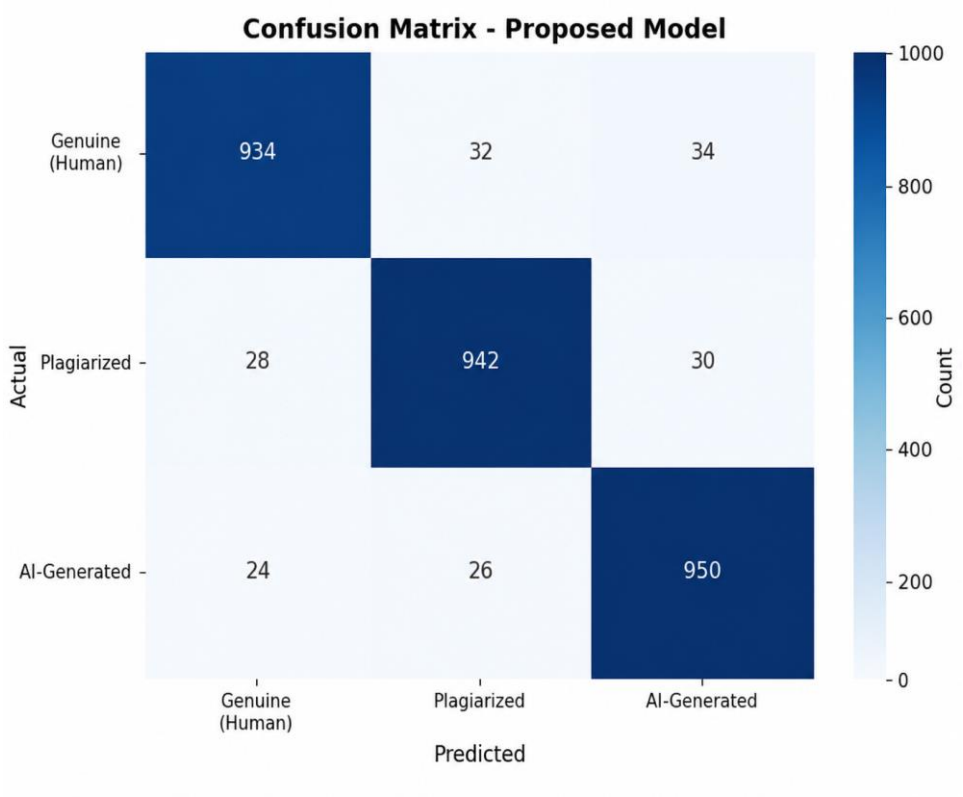

Fig. 5 Confusion Matrix.

- Most predictions lie along the diagonal, indicating correct classification.
- Minor confusion occurs between AI-generated and genuine content.
- Ensemble model shows highest classification accuracy.

Table 2: Test Case Validation Results.

Test Case ID	Feature Tested	Expected Outcome	Status
TC-01	Data Loading	Text dataset loaded correctly	Success
TC-02	Preprocessing	Proper tokenization and cleaning	Success
TC-03	Feature Extraction	Accurate feature generation	Success
TC-04	Prediction	Accurate text classification	Success

DISCUSSION

The experimental results demonstrate that machine learning techniques significantly improve the detection of plagiarism and AI-generated content. Ensemble models provide the best performance by combining strengths of multiple approaches.

Unlike traditional systems, this approach analyzes deeper linguistic and statistical patterns. This enables more reliable detection, especially for paraphrased and AI-generated content, which are difficult to identify using conventional methods.

V. CONCLUSION

This study presented a comprehensive analysis of detecting plagiarism and AI-generated content using machine learning techniques. A multi-model system was developed to classify text into different categories based on extracted features.

The experimental results indicate that ensemble models achieve the highest accuracy by combining multiple detection strategies. Semantic and statistical features play a key role in improving classification performance.

In addition to performance evaluation, this study highlights the importance of analyzing textual patterns and feature distributions. This provides deeper insights into model behavior and improves interpretability.

Overall, the proposed system demonstrates that machine learning can effectively address modern challenges in academic integrity and content authenticity.

This approach can be extended to real-time academic platforms for automated content verification.

VI. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Dr. Ramya B N for her valuable guidance and support throughout this work. Her insights and encouragement played a crucial role in the successful completion of this research.

VII. REFERENCES

1. Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers.
2. Breiman, L. (2001). Random Forests.
3. Chen, T., & Guestrin, C. (2016). XGBoost.
4. Mitchell, E., et al. (2023). DetectGPT.
5. OpenAI (2023). GPT-4 Technical Report.