iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79696

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 14 Issue IV Apr 2026- Available at www.ijraset.com

A Contextual Similarity Analysis System for Detecting Original Authored and Machine Generated Scientific Abstract

Srilakshmi G [1], Monika S [2], Surendhar M[3], Dr. Ezhilarasan M [4]
[1, 2, 3] *Student, Department of Information Technology, Puducherry Technological University, Puducherry, 605014, India*
[4] *Professor, Department of Information Technology, Puducherry Technological University, Puducherry, 605014, India*

Abstract: ***Ensuring academic integrity has become increasingly challenging due to the rapid growth of artificial intelligence, as machine-generated scientific abstracts closely resemble human-written content. To address this, this study presents a Contextual Similarity Analysis System designed to detect and classify original-authored and machine-generated scientific abstracts. Traditional text classification systems often struggle with capturing deep contextual relationships and handling complex linguistic patterns, resulting in reduced accuracy. To overcome these limitations, we propose a three-module architecture. The first module utilizes Natural Language Processing (NLP) techniques such as text cleaning, normalization, tokenization, and dataset balancing to prepare high-quality input data. The second module employs transformer-based models including DistilBERT and BERT combined with Convolutional Neural Networks (CNN) to generate contextual embeddings and extract meaningful features. The final module integrates advanced contextual modeling using RoBERTa with Bidirectional Long Short-Term Memory (BiLSTM) networks along with model comparison techniques to perform accurate classification. Experimental results demonstrate that the proposed system achieves high validation accuracy, with hybrid and advanced models outperforming baseline approaches while maintaining efficient performance. The integrated web application enables low-latency, real-time classification with confidence scores, making it a reliable solution for ensuring academic integrity in research and educational environments.***

Keywords: Scientific Abstract Classification, Contextual Similarity Analysis, DistilBERT, BERT+CNN, RoBERTa+BiLSTM, Natural Language Processing.

I. INTRODUCTION

A. *Background*

Ensuring academic integrity has become increasingly challenging with the rapid advancement of artificial intelligence, as machine-generated scientific abstracts closely resemble human-written content. Although natural language processing techniques have been applied to address this issue, many existing systems rely on surface-level feature extraction and traditional methods. These approaches fail to capture deep contextual and semantic relationships within text. Furthermore, most systems focus only on basic classification, lacking the ability to provide accurate, scalable, and real-time detection for academic applications.

B. *Problem Statement*

Despite advancements in natural language processing and machine learning, existing scientific abstract classification systems face significant limitations. A major challenge is the high similarity between human-written and AI-generated text, which makes accurate distinction difficult. Another critical issue is the inability to effectively capture complex linguistic patterns and domain-specific terminology present in scientific abstracts. Additionally, most current systems lack deep contextual understanding, as they perform direct classification without considering semantic dependencies between words and sentences, resulting in reduced accuracy and unreliable predictions.

C. *Objective*

This project, “Contextual Similarity Analysis System for Detecting Original-Authored and Machine-Generated Scientific Abstracts,” addresses these challenges by proposing an end-to-end multi-module architecture. The primary objective is to provide accurate, efficient, and context-aware classification of scientific abstracts.

By integrating advanced preprocessing techniques, transformer-based contextual embedding models, and hybrid deep learning approaches such as CNN and BiLSTM, the system aims to capture both semantic and sequential features of text. The final objective is to deliver reliable real-time predictions with high accuracy, thereby supporting academic integrity and research validation.

II. LITERATURE REVIEW

The study focuses on algorithmic dataset preparation methods, highlighting the importance of proper data splitting for machine learning models. It explains that traditional approaches such as random and manual splitting may lead to biased data representation and affect model performance. An algorithmic splitting technique [9] is proposed to ensure balanced training, validation, and testing datasets, improving model reliability and consistency.The research emphasizes the role of preprocessing techniques in text classification, including tokenization, stop-word removal, stemming, TF-IDF, and cosine similarity. These techniques [2] help in improving feature extraction and classification accuracy by reducing noise and enhancing text representation, but they are limited to traditional methods without deep contextual understanding. The study highlights the advancement of transformer-based models such as BERT and DistilBERT, focusing on knowledge distillation for efficient NLP processing. Distilled models [16] reduce computational complexity while maintaining performance, enabling faster and scalable text classification, though they lack detailed evaluation in specific real-world applications.Overall, the literature highlights a trade-off between accuracy and computational efficiency.

While deep learning and multimodal approaches improve recognition performance, they are not suitable for real-time deployment. This justifies the need for a lightweight, landmark-based feature extraction approach, as adopted in our system, to enable efficient and robust real-time gesture recognition.

The study focuses on BERT-based transfer learning for text classification, where pretrained transformer models are fine-tuned on specific datasets to improve classification performance. The model utilizes contextual embeddings and attention mechanisms to understand semantic relationships in text, achieving high accuracy in classification tasks. Transfer learning approach [10] reduces the need for large labeled datasets and improves efficiency, but it still requires fine-tuning and may lack domain-specific adaptation in certain cases. The research proposes a CNN-based model (Set-CNN) for short text classification using semantic extension techniques. It applies multi-channel convolution, including standard and atrous convolution, to capture both local and global semantic features. Convolutional neural network approach [17] enhances feature extraction and reduces noise through semantic extension, but the model may introduce complexity and depends on additional feature expansion mechanisms for better performance. The study presents a hybrid model combining BERT and CNN for long text classification, where BERT captures global contextual features and CNN extracts local important features such as key phrases. Hybrid BERT-CNN model [15] improves classification accuracy by integrating both global and local feature learning, but it increases computational complexity and requires more resources for training and deployment. Overall, the literature highlights the importance of combining contextual embedding with hybrid deep learning techniques.

While transformer-based and hybrid models improve classification accuracy, they increase computational complexity and may not be suitable for real-time scalable systems. This justifies the use of an efficient and lightweight model in the proposed system for accurate and scalable scientific abstract classification.

Recent advancements in short-text classification focus on transformer-based architectures such as RoBERTa combined with deep learning models for improved semantic representation. The proposed RoBERTa-TextRCNN approach utilizes contextual embeddings along with convolutional neural networks to enhance classification accuracy in short texts. These models effectively capture semantic relationships and improve feature extraction efficiency [19]. However, they still require careful tuning and large datasets for optimal performance. Hybrid deep learning approaches combining BiLSTM with convolutional layers have shown strong performance in text classification tasks.

These models leverage sequential learning and feature extraction simultaneously [4], enabling better understanding of contextual dependencies and textual patterns. BiLSTM captures both forward and backward dependencies, while CNN extracts local features. However, such models may suffer from increased computational complexity and longer training time.Recent research on RoBERTa-BiLSTM hybrid models highlights the importance of combining transformer-based embeddings with sequential models for improved sentiment analysis. In these approaches, RoBERTa generates contextual embeddings while BiLSTM processes temporal dependencies [8], leading to improved accuracy and contextual understanding. Despite these advantages, the model complexity and resource requirements remain a major challenge for real-time applications.

Overall, existing studies demonstrate that while transformer-based and hybrid models significantly improve classification performance, they often introduce higher computational cost and complexity. This creates a need for an optimized system that balances accuracy and efficiency. Hence, the proposed approach focuses on integrating multiple models to achieve better performance with reduced complexity.

III. PROPOSED SYSTEM

A. *System Architecture*

The proposed system is designed as an end-to-end pipeline for detecting and classifying scientific abstracts in real time. The architecture is divided into three major modules that operate sequentially to ensure accurate and efficient classification. Initially, the input scientific abstract is provided through a user interface and passed to the preprocessing layer, where the text is cleaned, normalized, and tokenized.

The processed data is then transformed into contextual embeddings using transformer-based models. These embeddings are further analyzed using hybrid deep learning techniques to capture both semantic and sequential relationships. Finally, the system performs model comparison and generates the classification result, which is displayed through a web-based interface along with confidence scores, enabling reliable and real-time prediction.

B. *Scientific Abstract Processing and Baseline Model Training*

Module I acts as the foundational processing layer of the system and is responsible for transforming raw scientific abstract input into structured data suitable for classification. The process begins with dataset preparation, where scientific abstracts are collected, validated, and labeled as human-written or AI-generated. Duplicate and incomplete records are removed to ensure data quality. The input text is then processed using Natural Language Processing (NLP) techniques such as lowercasing, tokenization, stop-word removal, and normalization to enhance feature quality.

The preprocessed text is converted into numerical representations using the DistilBERT tokenizer, which generates contextual embeddings. These embeddings are used to train the DistilBERT model, producing baseline classification results that distinguish between human-written and AI-generated abstracts.

C. *Hybrid Deep Learning Model Development*

Module II focuses on improving classification performance by integrating contextual embedding with feature-based learning techniques.

The preprocessed data from Module I is passed to transformer-based models such as BERT to generate contextual embeddings that capture semantic relationships within the text. These embeddings are then processed using Convolutional Neural Networks (CNN) to extract important local textual features and patterns. This step enhances the model's ability to identify distinguishing characteristics in scientific abstracts. The extracted features are used to train a hybrid BERT+CNN model, which combines global contextual understanding with local feature extraction. The model generates improved prediction results with higher accuracy compared to baseline approaches.

D. *Advanced Contextual Model and Ensemble Classification*

Module III serves as the final decision-making and evaluation layer of the system. It utilizes advanced contextual modeling through RoBERTa to generate deep semantic embeddings from the input text.

These embeddings are further processed using Bidirectional Long Short-Term Memory (BiLSTM) networks to capture sequential dependencies and contextual flow within scientific abstracts.

The system then performs model comparison by evaluating the outputs from DistilBERT, BERT+CNN, and RoBERTa+BiLSTM using performance metrics such as accuracy, precision, recall, and F1-score. Based on this evaluation, the best-performing model is selected to generate the final classification result. The output, along with confidence scores, is displayed through a web-based interface, ensuring accurate, reliable, and real-time prediction.

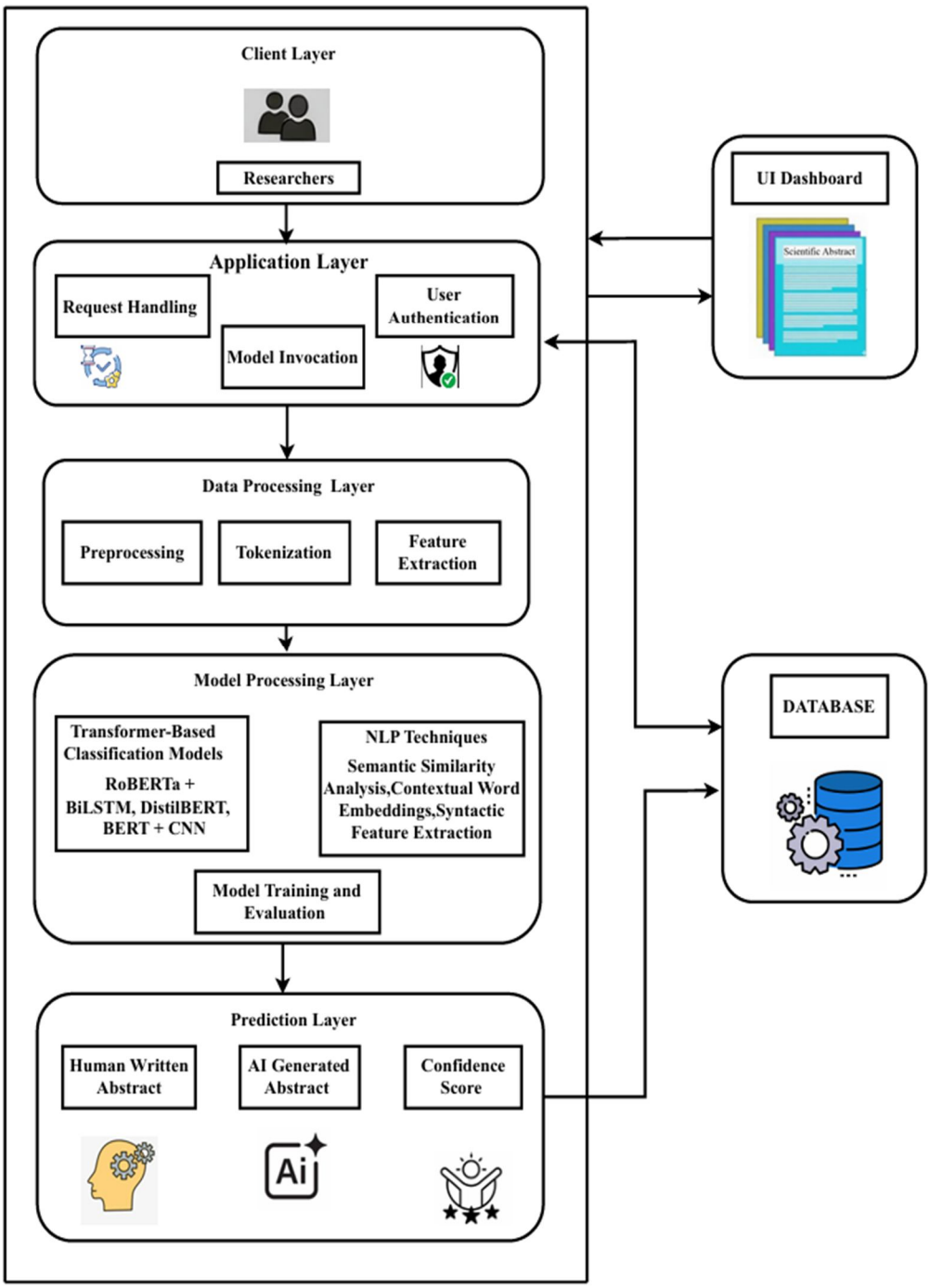


Fig. 1 High-Level Architecture of the Proposed System

IV. RESULTS AND DISCUSSION

A. *Comparative Analysis*

Traditional scientific text classification systems fall into two broad categories:

1) Rule-based or statistical methods using TF-IDF, cosine similarity, or keyword matching,

2) Basic machine learning classifiers that lack deep contextual understanding. Such systems perform adequately on simple datasets but fail when handling complex semantic structures, paraphrased content, or AI-generated text that closely mimics human writing. They also lack the ability to capture deep linguistic patterns and often produce less reliable predictions.

The proposed system advances beyond these limitations in three ways:

a) Contextual embedding – Instead of relying on surface-level features, the system uses transformer-based models (BERT, DistilBERT, RoBERTa) to capture deep semantic and syntactic relationships within scientific abstracts.

b) Hybrid deep learning – The integration of CNN and BiLSTM with transformer embeddings enables the system to capture both local textual patterns and sequential dependencies, improving classification accuracy.

c) Ensemble-based classification – Multiple models are compared using performance metrics, and the best-performing model is selected to generate accurate and reliable predictions.

Feature	Prior Systems (ML + Basic DL)	Proposed System
Input Data	Raw or basic preprocessed text	Preprocessed text with contextual embeddings
Context Understanding	Limited (CNN/LSTM)	Deep contextual (BERT, RoBERTa)
Feature Extraction	Single model-based features	Hybrid (Transformer + CNN + BiLSTM)
Classification Approach	Individual model prediction	Ensemble-based model comparison
Output	Basic classification result	Accurate result with confidence score

Table 1: Comparative Analysis with Existing Systems

B. Evaluation Metrics

To demonstrate both recognition quality and real-time usability, we report metrics across three layers of the system.

1) Offline Accuracy

a) Dataset

- Total Dataset: 28,667
- Training set : 22.933
- Validation/test Set: 5734

b) DistilBERT (Baseline Model)

- Validation accuracy: 98.2% at 3 epoch.
- Provides fast and lightweight classification with good generalization.
- Performs well on standard abstract classification but slightly lower accuracy compared to advanced hybrid models.

c) BERT + CNN (Hybrid Model)

- validation accuracy: 99% at 3 epoch.
- Combines contextual embeddings with convolutional feature extraction for improved performance.
- Captures both global semantic meaning and local textual patterns effectively.

d) RoBERTa + BiLSTM (Advanced Contextual Model)

- validation accuracy: 99.7% at 3 epoch.
- Achieves highest performance by capturing deep contextual and sequential dependencies.
- Provides robust classification results across complex and ambiguous abstracts.

e) Overall Evaluation

- Hybrid and ensemble-based approaches outperform standalone models in accuracy and reliability.
- The system maintains consistent performance across different abstract categories.
- Model comparison confirms that RoBERTa+BiLSTM provides the best results, while DistilBERT ensures efficiency.

2) Real-Time Performance Metrics

a) Preprocessing Latency:

- NLP preprocessing operations such as tokenization and normalization are completed within a few milliseconds per abstract, ensuring fast data preparation.

b) Model Training Environment:

- The models are trained using Google Colab with Tesla T4 GPU acceleration, which significantly reduces training time and improves model performance.

c) Model Inference Latency:

- DistilBERT performs classification in less than 100 ms on CPU, enabling quick prediction of whether the abstract is human-written or AI-generated.

d) End-to-End Delay:

- From user input to final output display: preprocessing (~20–30 ms) + model inference (<100 ms) + result rendering (~20 ms) = approximately 150–200 ms delay, suitable for real-time applications.

e) Throughput:

- The system can process multiple abstracts per second depending on input size and system configuration, ensuring efficient performance for practical usage.

3) Output Quality & Stability Metrics

a) Confidence Statistics:

- The system generates confidence scores for each prediction, with most outputs showing high confidence (>0.85), ensuring reliable classification results.

b) Model Accuracy:

- Transformer-based and hybrid models achieve high classification accuracy due to effective contextual and semantic feature extraction.

c) Model Reliability:

- The use of multiple models (DistilBERT, BERT+CNN, RoBERTa+BiLSTM) ensures stable predictions and reduces misclassification through model comparison.

d) UI Feedback and Visualization:

- The system displays classification results along with confidence scores in real time through an interactive web interface, allowing users to easily interpret outputs.

C. Graph output Comparative analysis with Existing

This bar chart compares the performance of the existing DistilBERT model and the proposed approach. The DistilBERT model achieves around 97% accuracy, showing strong baseline performance. The proposed model improves this further, reaching 98% validation accuracy through enhanced contextual and hybrid learning. This demonstrates a clear performance improvement and better classification capability of the proposed system.

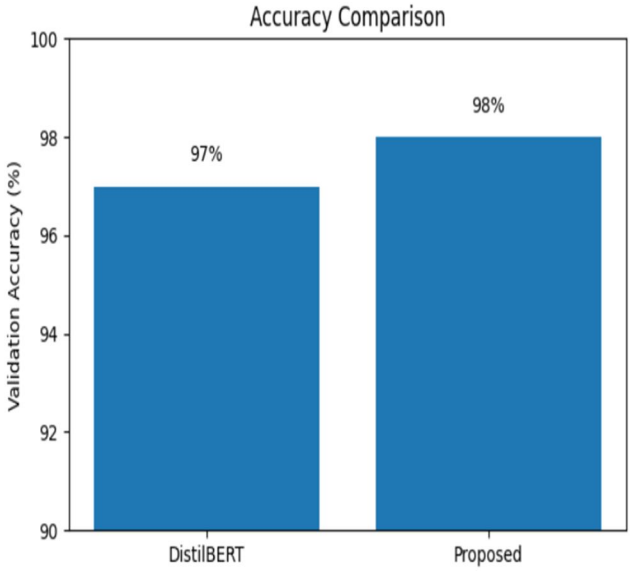


Fig 2: Accuracy Comparison

V. CONCLUSION

This project successfully develops an intelligent system for detecting and classifying scientific abstracts as human-written or machine-generated, thereby supporting academic integrity and reliable content validation. By transforming textual input into contextual embeddings using transformer-based models and enhancing feature extraction through hybrid deep learning techniques such as CNN and BiLSTM, the system effectively captures both semantic and sequential relationships within the text. The integration of DistilBERT, BERT+CNN, and RoBERTa+BiLSTM models ensures accurate and robust classification through model comparison strategies. Achieving high validation accuracy while maintaining efficient real-time performance, the system demonstrates its effectiveness as a scalable and reliable solution for academic and research applications.

VI. FUTURE SCOPE

Future enhancements include extending the system to support multi-language scientific abstract classification, enabling broader applicability across diverse research domains. Incorporating explainable AI techniques can improve transparency by highlighting key features influencing model predictions. Additionally, optimizing models using techniques such as quantization and lightweight transformer architectures can enhance performance on low-resource and edge devices. The integration of large-scale datasets and domain-specific fine-tuning can further improve classification accuracy. Continuous system updates through user feedback and adaptive learning mechanisms will help maintain performance and ensure long-term reliability.

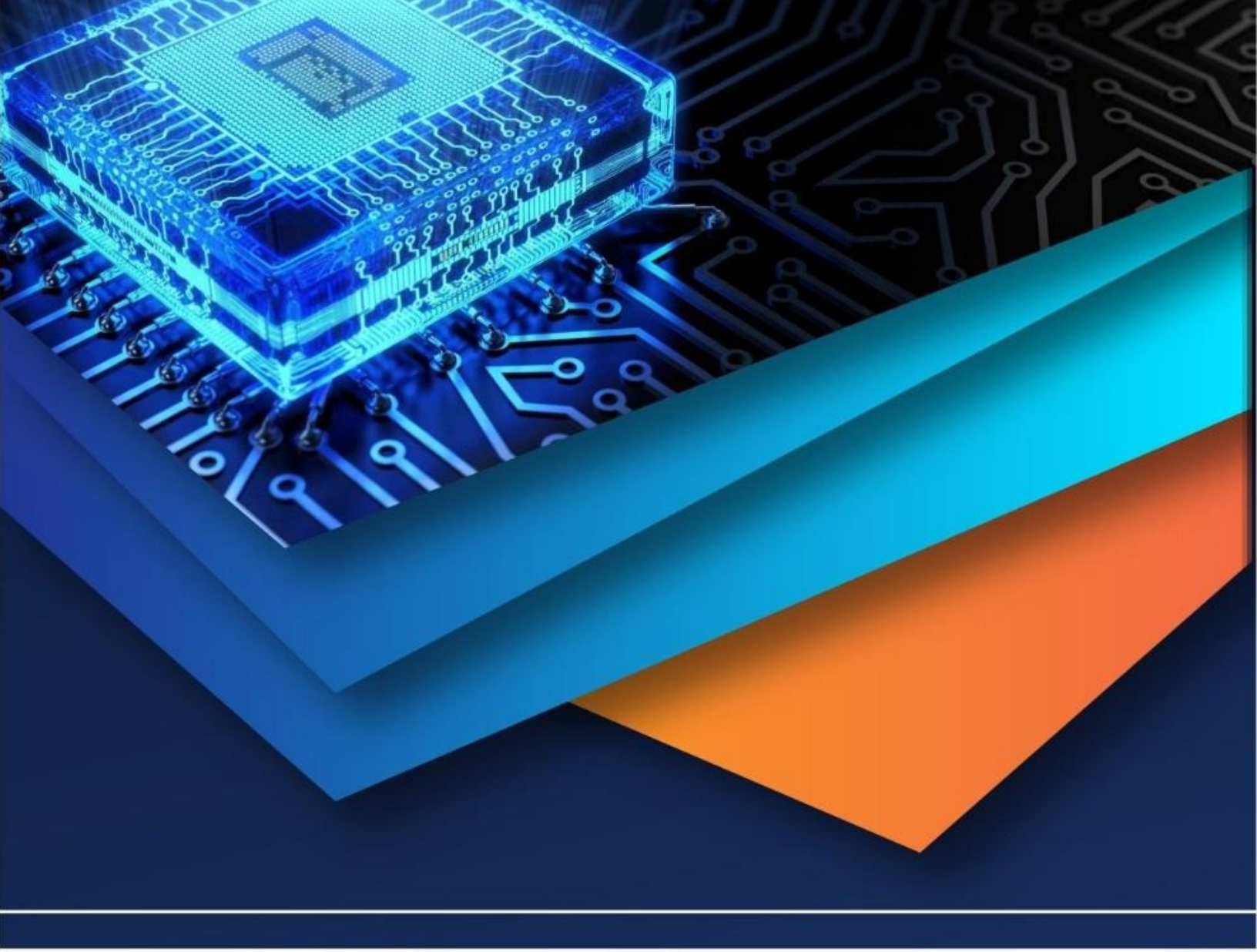
VII. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Puducherry Technological University , especially the Department of Information Technology, for providing the necessary infrastructure and support throughout the development of this project. Special thanks to our guide, Dr. M.Ezhilarasan, Professor - Puducherry Technological University for her valuable guidance, constant encouragement, and insightful suggestions that helped shape this research. We would also like to thank our peers and the volunteers who participated in the user testing phase, whose feedback greatly contributed to improving the system. Finally, we appreciate our families and friends for their unwavering support and motivation throughout this journey.

REFERENCES

[1] Ali Al Bataineh, Rachel Sickler, Kerry Kurcz, Kristen Pedersen, “AI-Generated Versus Human Text: Introducing a New Dataset for Benchmarking and Analysis,” IEEE Transactions on Artificial Intelligence, Vol. 6, No. 8, pp. 2241–2247, 2025.
[2] Ali Saleh Alammary,“BERT Models for Arabic Text Classification: A Systematic Review,” Applied Sciences, vol. 12, no. 11, p. 5720, June 2022.
[3] Ammar Ismael Kadhim, “An Evaluation of Preprocessing Techniques for Text Classification,” International Journal of Computer Science and Information Security (IJCSIS), vol. 16, no. 6, pp. 22–32, June 2018.
[4] Gang Liu and Jiabao Guo, “Bidirectional LSTM with Attention Mechanism and Convolutional Layer for Text Classification,” Neurocomputing, vol. 337, pp. 325–338, Feb. 2019.
[5] Hikmat Ullah Khan, Anam Naz, Fawaz Khaled Alarfaj, and Naif Almusallam, “Identifying Artificial Intelligence-Generated Content Using the DistilBERT Transformer and NLP Techniques,” Scientific Reports, vol. 15, p. 20366, May 2025.
[6] John Blake, Abu Saleh Musa Miah, Krzysztof Kredens, Jungpil Shin, “Detection of AI-Generated Texts: A Bi-LSTM and Attention-Based Approach,” IEEE Access, Vol. 13, pp. 71563–71574, 2025.
[7] Joshua Mawudem Gakpetor, Martin Doe, Michael Yeboah-Sarpong Damoah, Dominic Dalyngton Damoah, John Kingsley Arthur, and Michael Tetteh Asare, “AI-Generated and Human-Written Text Detection Using DistilBERT,” Proceedings of the 2024 IEEE SmartBlock4Africa Conference, Accra, Ghana, September 30 – October 4, 2024.
[8] Md Mostafizer Rahman, Ariful Islam Shiplu, Yutaka Watanobe, and Md Ashad Alam, “RoBERTa-BiLSTM: A Context-Aware Hybrid Model for Sentiment Analysis,” IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 9, no. 6, pp. 3788–3805, Dec. 2025.
[9] Khalid M. Kahloot and Peter Ekler, “Algorithmic Splitting: A Method for Dataset Preparation,” IEEE Access, vol. 9, pp. 125229–125237, September 6, 2021.
[10] Rukhma Qasim, Waqas Haider Bangyal, Mohammed Abdullah Alqarni, and Abdulwahab Ali Almazroi,“A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification,” Journal of Healthcare Engineering, vol. 2022, Article ID 3498123, Jan. 2022.
[11] Shun Moriya and Chihiro Shibata,“Transfer Learning Method for Very Deep CNN for Text Classification and Methods for Its Evaluation,” in Proc. IEEE 42nd Int. Conf. Computer Software and Applications (COMPSAC), Tokyo, Japan, 2018, pp. 153–154.
[12] Tugba Celikten and Aytug Onan, “Exploring Text Similarity in Human and AI-Generated Scientific Abstracts: A Comprehensive Analysis,” IEEE Access, Vol. 13, pp. aa Diyana Aldeen Ahmed, Thura Abbas, and Ayyad Rodhan Abbas, “Review of Detecting Text 74313–74322, 2025.
[13] Toka A. Mohamed, Mohamed H. Khafgy, Ahmed B. Elsedawy, And Ahmed S. Ismail, ” A Proposed Model for Distinguishing Between Human- based and ChatGPT Content in Scientific Articles ,” IEEE Access Vol.12, pp. 121251-121259, 22 Auguest 2024.
[14] Wei Shi, Miao Song, and Yong Wang, “Perturbation-Enhanced-Based RoBERTa Combined with BiLSTM Model for Text Classification,” Proc. ICETIS Conference, pp. 294–298, 2022.
[15] Xinying Chen, Peimin Cong, and Shuo Lv,“A Long-Text Classification Method of Chinese News Based on BERT and CNN,” IEEE Access, vol. 10, pp. 34046–34057, 2022.

[16] Yazdan Zandiye Vakili, Avisa Fallah, and Hedihe Sajedi, "Distilled BERT Model in Natural Language Processing," Proceedings of the 2024 14th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, November 19–20, 2024.
[17] Yajian Zhou, Jiale Li, Junhui Chi, Wei Tang, and Yuqi Zheng,"Set-CNN: A Text Convolutional Neural Network Based on Semantic Extension for Short Text Classification," Knowledge-Based Systems, vol. 257, p. 109948, 2022.
[18] Yunxiang Zhang and Zhuyi Rao, "n-BiLSTM: BiLSTM with n-gram Features for Text Classification," Proc. IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), pp. 1056–1059, 2020.
[19] Zixian Guo, Lu Han, and Ligu Zhu, "Research on Short Text Classification Based on RoBERTa-TextRCNN," Proc. 2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI), pp. 845–850, 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY