

# Plagiarism Detection Using Machine Learning

Priyanka, Basavashree, MD Saheil Pasha, Nikhil<sup>1</sup>  
Mrs. Sadhana.C<sup>2</sup>

<sup>1</sup>UG Students of Department of Electronics & Communication,

<sup>2</sup>Faculty of Department of Electronics & Communication,

Lingarajappa Engineering College Bidar, Karnataka.

Emails:sadhana@gmail.com

## ABSTRACT

Plagiarism has become a critical issue in academic and professional environments due to the easy availability of digital content. Traditional plagiarism detection techniques based on keyword matching and string similarity often fail to detect paraphrased and semantically similar content. This paper presents a machine learning-based plagiarism detection system that integrates natural language processing (NLP) techniques and cosine similarity for accurate detection of plagiarized content. The proposed system preprocesses text using tokenization, stop-word removal, and lemmatization, followed by feature extraction using TF-IDF vectorization. Online content is retrieved using Google Custom Search API and compared with user-submitted documents. Experimental results demonstrate that the proposed approach effectively identifies exact and paraphrased plagiarism with improved accuracy and efficiency. The system provides detailed plagiarism reports and visualizations, making it suitable for academic and professional use.

**Keywords—**Plagiarism Detection, Machine Learning, Natural Language Processing, TF-IDF, Cosine Similarity, Google Custom Search API.

## I. INTRODUCTION

Plagiarism is defined as the act of using another individual's work without proper acknowledgment. With the rapid growth of digital resources and online content, plagiarism has become increasingly prevalent in academic, creative, and professional domains. Traditional plagiarism detection approaches, such as manual review and simple string matching, are time-consuming and ineffective when dealing with paraphrased or restructured content. Machine Learning (ML) and Natural Language Processing (NLP) techniques provide advanced methods for analyzing semantic similarity between documents. These techniques enable the detection of hidden similarities beyond surface-level word matching. This paper proposes a plagiarism detection system using ML-based feature extraction and similarity measurement techniques to improve the detection of both exact and paraphrased plagiarism.

The main contributions of this work include:

- ◆ Development of an automated plagiarism detection framework using NLP and ML.
- ◆ Integration of Google Custom Search API for real-time web-based comparison.
- ◆ Use of TF-IDF and cosine similarity for accurate similarity measurement.
- ◆ Generation of detailed plagiarism reports with visualization.

## II. RELATED WORK

Several researchers have proposed different approaches for plagiarism detection. Sonawane and Prabhudeva [1] used Karp-Rabin and string matching algorithms to detect copied content. Chitra and Rajkumar [2] applied paraphrase recognition techniques to improve plagiarism detection at sentence and passage levels. Brainard and Seng [3] utilized Rabin-Karp and Jaro-Winkler distance algorithms to detect similarities in text documents.

Parker and Hamblen [4] discussed early computer algorithms for plagiarism detection to support academic integrity. Engels et al. [5] proposed feature-based neural networks using doc2vec embeddings for plagiarism detection. Awale et al. [6] applied machine learning techniques to detect plagiarism in programming assignments and achieved high accuracy using XGBoost.

Compared to existing approaches, the proposed system focuses on combining TF-IDF, cosine similarity, and online content retrieval for improved real-time plagiarism detection.

### III. PROPOSED METHODOLOGY

#### A. Data Collection

User-submitted documents are collected in text, PDF, or Word format. In addition, online documents are retrieved using Google Custom Search API for comparison.

#### B. Text Preprocessing

The preprocessing stage includes:

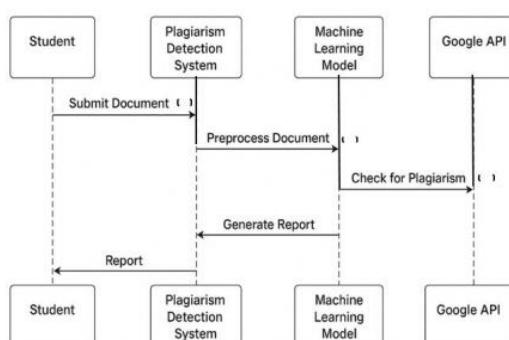
- Tokenization
- Lowercasing
- Stop-word removal
- Stemming/Lemmatization
- Removal of punctuation and special characters

#### C. Feature Extraction

TF-IDF (Term Frequency–Inverse Document Frequency) is used to convert text documents into numerical feature vectors.

#### D. Similarity Measurement

Cosine similarity is applied to measure the similarity between the TF-IDF vectors of user documents and retrieved online sources. A similarity threshold is used to classify documents as plagiarized or original.



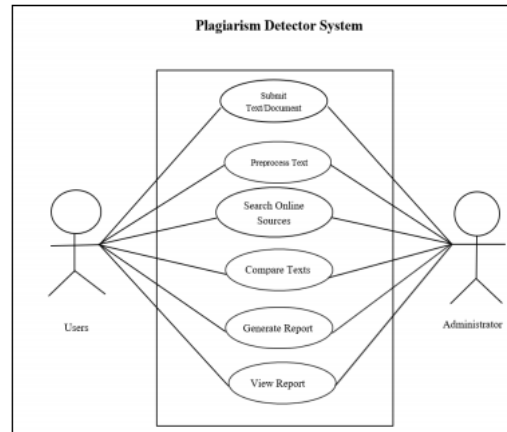
### IV. System Architecture

The system consists of the following modules:

- Document Upload Module
- Preprocessing Module

- Online Search Module (Google Custom Search API)
- Vectorization Module (TF-IDF)
- Similarity Computation Module (Cosine Similarity)
- Report Generation Module

## V. IMPLEMENTATION



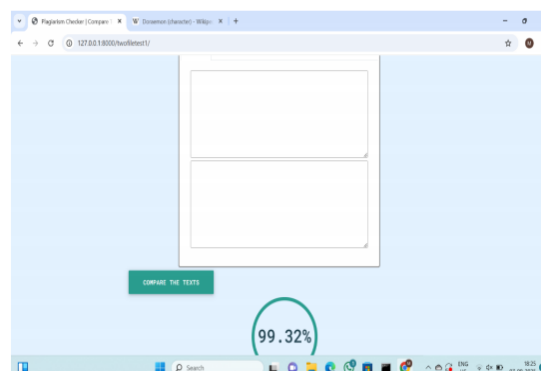
The system is implemented using Python and the following libraries:

1. Scikit-learn for TF-IDF and similarity computation
2. NumPy for numerical operations
3. Pandas for data handling
4. NLTK/spaCy for text preprocessing

Google Custom Search API is used to retrieve relevant online sources. The web interface is developed using HTML, CSS, and Bootstrap, and the backend is implemented using Django/Flask

## VI. RESULTS AND DISCUSSION

The system was tested using multiple documents containing both original and plagiarized content. The results show that the proposed system effectively detects exact and paraphrased plagiarism. The plagiarism percentage and matched sources are displayed in a detailed report along with visualizations. The experimental evaluation demonstrates improved detection accuracy compared to traditional keyword-based methods.



## VII. CONCLUSION

This paper presented a machine learning-based plagiarism detection system that combines NLP techniques, TF-IDF vectorization, cosine similarity, and online content retrieval. The system effectively detects both exact and paraphrased plagiarism and generates detailed plagiarism reports.

### VIII. FUTURE WORK

Future work includes integrating deep learning-based semantic models such as BERT, supporting multilingual plagiarism detection, and improving scalability for large document collections.

### REFERENCES

- [1] Ebrahim, F. and Joy, M. (2023). Source code plagiarism detection with pretrained model embeddings and automated machine learning. In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, pages 301–309.
- [2] Eppa, A. and Murali, A. (2022). Source code plagiarism detection: A machine intelligence approach. In 2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAEECC), pages 1–7. IEEE.
- [3] Cheers, H., Lin, Y., and Smith, S. P. (2021). Academic source code plagiarism detection by measuring program behavioral similarity. IEEE Access, 9:50391–50412.
- [4] Alan Parker and James O. Hamblen, "Computer Algorithms for Plagiarism Detection" in IEEE transaction on Education, Vol. 32, no. 2, 2020.
- [5] Chitra, A. and Rajkumar, Anupriya, "Plagiarism Detection Using Machine Learning Based Paraphrase Recognizer" Journal of Intelligent Systems, vol. 25, no. 3, 2016, pp. 351-359. <https://doi.org/10.1515/jisys-2014-0146>.
- [6] Kiran Shivaji, Sonawane S, Prabhudeva. (2015). Plagiarism Detection by using Karp Rabin and StringMatching Algorithm Together. International Journal of Computer Applications. 115. 37-41. <https://doi.org/10.5120/20294-2734>.