iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69630

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Hybrid NLP and Machine Learning Framework for Detecting Human and AI-Written Texts

Mrs. B. Haritha[1], G. Aswini[2], K. Suvarna Lakshmi[3], P. Sailaja[4], P. Aparna[5], I. Haritha[6]
[1]*MTech, Computer Science & Engineering, Bapatla Women's Engineering College, Bapatla, AP, India*
[2, 3, 4, 5, 6]*BTech, Computer Science & Engineering, Bapatla Women's Engineering College, Bapatla, AP, India*

Abstract: ***With the rise of powerful AI language models like GPT-4 and LLaMA, distinguishing between AI-generated and human-written text has become increasingly challenging. This project presents a detection system that utilizes Natural Language Processing (NLP) and Machine Learning (ML) to identify AI-generated content. It integrates deep BERT embeddings with carefully crafted linguistic features such as perplexity, sentence structure, sentiment, and word usage. These features train two classifiers -XGBoost and Support Vector Machine (SVM)—which are combined into an ensemble model for enhanced accuracy. Trained on a balanced dataset of AI and human-written texts, the ensemble model achieved up to 93% accuracy, while XGBoost and SVM individually attained 84% and 81%, respectively. The system also includes a user-friendly interface for real-time text analysis and generates an HTML report detailing predictions and confidence scores. This solution provides an effective tool for educators, researchers, and institutions to detect AI-generated text and promote the ethical use of AI technologies.***
Keywords: AI-generated text, BERT, NLP, Machine Learning, SVM, XGBoost, Text Classification, Linguistic Features.

I. INTRODUCTION

The rapid advancement of AI, particularly in NLP, has led to advanced language models like OpenAI's GPT-4, Meta's LLaMA, and Google's Gemini, which produce coherent, grammatically sound, and stylistically refined text that closely mimics human writing. While these innovations enhance automation and communication in fields like education, journalism, legal writing, and social media, they raise concerns about content authenticity and ethical use due to risks like misinformation, academic dishonesty, and identity fraud. Current detection tools, such as plagiarism software and rule-based systems, struggle to identify AI-generated text because of its originality and fluency. To address this, we propose a hybrid detection system that integrates BERT-derived semantic insights with carefully selected linguistic features, training XGBoost and SVM models in an ensemble approach to improve accuracy, complemented by an intuitive interface offering real-time predictions and confidence scores to promote transparency and responsible AI use.

II. RELATED WORK

Previous work in this area includes perplexity-based detection, stylometric analysis, and neural classifiers like DetectGPT. While these approaches show promise, they often lack interpretability or adaptability across models. Our approach combines traditional linguistic analysis with BERT embeddings and ensemble learning for enhanced robustness and explainability.

III. SYSTEM ARCHITECTURE

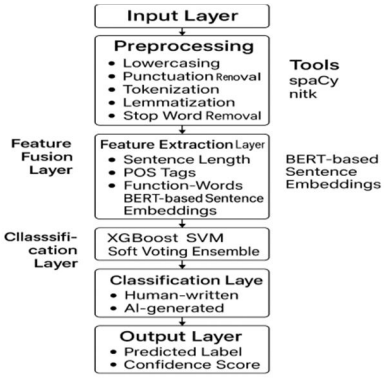


Fig1: Flow Chart

The system designed to detect AI-generated text combines several modules, beginning with text preparation, including lowercasing, word segmentation, and punctuation elimination, followed by analyzing linguistic traits such as sentence length, vocabulary diversity, sentiment, casual word usage, and pronoun frequency. Semantic features are extracted using the bert-base-uncased model and blended with linguistic data to create a cohesive feature array. Classification employs XGBoost and SVM models, integrated via a soft voting ensemble to enhance accuracy, producing a binary outcome (human-written or AI-created) with confidence scores. An HTML report tool provides visual insights and supports debugging for greater clarity and accountability.

IV. METHODOLOGY

A. *Feature Engineering*

The system's success hinges on its approach to feature engineering, which focuses on identifying clear and nuanced text traits that set human-written content apart from AI-generated output. Blending style-based and meaning-based features creates a clear and detailed picture of the text, supporting precise classification.

1) Linguistic Features: These features focus on syntactic, stylistic, and statistical traits that often vary between human and AI writing styles.

2) Custom Perplexity Estimation: A simplified, rule-based perplexity metric assesses how predictable a text sequence is. Since AI-generated content tends to follow highly probable language patterns, it typically results in lower perplexity scores.

3) Sentence Complexity and Variation: The system examines indicators like the typical length of sentences and the range of sentence structures, including how deeply words connect grammatically and the frequency of supporting clauses. Human writing often weaves a mix of sentence types, creating a natural flow, while AI-produced text can seem overly consistent or refined in its arrangement.

4) Word usage diversity: Word usage diversity reflects how varied a text's vocabulary is, helping our system distinguish human writing from AI-generated content. People tend to choose diverse words shaped by their unique style and situation, while AI often leans on familiar words or repeated phrases, even when sounding smooth. The system measures this with tools like the unique-to-total word ratio, which shows the proportion of distinct words, and a metric that gauges vocabulary repetition. These clues enable the system to detect patterns revealing whether a person or AI wrote the text

5) Sentiment polarity and subjectivity: Sentiment polarity and subjectivity highlight a text's emotional tone and degree of personal expression, helping the system separate human writing from AI-generated content. Polarity indicates whether the text feels positive, negative, or neutral; subjectivity shows if it leans toward opinions or facts. Human writing often bursts with authentic emotion or personal viewpoints, mirroring the author's purpose. By contrast, AI text may feel emotionally subdued or neutral, unless crafted differently. With tools like TextBlob or VADER, the system detects these emotional signals, or their absence, to identify whether a human or AI authored the text.

6) Informality and personal tone: Informality and personal tone act as vital hints in identifying whether a text is human-written or AI-generated. Humans often weave casual phrases, slang, contractions, or pronouns like "I", "we", or "my" into their writing to create a relaxed or expressive vibe, lending the text a warm, authentic feel. In contrast, AI-generated text usually comes across as more formal, polished, or consistent, unless instructed to sound conversational. The system counts how frequently casual words and personal pronouns appear to assess the text's tone. A high number of these features suggests human authorship, while their rarity typically indicates AI creation.

7) Semantic Features: Semantic features are crucial for capturing the deeper meaning of text beyond surface-level patterns. In this project, we extract semantic information using BERT (Bidirectional Encoder Representations from Transformers), specifically leveraging the CLS (classification) token embedding. This embedding summarizes the overall context of the input text by capturing the relationships between all words in the sentence. Unlike traditional word-level features, BERT embeddings provide a robust representation that understands grammar, context, and meaning across different parts of the text. By inputting these semantic vectors into the classifier, the system can more effectively detect subtle differences in how humans and AI generate meaningful content, thereby improving the accuracy of AI-generated text detection.

B. *Model Training*

The model training phase of this project involved building a balanced and diverse dataset consisting of 400 human-written texts and 400 AI-generated samples from models like GPT-4 and LLaMA. These texts covered various writing styles, including formal, informal, creative, and technical. The combined feature vectors, generated by fusing BERT embeddings with handcrafted linguistic features, were used to train two separate classifiers: XGBoost and Support Vector Machine (SVM).

Both models underwent hyperparameter tuning using GridSearchCV to optimize performance and prevent overfitting. To improve prediction reliability and accuracy, an ensemble approach was applied using soft voting, where the final decision was based on the average of the prediction probabilities from both models. This ensemble strategy effectively leveraged the strengths of both classifiers, resulting in more stable and accurate detection of AI-generated text.

V. RESULTS AND ANALYSIS

The system includes a user interface for real-time predictions. An HTML file is generated after each prediction session, displaying:

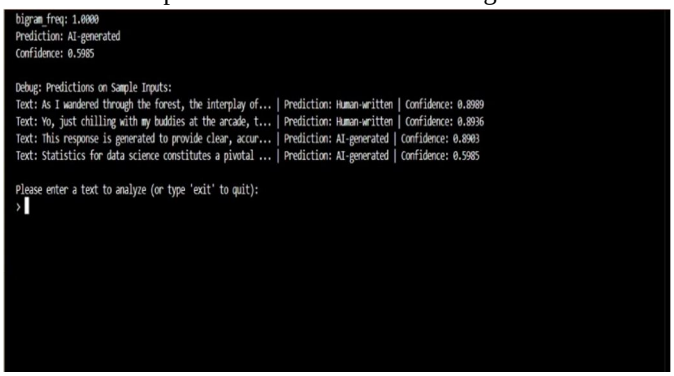


Fig.2: Input as text

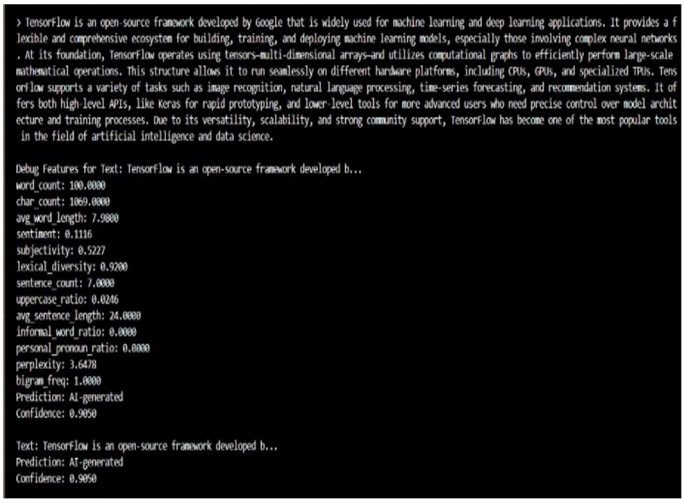


Fig.3: Text Result

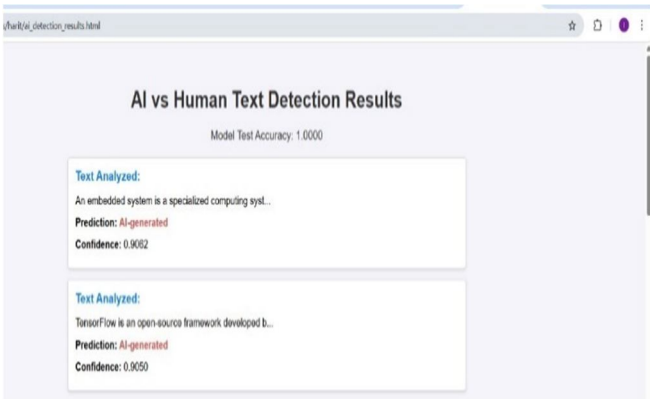


Fig.4: Result in Web Browser

Fig.5: Input as text

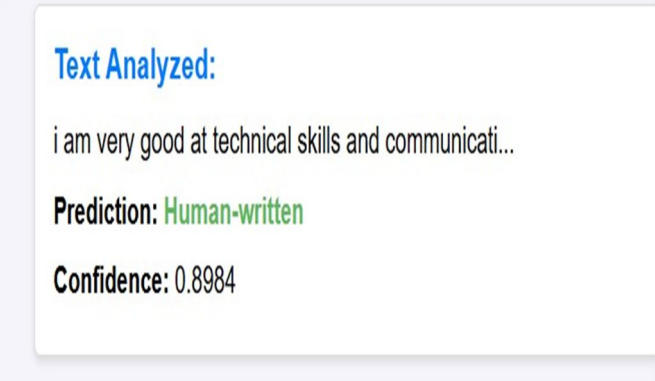


Fig.6: Result in Web Browser

The individual classifiers—XGBoost and SVM—demonstrated solid performance, with XGBoost slightly outperforming SVM in all metrics. However, the ensemble model showed a significant improvement in all areas. By combining the strengths of both classifiers using a soft-voting strategy, the ensemble achieved a high accuracy of 93%, indicating a strong capability to differentiate between AI-generated and human-written texts.

The improvement in performance through ensemble learning confirms the effectiveness of combining semantic and linguistic features. BERT embeddings provided deep contextual understanding, while handcrafted linguistic features captured style, tone, and structure elements often missed by AI. This dual-feature approach allowed the model to detect subtleties such as emotional expression, sentence variety, and vocabulary richness, which are typically more nuanced in human writing.

The system also includes a debugging module and HTML-based report generation, enabling users to view prediction results with confidence scores and analyze text features. This enhances interpretability and usability in real-world applications such as education, publishing, and digital content verification.

VI. CONCLUSION

This project effectively detects whether text is AI-generated or human-written. It uses BERT for meaning and linguistic features for style, spotting AI's polished text versus humans' varied, emotional writing. Combining XGBoost and SVM in an ensemble ensures accurate predictions, trained on 400 AI and 400 human texts. The model performs well and outputs results in an easy-to-read HTML report.

VII. FUTURE SCOPE

The system can be improved by adding multiple languages and coding programs. Including languages like Spanish or Hindi with models like mBERT will help detect AI text globally. Training on code (e.g., Python) from humans and AI will spot AI-generated programs by checking code patterns. This needs more data, new features, and updated XGBoost/SVM or CodeBERT models. These changes will make the system useful for worldwide text and coding applications.

REFERENCES

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019).BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL. https://arxiv.org/abs/1810.04805
[2] Solaiman, I., et al. (2019).Release Strategies and the Social Impacts of Language Models. OpenAI. https://arxiv.org/abs/1908.09203
[3] Zellers, R., Holtzman, A., Rashkin, H., et al. (2019). Defending Against Neural Fake News. NeurIPS. https://arxiv.org/abs/1905.12616
[4] Zhang, Y., et al. (2021).Detecting AI-Generated Text: A Survey. arXiv. https://arxiv.org/abs/2107.06499
[5] OpenAI (2023). GPT-4 Technical Report. https://openai.com/research/gpt-4
[6] Kreps, S., McCain, R. M., & Brundage, M. (2022). All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. Journal of Experimental Political Science. https://doi.org/10.1017/XPS.2021.29
[7] Shu, K., et al. (2022).Fake News Detection on Social Media: A Data Mining Perspective (Updated Review). SIGKDD Explorations.https://arxiv.org/abs/1708.01967
[8] Liu, J., et al. (2023).DetectGPT: Zero-Shot Detection of Machine-Generated Text via Probability Curvature. arXiv. https://arxiv.org/abs/2301.11305
[9] Weidinger, L., et al. (2021).Ethical and Social Risks of Harm from Language Models. arXiv. https://arxiv.org/abs/2112.04359

AUTHOR'S PROFILES

Mrs. B. Haritha is working as an Assistant Professor in the Department of CSE, BWEC, Bapatla.

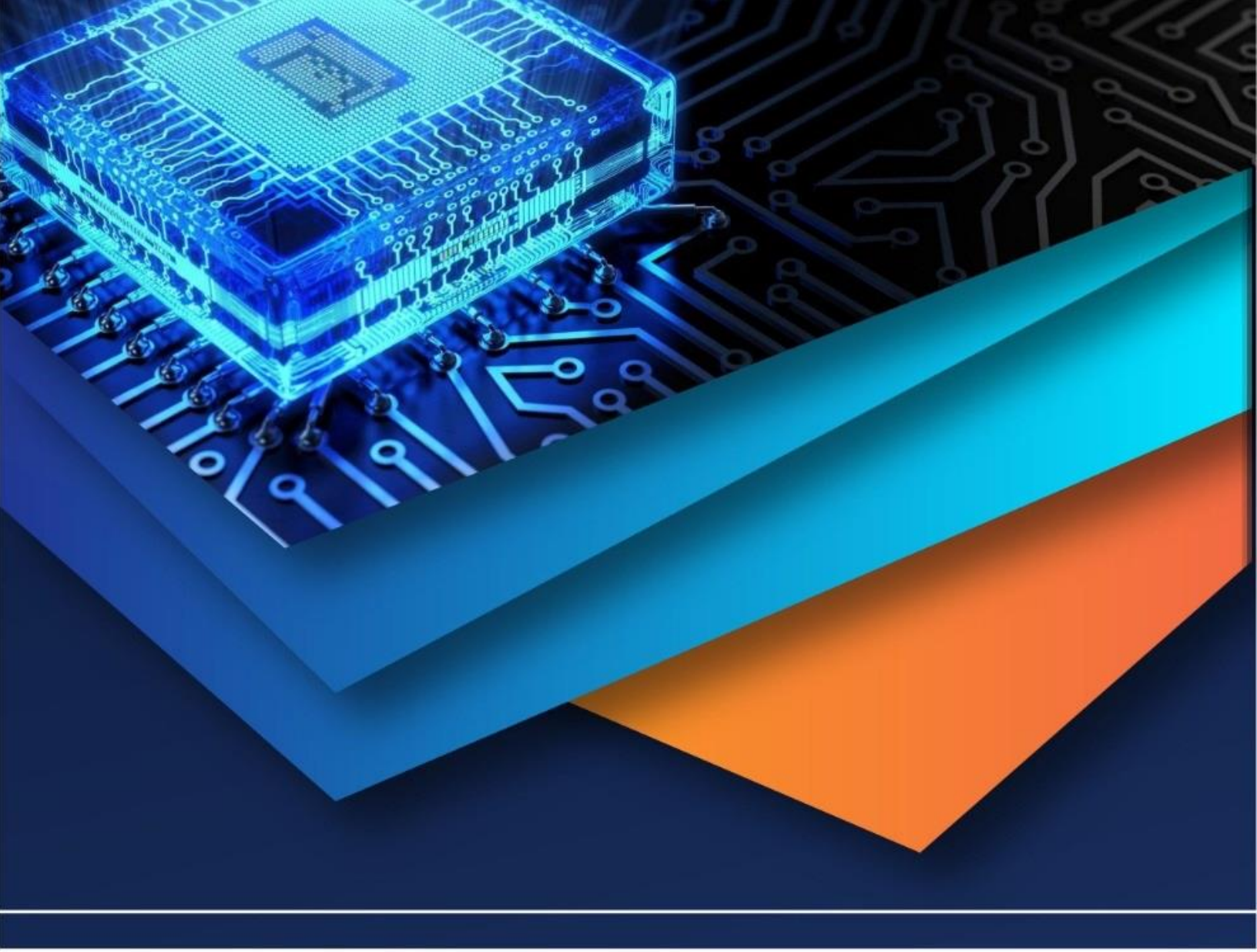
G. Aswini B.Tech with specialization in Artificial Intelligence and Machine Learning at Bapatla Women's Engineering College, Bapatla.

K. Suvarna Lakshmi B.Tech with specialization in Artificial Intelligence and Machine Learning in Bapatla Women's Engineering College, Bapatla.

P. Sailaja B.Tech with specialization in Artificial Intelligence and Machine Learning at Bapatla Women's Engineering College, Bapatla.

P. Aparna B.Tech with specialization in Artificial Intelligence and Machine Learning at Bapatla Women's Engineering College, Bapatla.

I.Haritha B.Tech with specialization in Artificial Intelligence and Machine Learning at Bapatla Women's Engineering College, Bapatla.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY