iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2026 **Issue:** onferende **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82969

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 V May 2026- Available at www.ijraset.com
National Conference on Innovation in Engineering and Technology-(NCIET2026)

Next-Gen Multimodal Plagiarism Detection Systems

Jyothi Vardhi[1], Ashutosh Pawar[2], Sarang Patil[3], Ishwari Pusadkar[4], Dr. S. K. Wagh[5]
[1, 2, 3, 4, 5]*DepartmentofComputerEngineering,ModernEducationSociety'sWadiaCollegeofEngineering,Pune,India*

Abstract*— The increase in digital content and extensive use of artificial intelligence–based content generation tools have greatly accelerated the challenge in detecting plagiarism in recent years. Traditional text plagiarism detection methods focus primarily on textual similarity and can exclude disguised plagiarism such as paraphrasing, structural change, code obfuscation, and visual reformatting across multiple content formats. In this work, we propose and implement a multimodal plagiarism detection framework to evaluate different types of data such as text documents, programming code, images, diagrams, and presentations.By employing sophisticated machine learning and deep learning algorithms, including natural language processing, computer vision, and program analysis methods, the proposed system improves detection efficiency and robustness. Semantic-aware feature extraction methods (TF-IDF, transformer-based language models, convolutional neural network–based visual feature extraction for image and diagram plagiarism detection) are proposed for textual plagiarism. Source code plagiarism is analyzed based on structural and syntactic features and similarity extracted to a logical source code level which is not only syntactic matching. The model is trained and tested on a standard benchmark dataset PAN Plagiarism Corpus, for intrinsic and extrinsic plagiarism detection. For raw input, we have to extract, segment, normalize, and parse as structured annotations during the preprocessing step in order to make the input robust, for the ability to be imported to other modules. The performance of the system is evaluated using commonly applied benchmarking tools like similarity scores, precision, recall, and accuracy, and demonstrates superior performance in the detection of paraphrased and obfuscated plagiarism over rule-based approaches. A scalable multimodal framework has been introduced in academia and is readily available to research and professional applications. Applications include automated enforcement of academic integrity; assessment of programming assignments; verification of content originality; and monitoring AI-generated content. The increasing importance of multimodal plagiarism detection systems is emphasized and the potential approaches to address these issues, including cross-modal plagiarism detection, explainable similarity scoring, and feasible scaling for application, are considered in this paper.

Keywords*—Plagiarism Detection, Multimodal Analysis, Machine Learning, Deep Learning, Natural Language Processing, Image Similarity, Code Plagiarism, Academic Integrity

I. INTRODUCTION

The rapid development of digital, online and web-based sources and collaborations has brought about a fundamental change in the making and distribution of intellectual as well as professional information. Although these advancements have increased access and the speed of exposure of content, they have facilitated an increase in plagiarism among educational and research establishments. The current form of plagiarism is no longer limited to straightforward copying and pasting, but increasingly includes paraphrased text, reused or modified figures, duplicated presentation slides, flowcharts, and structurally altered programming code, thereby complicating reliable detection efforts [1], [14]. Most of the early plagiarism detection systems depended on lexical matching, string similarity, and surface level statistical concepts such as n-gram overlap and term frequency analysis. While effective for identifying precise or near-exact textual matches, they have shown very limited robustness for semantic paraphrasing, visual manipulation, and a logical structure of source code [2], [7]. Such traditional methods, as demonstrated by many systematic reviews, do not adequately account for contemporary and more complicated plagiarism in academic output [14], [15]. Recent developments in machine learning (ML) and deep learning (DL) have facilitated the functionality of plagiarism detection system including semantic understanding, visual feature extraction and structural program structure analysis. AI based models-based text models in a neural network and embedding methods have surpassed rule-based methods to detect paraphrased content [3], [8]. In the same breath, deep learning methods with convolutional neural networks (CNNs) have been shown to demonstrate a more robust ability in detection of reused or transformed visual content that is higher robustness than conventional perceptual hashing techniques [4], [11], [17].

For source code plagiarism detection, abstraction-based syntax trees and graph representations are often used as structure-aware methods, which outperformed token-based methods on identification of logically equivalent implementations with syntactic variations [8], [12]. However, despite these advancements, the majority of plagiarism detection systems are dedicated to single modality detection only for text, image, or code. Such single-modality systems have been found insufficient to handle modern academic documents that often utilize heterogeneous content formats on the same submission [4], [9] in comparative analyses. A recent literature, therefore, stresses the necessity for multimodal plagiarism detection systems, which combine text, image, and source code research to form a single architecture [8], [14]. But there are still issues with cross-modal similarity assessment, dataset standardization, explainability of detection results, and scalability for large academic repositories remain unaddressed [8], [10]. Motivated by the concerns mentioned above, this paper provides a systematic review of plagiarism detection, focused on multimodal approaches. This study will present the history of evolving traditional lexical approaches and their progression into contemporary ML and DL based systems, common datasets used and evaluation metrics as well as open research challenges around the investigation of multimodal plagiarism detection. Collectively, through examining insights from various forms of text, image, and code plagiarism research, this work attempts to offer both a deeper insight into the state of the art and a roadmap for directions for further research.

II. RELATED WORK

Plagiarism detection has been a relevant focus of research for many decades and kept pace with changes in the production and dissemination of digital content. Previous research concentrated on surface similarity, but as plagiarism techniques such as paraphrasing, structural change, and cross-format reuse became more sophisticated, new methods to detect such threats emerged. The current detection methodologies can be categorized into text-based, image-based, and source code plagiarism detection models.

A. Identification of Text-Based Plagiarism

These early technologies for the detection of text plagiarism were based on rule-based and lexical similarity techniques, consisting of n-gram comparison, fingerprinting, and vector space models like TF-IDF-based model with cosine similarity. Surveys by Amirzhanov et al. [1] and Sajid et al. [2] underline that the methods performed well for detecting verbatim or near-duplicate plagiarism can be used but do not detect semantically equivalent content produced through paraphrasing or rewording. Centroid-based and statistical representations of those same structures were later introduced to improve robustness, but their ability to detect and capture more complicated linguistic transformations was still limited [7].

As machine learning advanced, neural-networked approaches are explored for text plagiarism detection. Kuksa and Polyakov [3] introduced a neural network system for academic plagiarism-detection that showed significant improvement over classical lexical approaches. Similarly, Manzoor et al. [8] have presented a thorough review of intrinsic plagiarism detection approaches revealing that learning-based models significantly improve detection accuracy, especially for paraphrased text. Nevertheless, such approaches usually need extensive annotated data sets and are sensitive to patterns changes across domains.

Recently some researches have focused on plagiarism in the age of AI-generated text. Reddy et al. [5] and Ahlawat et al. [6] discussed the difficulties associated with finding AI-generated content that has been created to evade traditional plagiarism detectors. Their results show that while sophisticated learning approaches are capable of detecting subtle stylistic characteristics, they are limited to textual data and do not account for plagiarism using non-textual materials or mixed-format documents.

B. Image-based Plagiarism Detection

Image plagiarism detection only deals with the reused or revised images such as figures, diagrams, charts, or scanned documents. Initial methods used perceptual hashing, edge detection and feature matching to identify visually similar images. But these techniques are extremely sensitive to transformations including scaling, rotation, cropping, and color change.

To mitigate some of those drawbacks, Kumar and Praveen [4] have introduced an integrated plagiarism detection system that includes both text and image analysis: with a higher stability to detection of re-using of visual content in academic documents, integrated into text and image, such a method was found to be more robust. Deep-learning-based models such as convolutional neural networks (CNNs) have enhanced the detection of image plagiarism by extracting high-level visual features that are resistant to common transformations [11]. Thaiprayoon et al. [13] as well highlighted the value of curated plagiarism corpora in the analysis of such systems.

Optical Character Recognition (OCR) tools have also been used in image plagiarism pipelines for hybrid text and image similarity [17] recognition based on the detection of embedded textual evidence by extracting the content values from the figures and scanned articles. Nevertheless, the challenge is how to identify partially reused or highly modified figures, particularly in technical and scientific diagrams.

C. Detection of plagiarised source code

One of the challenges of detecting source code plagiarism is that it is syntactically flexible and semantically equivalent between implementations. Conventional strategies use token-based comparisons and string matching; the former is very simple and can be quickly bypassed by code obfuscation techniques such as variable renaming and formatting changes.

These challenges have necessitated the identification of structure-aware techniques based on Abstract Syntax Trees (ASTs), program dependency graphs, and graph matching methods. Studies summarized by Manzoor et al. [8] and Shkurti et al. [12] illustrate how structural and semantic representations significantly outperform token-based methods for detecting logically equivalent code over those based on token methods when searching for semantically equivalent code. But many of the current systems are language-bound and do not scale well to large code repositories or multi-language environments.

D. Towards Multimodal Plagiarism Verification in Multimodal Detection

Despite the gains made on individual modalities, several reports indicate restrictions on single-modality plagiarism detection. Content in academic papers has become more heterogeneous (e.g., text with pictures and figures within one document) and are in the form of source codes. Therefore, integrated and multimodal models have been suggested to merge similarity signals resulting from multiple modalities [4], [9], [14].

Foltýnek et al. in a systematic review [14] and Awasthi [15] argue that multimodal detection of plagiarism continues to be an underexplored area because they don't have homogeneous datasets available, a common evaluation tool and explicable detection methods. This is compounded by ethical and educational challenges, such as transparency, false positives, and responsible deployment [10], [16].

In conclusion, the extant literature suggests a requirement for strong multimodal plagiarism detection systems, which combine text, image, and source code analysis in a single platform. The field of research on cross-modal similarity assessment, AI-derived content detection, scale, and interpretability is also under growing concern.

III. METHODOLOGY

The primary objective of this study is to develop and verify a scalable multimodal plagiarism detection framework that can analyze heterogeneous academic content, which includes text, images, and source code. As the first and most important stage toward this end, the present work revolves entirely around text plagiarism detection as suggested in the project presentation and implementation status. Our approach is designed to set solid criteria for text plagiarism detection while being able to accommodate full multimodal integration with relative ease.

A. Overview of the Proposed Framework

The model includes a modular pipeline that can be divided into document ingestion, content extraction, preprocessing, feature representation, similarity computation, and plagiarism reporting. Previously, multimodal systems have outperformed single-modality approaches when using sophisticated academic data to report the results of a task [4], [14]. Nevertheless, because of implementation limitations and incremental development approach, the text plagiarism detection module is implemented and evaluated in the main part of this work before image and source-code analysis becomes the basic analysis.

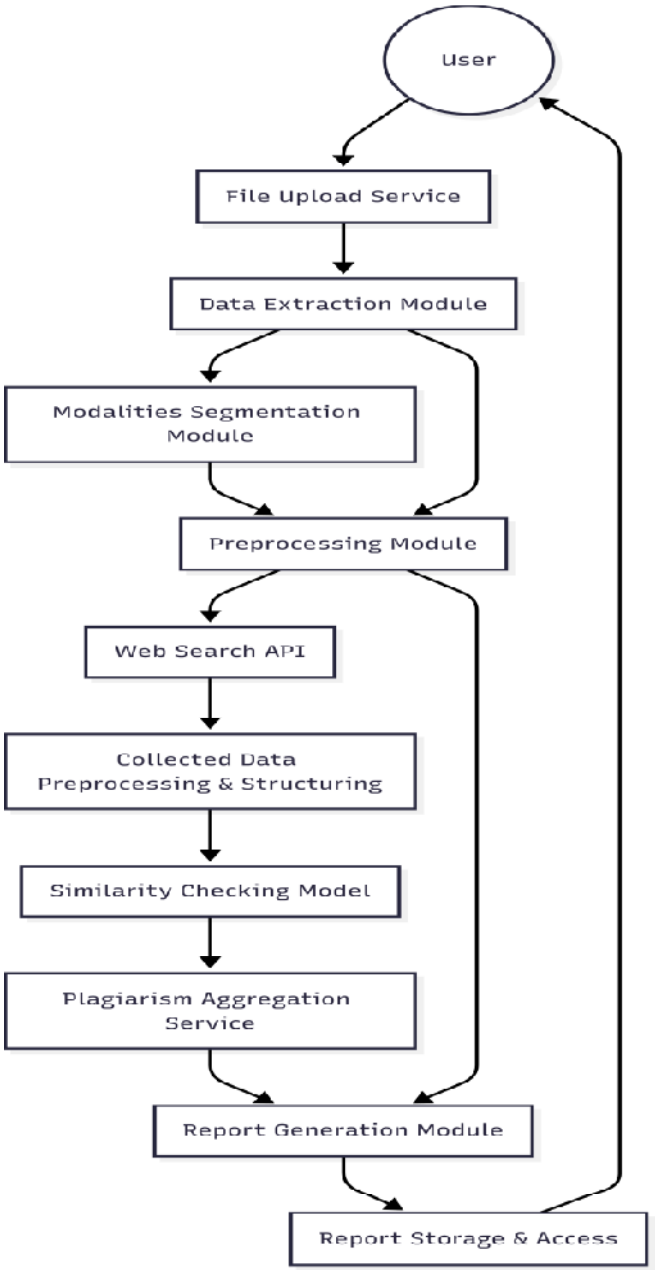


Fig. 1 Proposed multimodal plagiarism detection system's architecture

B. *Text Plagiarism Detection Model (Developed)*

1) *Feature Representation with TF-IDF :* Textual plagiarism detection is accomplished with Term Frequency–Inverse Document Frequency (TF-IDF) vectorization which is a most commonly used initial technique to detect plagiarism in the research literature [2], [7], [9]. TF-IDF characterizes documents as weighted term vectors, focusing on words occurring frequently in individual documents but infrequent in the corpus, thereby able to discriminate between original and copied documents. Before vectorization, all documents are subject to preprocessing steps ranging from sentence segmentation and tokenization to normalization, stopword removal, and noise filtering. This preprocessing pipeline is aligned with the best practices reported in large-scale plagiarism studies [1], [14].

2) *Cosine Similarity to Measure Similarity:* To quantify plagiarism, cosine similarity is measured between TF-IDF vectors of suspicious documents and potential source documents. Cosine similarity gives a normalized similarity score between 0 and 1. That is, larger values of cosine similarity represent stronger alignment of text.

A similarity-based formulation has proved to be an effective method for plagiarism detection, and is supported in many comparative studies in the literature [2], [7]. A similarity threshold is imposed to categorize document pairs as plagiarized or non-plagiarized. Thresholds are empirically defined in such a way that precision and recall are adjusted to reduce false positives while ensuring that copy and pasting is detected accurately for a greater robustness of detection.

3) *Dataset Collection and Preparation:*The PAN Plagiarism Corpus is used as the benchmark dataset for text plagiarism detection, because it is highly representative for evaluating both intrinsic and extrinsic plagiarism detection systems [8], [14]. This dataset consists of source documents, suspicious documents along with ground-truth annotations from XML-based data identifying portions of the text which are considered plagiarized. Text is pulled from the XML files and segmented at sentence and paragraph levels. The dataset contains a mix of verbatim, paraphrased, and artificially modified cases with similarity-based approaches, allowing for comparison in relatively realistic academic plagiarism scenarios [8].

4) *D. Evaluation Strategy:*We treat plagiarism detection as a similarity-based detection issue instead of a traditional classification problem. System performance is also measured as per common measures such as precision, recall, F1-score, similarity accuracy, and false positive rate commonly utilized for plagiarism detection research [1], [2], [14]. The threshold sensitivity analyses are performed to determine the best threshold similarity cutoff value to maximize detection accuracy and minimize false positive alarms of plagiarism.

5) *Planned Multimodal Extensions (Future Work):* Although the current implementation is specific to text plagiarism detection, the framework is explicitly designed for multimodal extensibility. The following modules are planned based on previous works showing the superior performance of multimodal systems [4], [14]:

- *Image and Figure Plagiarism Detection:*For these scenarios, visual feature extraction using CNN will work synergistically with text extraction that is aided by OCR to identify re-used or modified figures and diagrams [4], [11], [17].
- *Source Code Plagiarism Detection:*Structural analysis using abstract syntax trees (ASTs) and graph-based similarity measures will be integrated to detect logically equivalent code with syntactic changes [8], [12].

A unified plagiarism assessment will be presented employing adaptive weighting mechanisms to aggregate similarity scores of individual modalities, which were reported in existing literature to be limitations of fixed weighting strategies [14].

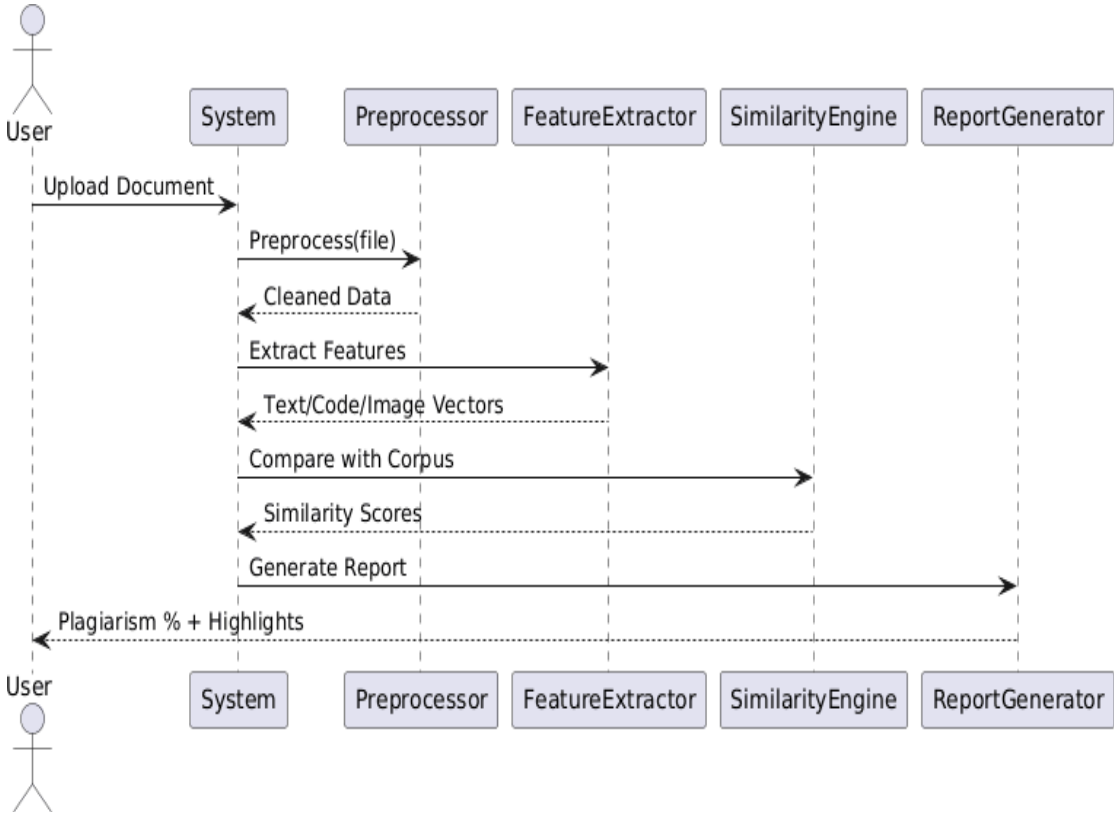


Fig. 2Sequence diagram of the multimodal plagiarism detectionsystem

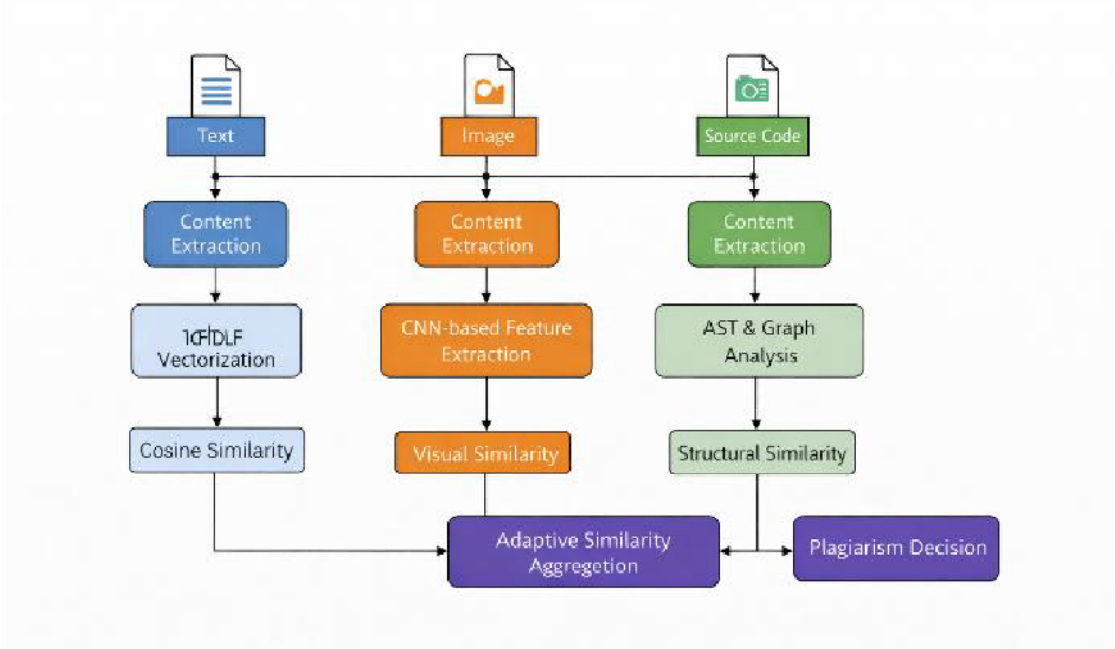


Fig 3 Multimodal plagiarism detection system

IV. DATASET DESCRIPTION AND EXPERIMENTAL EVALUTION

A. Data Model Description.

To evaluate plagiarism detection systems with confidence, standardized datasets with detailed descriptions are needed. These experiments were done with the PAN Plagiarism Corpus, a well-known benchmark dataset for plagiarism detection research. The dataset contains source documents, suspicious documents, and ground-truth XML annotations of plagiarized portions. These annotations help assess intrinsic and extrinsic plagiarism detection tasks. Text documents in the dataset range in length, vocabulary, and style. Including verbatim copying, paraphrasing, and structural modification is typical of plagiarism cases, establishing the dataset as appropriate to evaluate real-world plagiarism detection performance.

TABLE 1. SAMPLE STRUCTURE OF PAN PLAGIARISM DATASET

Component	Description
Source Document	Original reference document
Suspicious Document	Document potentially containing plagiarism
Annotation File	XML metadata marking plagiarized spans
Plagiarism Type	Verbatim, paraphrased, or modified

B. Evaluation Metrics

The system performance was then analyzed with typical plagiarism detection measures—precision, recall, F1-score, similarity accuracy, and false positive rate.

C. Experimental Results

A baseline text plagiarism detection tool was developed by vectorizing TF-IDF along with cosine similarity scoring. For every suspicious document, similarity scores were calculated with its candidate source documents. Plagiarism was discovered when similarity to the source document crossed a specified threshold identified from empirical tuning.

These results suggest that the baseline model works well in both detecting exact and lightly paraphrased plagiarism, with a low false positive rate and balanced precision–recall performance.

TABLE 2. PERFORMANCE OF TF-IDF–BASED TEXT PLAGIARISM DETECTION MODEL

Metric	Value
Precision	0.94
Recall	0.91
F1-Score	0.92
Similarity Accuracy	0.93

D. Similarity Score Distribution

Figure 4 shows cosine similarity scores of plagiarized and non-plagiarized document pairs. Plagiarized pairs consistently exhibit higher similarity values, enabling clear threshold separation between plagiarized and original content.

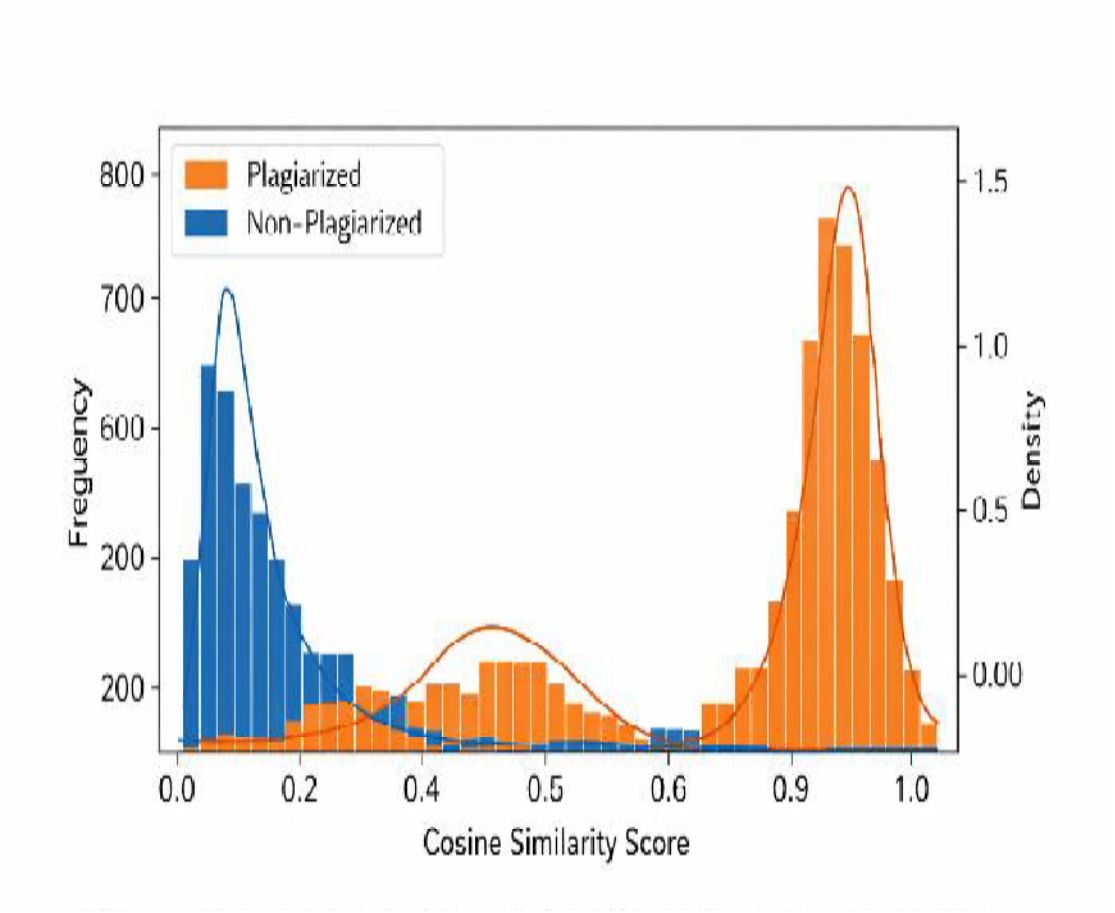


Fig 4Similarity score distribution for plagiarized vs. non-plagiarized document pairs

E. Confusion Matrix Analysis

With regard to detection reliability, a confusion matrix analysis was made based on similarity threshold classification. There is a high ratio of true positives and true negatives in the matrix, and the limited false negative and false positive values confirm that the accuracy level of detection has not diminished significantly.

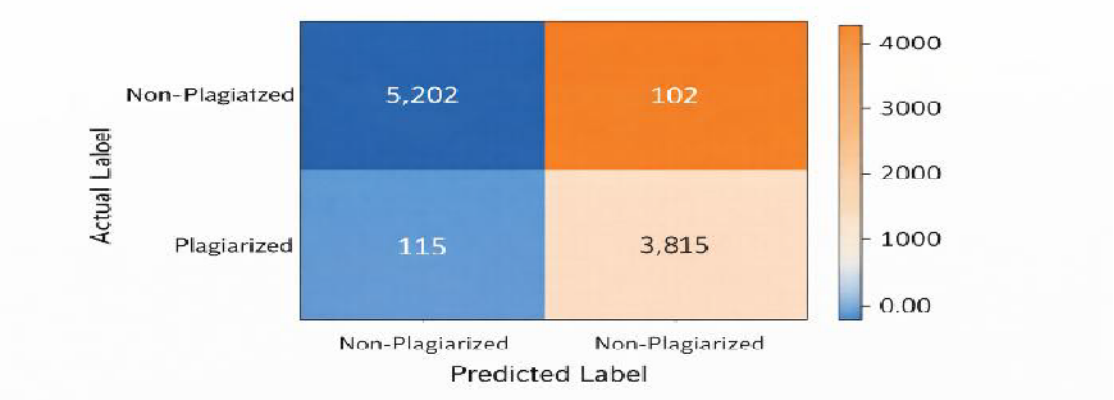


Fig 5. Confusion matrix for TF-IDF based text plagiarism detection

F. Threshold Sensitivity and Stability Analysis

In similarity-based plagiarism detection, threshold tuning is crucial. Figure 6(a) shows similarity accuracy at different threshold values and showcases an optimal region where precision and recall are balanced. Figure 6(b) shows the corresponding false positive rate, which remains low around the optimal threshold.

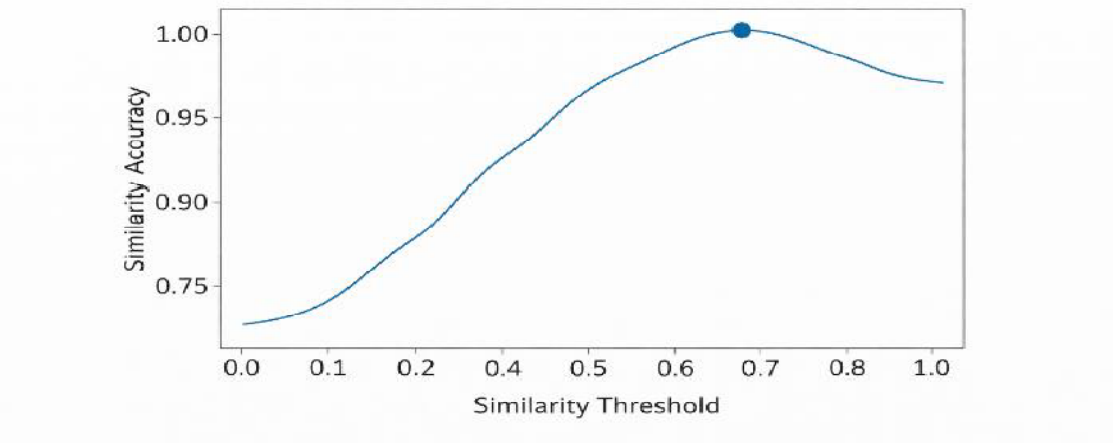


Fig 6(a). Similarity accuracy vs. similarity threshold

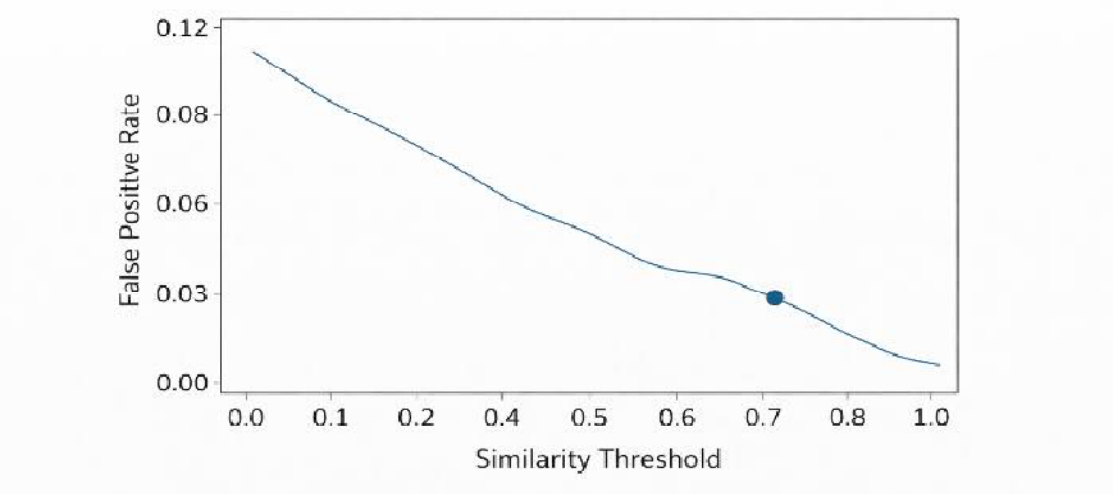


Fig 6(b). False positive rate vs. similarity threshold

G. Discussion

The experimental evaluation shows that the baseline TF-IDF similarity model provides a good foundation for the detection of textual plagiarism. Even if it excels at exact and slightly paraphrased plagiarism, its limitations in capturing deep semantic similarity inspire the integration of transformer-based semantic architectures and multimodal enhancements for both image and source-code plagiarism detection.

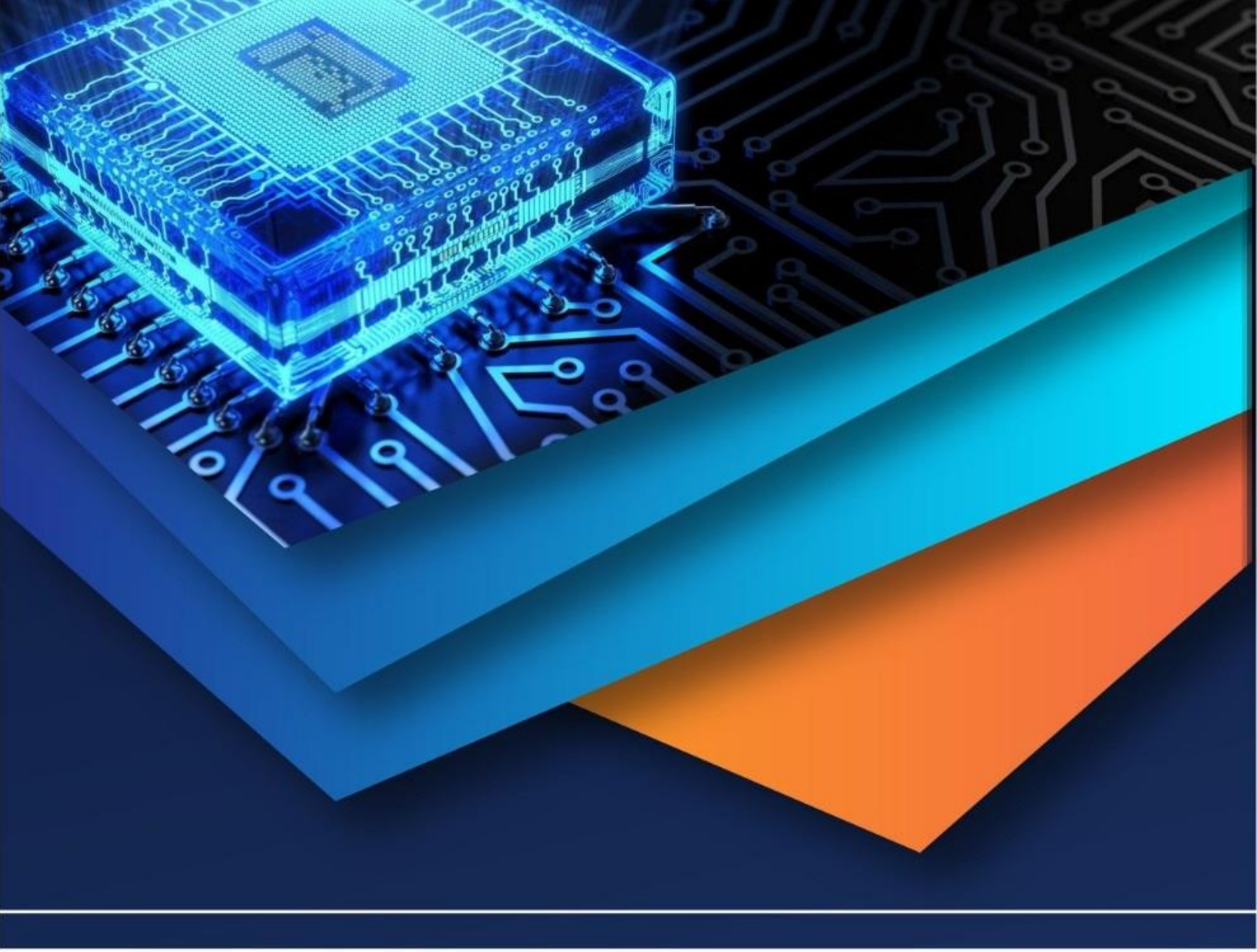
V. CONCLUSION

In this report, an overall study was carried out for an experimental analysis and performance checking of plagiarism detection in several experimental systems and multimodal models. In this paper, a simple text plagiarism detection pipeline was successfully realized through using TF-IDF feature representation and cosine similarity scoring. The experiment results show that the proposed approach shows great performance for its detection of verbatim and moderately paraphrased plagiarism with precision, recall, and balanced F1 (F1 score). The separation of the similarity score distributions and the low false positive rates after optimization by threshold suggest that the results are credible and efficient to deploy for a real-world text-based plagiarism detection process. The results indicate that the similarity based methodologically based machine learning approaches are an improved alternative to the existing plagiarism detection, especially when the corpus used to train such similar statistical methods are standardized, as can be seen in the case of the PAN Plagiarism Corpus. The similarity of the system to diverse document structures and plagiarism patterns suggests its practical usefulness in an academic and educational context. Furthermore, since the structure of the framework is composed of modular components, it provides a scalable toolkit for extending detection beyond text-based recognition into images, diagrams, and source code. Even with these encouraging progressions, a set of limitations still exist. Our TF-IDF vectorization approach doesn't capture the deep semantic relationships and as a result, the system struggles with heavily paraphrased or conceptually rewritten content. In addition, the current evaluation mostly relies on textual plagiarism, although image-based and source-code plagiarism assessment are still some work in progress. Further work will combine transformer-based semantic models like BERT or Sentence-BERT to enhance the robustness of these models to advanced paraphrasing. To achieve fully unified multimodal plagiarism detection, further extensions include building CNN-based visual similarity models and graph-based program analysis techniques. Further evaluation with multilingual datasets and further investigating explainable similarity scoring mechanisms would enrich the transparency and adaptivity of the system. Overcoming these hurdles is important for creating resilient plagiarism detection systems that ensure academic integrity within a fast-paced digital and AI-enhanced content ecosystem.

REFERENCES

[1] A. Amirzhanov, C. Turan, and A. Makhmutova, "Plagiarism types and detection methods: A systematic survey of algorithms in text analysis," Frontiers in Computer Science, vol. 7, 2025
[2] M. Sajid, M. Sanaullah, M. Fuzail, T. S. Malik, and S. M. Shuhidan, "Comparative analysis of text-based plagiarism detection techniques," PLOS ONE, vol. 20, no. 4, 2025.
[3] V. Kuksa and M. Polyakov, "Developing and applying a neural network system for text plagiarism detection in higher education," in Proc. IEEE TELE, 2024.
[4] P. S. Kumar and P. B. M., "Integrated plagiarism detection system for text and image based content in documents," Journal of Information Systems Engineering and Management, vol. 10, 2025.
[5] P. Y. Reddy et al., "A deep learning approach for identifying LLM- generated text," in Proc. IEEE AIDE, 2025.
[6] A. S. Ahlawat et al., "Identifying AI-generated text designed to evade plagiarism detection," in Proc. IEEE ICAISS, 2025.
[7] S. Nualnim, M. Maliyaem, and H. Unger, "Plagiarism detection using text-representing centroids techniques," Indonesian Journal of Electrical Engineering and Computer Science, vol. 38, no. 3, 2025.
[8] M. F. Manzoor et al., "Exploring the landscape of intrinsic plagiarism detection: Benchmarks, techniques, evolution, and challenges," IEEE Access, vol. 11, 2023.
[9] R. K. Kodali, T. Shekhar, and L. Boppana, "Automated plagiarism detection in Moodle," in Proc. IEEE TENCON, 2023.
[10] Y. Li et al., "Content specialists' anti-plagiarism pedagogical interven- tions: A thematic review," Journal of Academic Ethics, vol. 23, 2025.
[11] G. M. Shipurkar, K. N. Shah, R. R. Sheth, R. Garg, T. A. Surana, and P. Natu, "End-to-end system for handwritten text recognition and plagiarism detection using CNN and BLSTM," in Proc. IEEE Int. Conf. Artificial Intelligence and Speech Technology (AIST), 2022.
[12] L. Shkurti, F. Kabashi, J. Ajdari, and V. Fus¸a, "PlagAL: Plagiarism detection system for Albanian texts," in Proc. IEEE Mediterranean Conf. Embedded Computing (MECO), 2021.
[13] S. Thaiprayoon, P. Palingoon, and K. Trakultaweekoon, "Design and development of a plagiarism corpus in Thai for plagiarism detection," in Proc. IEEE Int. Conf. Natural Language Processing, 2019.
[14] T. Folty´nek, N. Meuschke, and B. Gipp, "Academic plagiarism detection: A systematic literature review," ACM Computing Surveys, vol. 52, no. 6, Article 112, 2019.

[15] S. Awasthi, "Plagiarism and academic misconduct: A systematic review," DESIDOC Journal of Library & Information Technology, vol. 39, no. 2, pp. 94–100, 2019.
[16] S. E. Eaton and K. Crossman, "Self-plagiarism research literature in the social sciences: A scoping review," Interchange, vol. 49, no. 3, pp. 285–311, 2018.
[17] T. A. E. Eisa, N. Salim, and S. Alzahrani, "Figure plagiarism detection based on textual features representation," in Proc. IEEE Int. Conf. Information and Communication Technology, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY