IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

OFFLINE TEXT MATCHING AND PLAGIARISM DETECTION

Akansha Uike[1], Shweta Tayade[2], Ankita Gulde[3], Aakanksha Tiwari[4], Priyanka Patil[5]
[1]Student, [2]Student, [3]Student, [4]Assistant Professor, [5]Assistant Professor
Department of Computer Science and Engineering
Prof. Ram Meghe College of Engineering and Management, Amravati, India

Abstract: Plagiarism has emerged as a significant challenge in academic and content-driven environments due to the widespread availability of digital information. Most existing plagiarism detection tools operate as online services and depend on cloud-based databases, which raises concerns related to data privacy, subscription costs, and accessibility in low-connectivity environments. This paper presents the design and implementation of an offline plagiarism detection system using machine learning techniques. The proposed system performs text preprocessing, feature extraction using TF-IDF, and similarity computation using cosine similarity to identify potential plagiarism within a locally maintained document repository. The system generates detailed similarity reports highlighting overlapping content and providing similarity scores for interpretability. Experimental evaluation demonstrates that the proposed approach effectively detects direct and partial plagiarism in offline environments with acceptable performance and accuracy. The solution offers a privacy-preserving and cost-effective alternative to online plagiarism detection platforms, making it suitable for educational institutions and organizations with limited internet access.

Index Terms - Plagiarism Detection, Offline Systems, Machine Learning, TF-IDF, Cosine Similarity, Text Similarity, NLP

I. INTRODUCTION

The rapid digitization of academic resources and content creation platforms has increased the risk of plagiarism [5]. Students and content creators can easily copy, modify, or paraphrase material from existing sources. Ensuring originality is essential for maintaining academic integrity and content authenticity. Traditional manual plagiarism detection is time-consuming and unreliable when dealing with large document collections.

While several online plagiarism detection tools are widely used, their reliance on internet connectivity and third-party servers raises concerns related to privacy, cost, and data control [6]. Institutions in low-connectivity environments face accessibility issues, and confidential documents may not be suitable for upload to external platforms. These challenges motivate the development of offline plagiarism detection systems that operate locally and preserve data confidentiality.

This paper proposes an offline plagiarism detection system using machine learning-based text similarity techniques. The system processes documents locally, compares them with a local repository, and generates interpretable similarity reports. The approach aims to provide a practical, privacy-preserving solution suitable for real-world academic environments.

The novelty of the proposed system lies in its fully offline architecture, which ensures data privacy and eliminates dependency on internet connectivity. Unlike existing cloud-based plagiarism detection tools, the system provides a lightweight and cost-effective solution using TF-IDF and cosine similarity. This

makes it suitable for educational institutions, particularly in low-connectivity environments, while maintaining acceptable detection performance.

II. RELATED WORK

Existing plagiarism detection approaches can be broadly categorized into traditional text matching methods and machine learning-based similarity detection techniques. Traditional approaches rely on exact string matching, keyword overlap, and n-gram comparison [6]. These methods effectively detect direct copying but struggle with paraphrased content.

Machine learning and NLP techniques have enabled more robust detection through feature extraction methods such as TF-IDF and similarity measures such as cosine similarity [1]. Semantic similarity approaches using word embeddings and contextual models further improve detection of paraphrased plagiarism but require significant computational resources and are commonly deployed in cloud environments.

Most commercial plagiarism detection tools operate online and provide access to large web-scale databases. However, they introduce concerns related to data privacy, recurring subscription costs, and limited customization. Limited research focuses on practical, offline-capable plagiarism detection systems that balance accuracy with resource efficiency. This paper addresses this gap by presenting a lightweight, offline solution suitable for institutional deployment.

III. PROPOSED METHODOLOGY

The proposed system follows a modular offline workflow:

3.1 Data Collection and Local Repository

A local document repository is created using previously submitted assignments, reports, and reference documents. The repository is curated to ensure relevance and diversity.

3.2 Text Preprocessing

Documents undergo preprocessing, including tokenization, normalization, stop-word removal, and stemming/lemmatization. This reduces noise and standardizes text representation.

3.3 Feature Extraction

TF-IDF is used to convert text into numerical vectors [1]. Repository documents are vectorized in advance and stored locally for efficient reuse.

3.4 Similarity Computation

Cosine similarity measures the similarity between the input document vector and repository vectors [13]. Documents exceeding a predefined similarity threshold are flagged as potential plagiarism cases.

The system generates a detailed report with similarity percentages, ranked matched documents, and highlighted overlapping segments to support human interpretation.

IV. SYSTEM ARCHITECTURE AND IMPLEMENTATION

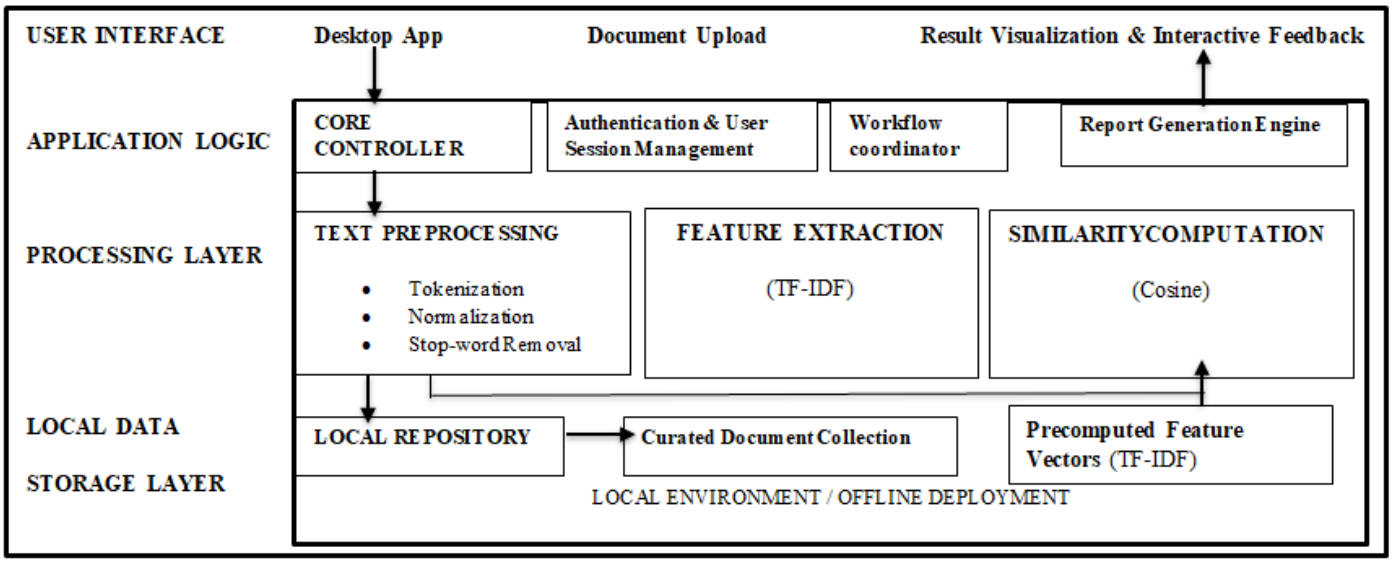

Fig. 1: System Architecture

As shown in Fig. 1, the system adopts a layered architecture comprising a user interface layer, application logic layer, processing layer, and local data storage layer. The implementation uses open-source tools to ensure cost-effectiveness and offline deployment. Feature vectors for repository documents are precomputed to optimize performance. The user interface supports document upload, result visualization, and report export.

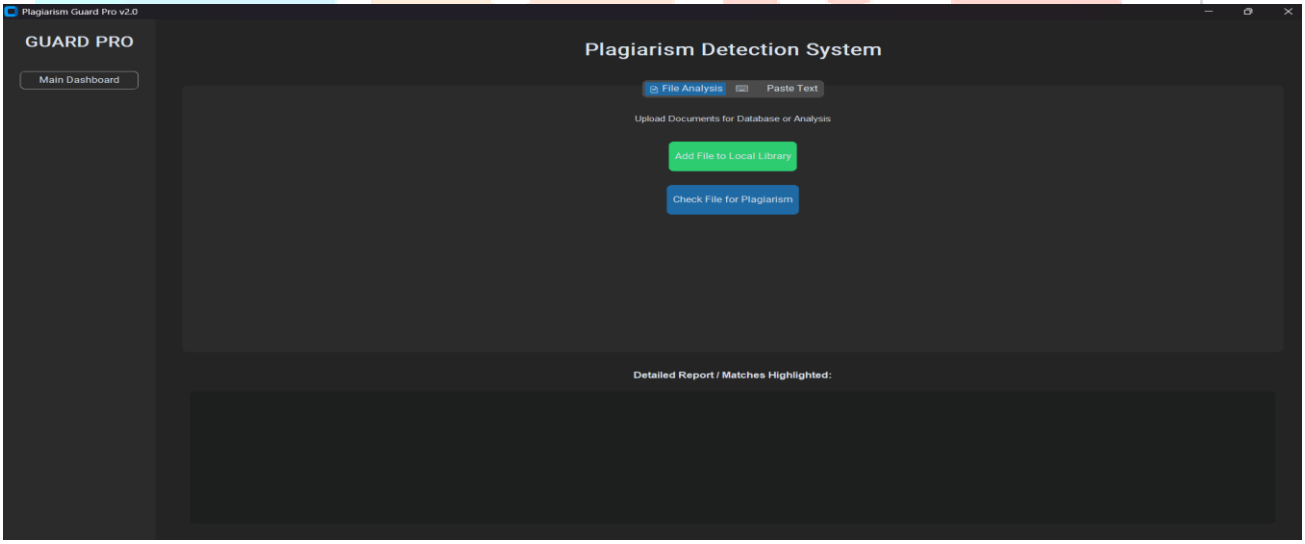

Fig. 2: User Interface

As shown in Fig. 2, the system is designed to be modular, allowing future enhancements such as semantic similarity models and multi-language support. All processing is performed locally to preserve data privacy and independence from external services.

V. EXPERIMENTAL EVALUATION

The system is evaluated using a test dataset containing approximately 25 documents, including original documents, direct copy-paste cases, partial copying, and paraphrased content. Performance metrics include processing time, detection accuracy, and false positive/negative rates.

Results indicate high accuracy for detecting direct and partial plagiarism. Performance remains acceptable for small to medium-sized repositories in offline environments. Detection of heavily paraphrased content is limited due to reliance on surface-level similarity features, highlighting opportunities for future enhancement.

VI. RESULTS AND DISCUSSION

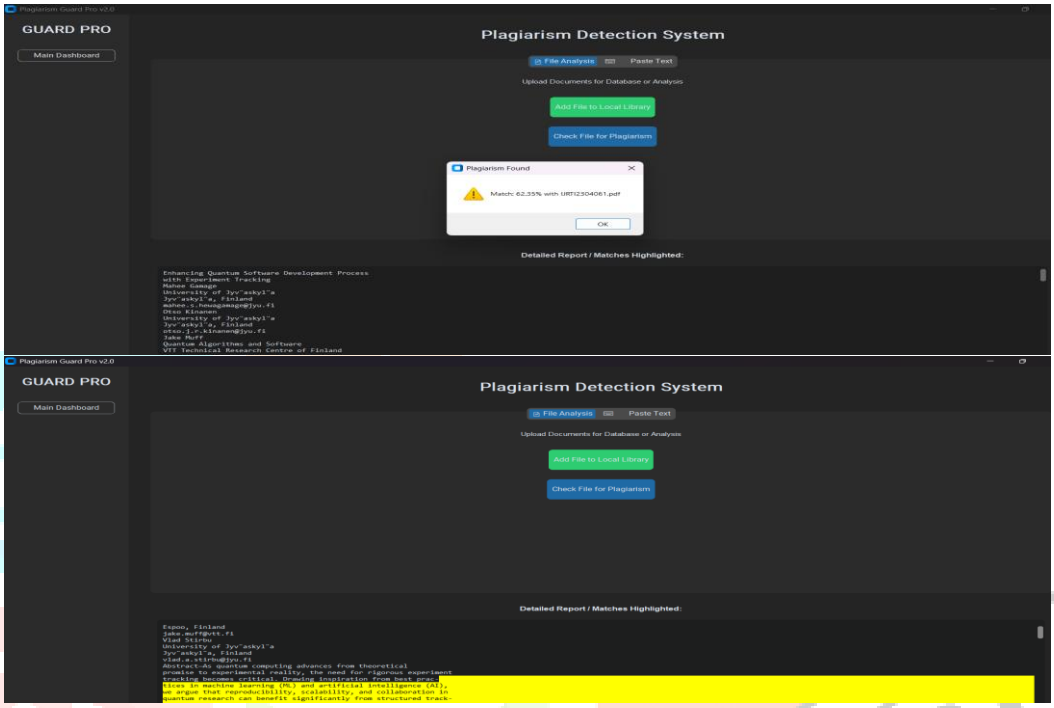

Fig. 3: Workflow of the System

As shown in Fig. 3, the experimental results demonstrate that TF-IDF and cosine similarity provide reliable detection of lexical overlap [14]. The offline architecture ensures data privacy and consistent availability. User testing indicates that the system is easy to use and integrates well into academic workflows. However, the lack of access to global online databases limits detection to the local repository. The system is best positioned as a decision-support tool rather than a complete replacement for online platforms or human judgment.

VII. ADVANTAGES AND LIMITATIONS

Advantages:

- Fully offline operation with strong data privacy
- Cost-effective using open-source tools
- Customizable local repository
- Interpretable similarity reports

Limitations:

- Limited detection of deep semantic plagiarism
- Repository coverage limits detection scope
- Performance constraints with very large repositories

VIII. FUTURE WORK

Future enhancements include integrating semantic similarity models for paraphrase detection, supporting multiple languages, extending detection to images and source code, and providing optional cloud synchronization for backup and collaboration. Mobile application integration can further improve accessibility.

IX. CONCLUSION

This paper presented an offline plagiarism detection system using machine learning-based text similarity techniques. The system effectively detects direct and partial plagiarism within a local repository while preserving data privacy and operating without internet connectivity. The proposed solution offers a practical and cost-effective alternative to online plagiarism detection tools, particularly suitable for institutions with connectivity or privacy constraints. Future enhancements can further improve detection accuracy and broaden system applicability.

X. REFERENCES

[1] Manning, C. D., Raghavan, P., & Schütze, H. (2009). *Introduction to Information Retrieval*. Cambridge University Press.

[2] Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed. draft). Pearson.

[3] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*.

[4] Potthast, M., Gollub, T., Rangel, F., & Stein, B. (2014). Improving the reuse detection of plagiarized content. *Proceedings of CLEF*.

[5] Clough, P. (2003). Old and new challenges in automatic plagiarism detection. *National Plagiarism Advisory Service Technical Report*.

[6] Gupta, D., & Lehal, G. S. (2009). A survey of text plagiarism detection techniques and tools. *Journal of Emerging Technologies in Web Intelligence*.

[7] Aggarwal, C. C., & Zhai, C. (2012). *Mining Text Data*. Springer.

[8] Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

[9] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

[10] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*.

[11] Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*.

[12] Robertson, S. E., & Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[13] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*.

[14] Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the 6th New Zealand Computer Science Research Student Conference*.

[15] Kondrak, G. (2005). N-gram similarity and distance. In *International Symposium on String Processing and Information Retrieval (SPIRE)*. Springer.