International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

# NLP-Based Plagiarism Detection using TF-IDF and Cosine Similarity System

**Aravind.L.Sino , Dr.T.R.NISHADAYANA**
PG Student, Department of Computer Science,Vels Institute of Science,Technology And Advanced Studies (VISTAS),Pallavaram, Chennai-600117,Tamil Nadu, India.
Assistant Professor, Department of Computer Science ,Vels Institute of Science,Technology And Advanced Studies (VISTAS),Pallavaram, Chennai-600117,Tamil Nadu, India.

**https://doi.org/**10.55041/ijsmt.v2i5.005

**Cite this Article**: Aravind.L.Sino, (2026). NLP-Based Plagiarism Detection using Tf-IDF and Cosine Similarity System. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i5.005



**ABSTRACT:** This project presents the development of a plagiarism detection system using Natural Language Processing (NLP) techniques. In the modern digital era, a vast amount of information is easily accessible through the internet, which makes copying and reusing content very common. Especially in academic environments, students and researchers may unintentionally or intentionally copy content from various sources. Detecting such plagiarism manually is a difficult, time-consuming, and error-prone process. Therefore, there is a strong need for an automated system that can efficiently identify similarities between documents and ensure originality.

The main aim of this project is to design and implement a system that can analyze text and detect plagiarism by comparing documents. The system uses TF-IDF (Term Frequency–Inverse Document Frequency), which helps in converting textual data into numerical vectors by identifying the importance of words in a document. After this, cosine similarity is used to measure the similarity between two or more text documents based on their vector representation. This approach provides an effective way to determine how closely related two documents are.

The system supports multiple input formats such as TXT, PDF, and DOCX files. It also allows users to directly enter text for comparison. Once the input is provided, the system extracts the text and performs preprocessing steps such as converting text to lowercase, removing special characters, eliminating stopwords, and applying lemmatization. These steps help in cleaning the text and improving the accuracy of similarity detection.

After preprocessing, the cleaned text is transformed into TF-IDF vectors, and cosine similarity is calculated between documents or sentences. The system then generates output in the form of similarity percentage, which indicates the level of plagiarism. It also highlights similar or plagiarized parts of the text, making it easier for users to identify copied content. In addition, the system provides visual representations such as graphs, charts, and heatmaps to give a better understanding of the similarity results.

Overall, this project offers a simple, efficient, and user-friendly solution for plagiarism detection. It can be used in educational institutions, research work, and content creation fields to maintain originality and integrity. Although the system has some limitations, such as difficulty in detecting paraphrased content,
it provides a strong foundation for further improvements using advanced machine learning and artificial intelligence techniques.

KEYWORDS: Plagiarism Detection, Natural Language Processing, TF-IDF, Cosine Similarity, Streamlit, Text Mining.

## I. INTRODUCTION

Plagiarism is the act of copying or using someone else's work, ideas, or content without proper acknowledgment and presenting it as one's own. It is considered a serious issue in academic, research, and professional fields, as it violates ethical standards and intellectual property rights. With the rapid growth of digital content and easy access to information through the internet, the chances of plagiarism have significantly increased.

Detecting plagiarism is very important to maintain academic integrity and originality in work. Educational institutions, researchers, and content creators rely on plagiarism detection systems to ensure that the submitted work is genuine and not copied from other sources. It also helps in promoting creativity and discouraging unethical practices such as copying assignments or research papers.

Manual plagiarism detection is a time-consuming and inefficient process, especially when dealing with large volumes of data. It requires comparing multiple documents line by line, which is not practical in real-world scenarios. Moreover, manual methods are prone to errors and may fail to identify hidden similarities or partial copying. Hence, there is a strong need for automated systems that can quickly and accurately detect similarities between documents.

To address this issue, the proposed system uses Natural Language Processing (NLP) techniques to detect plagiarism. The system analyzes text by converting it into numerical form using TF-IDF (Term Frequency–Inverse Document Frequency) and then calculates similarity using cosine similarity. It supports multiple file formats such as TXT, PDF, and DOCX, and provides results in the form of similarity percentage, highlighted text, and visual graphs. This makes the system efficient, user-friendly, and suitable for practical applications in academic and professional environments.

## II. RELATED WORK

The study of plagiarism detection has gained significant importance over the years due to the rapid growth of digital content and easy access to information. Various tools and techniques have been developed to identify similarities between documents and ensure originality. This section reviews existing plagiarism detection tools, the techniques used, and their limitations.

### ◆ Existing Plagiarism Detection Tools

Several plagiarism detection tools are currently available and widely used in academic and professional environments. Popular tools such as Turnitin, Grammarly, and Copyscape are designed to compare submitted content with large databases of web pages, research papers, and previously submitted documents.

These tools provide similarity reports, highlighting matched content and giving a percentage of similarity. They are commonly used in educational institutions to evaluate assignments and research papers. Some tools also provide suggestions for improving originality and proper citation.

However, many of these tools are paid services and may not be accessible to all users. Additionally, they often require internet access and depend on external databases for comparison.

### ◆ Techniques Used in Plagiarism Detection

Plagiarism detection systems use different techniques to compare and analyze text. The two main approaches are string matching and NLP-based techniques.

#### ✓ String Matching

String matching is one of the simplest methods used in plagiarism detection. In this approach, the system directly compares sequences of characters or words between documents. If identical or similar strings are found, it indicates possible plagiarism.

This method is easy to implement and works well for detecting exact copies of text. However, it is not effective when the text is slightly modified, such as changing words, sentence structure, or synonyms. Therefore, it has limited capability in detecting advanced forms of plagiarism.

#### ✓ NLP-Based Approaches

Natural Language Processing (NLP) techniques provide a more advanced and efficient way to detect plagiarism. Instead of directly comparing strings, NLP methods analyze the meaning and structure of the text.

Techniques such as tokenization, stopword removal, and lemmatization are used to preprocess the text. After preprocessing, methods like TF-IDF are used to convert text into numerical vectors, and cosine similarity is applied to measure similarity between documents.
NLP-based approaches are more flexible and can detect partial similarity even when the text is modified. They provide better accuracy compared to simple string matching methods and are widely used in modern plagiarism detection systems.

- **Limitations of Existing Systems**

Despite the availability of various tools and techniques, existing plagiarism detection systems have several limitations:

1. MANY TOOLS ARE **PAID** AND NOT ACCESSIBLE TO ALL USERS.
2. SOME SYSTEMS DEPEND ON **INTERNET CONNECTIVITY** AND EXTERNAL DATABASES.
3. DIFFICULTY IN DETECTING PARAPHRASED OR REWORDED CONTENT.
4. LIMITED ABILITY TO UNDERSTAND **CONTEXT AND SEMANTIC MEANING** OF TEXT.
5. MAY PRODUCE **FALSE POSITIVES** WHEN COMMON PHRASES ARE DETECTED AS PLAGIARISM.
6. HANDLING LARGE DOCUMENTS MAY REQUIRE MORE PROCESSING TIME AND RESOURCES.

## III. METHODOLOGY

1. THE PROPOSED PLAGIARISM DETECTION SYSTEM USES NATURAL LANGUAGE PROCESSING (NLP) AND MACHINE LEARNING TECHNIQUES TO IDENTIFY TEXTUAL SIMILARITY. INITIALLY, DOCUMENTS IN **TXT, PDF, AND DOCX** FORMATS ARE UPLOADED AND TEXT IS EXTRACTED USING FILE HANDLING LIBRARIES.

2. THE EXTRACTED TEXT UNDERGOES PREPROCESSING, INCLUDING LOWERCASE CONVERSION, REMOVAL OF SPECIAL CHARACTERS, STOP-WORD REMOVAL, TOKENIZATION, AND LEMMATIZATION TO IMPROVE TEXT QUALITY.

3. NEXT, THE CLEANED TEXT IS CONVERTED INTO NUMERICAL VECTORS USING **TF-IDF (TERM FREQUENCY–INVERSE DOCUMENT FREQUENCY)** FOR FEATURE EXTRACTION. SIMILARITY BETWEEN DOCUMENTS IS THEN MEASURED USING THE **COSINE SIMILARITY** ALGORITHM.

4. A THRESHOLD VALUE IS APPLIED TO IDENTIFY PLAGIARIZED CONTENT, AND THE PLAGIARISM PERCENTAGE IS CALCULATED BASED ON SIMILARITY SCORES. FINALLY, THE RESULTS ARE DISPLAYED THROUGH AN INTERACTIVE **STREAMLIT** INTERFACE WITH VISUALIZATIONS SUCH AS GAUGE METERS, HEATMAPS, AND RANKING CHARTS FOR BETTER ANALYSIS.

## IV. EXPERIMENTAL RESULTS

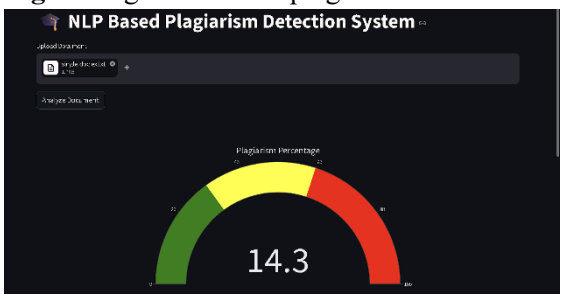
The application provides three main features:

1. Single document plagiarism detection
2. Multiple document comparison
3. Text-to-text similarity analysis

International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

**Fig 1**:Interface


**Fig 2:** Single document plagiarism detection




## Highlighted text

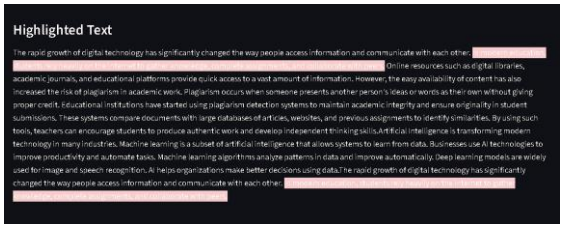
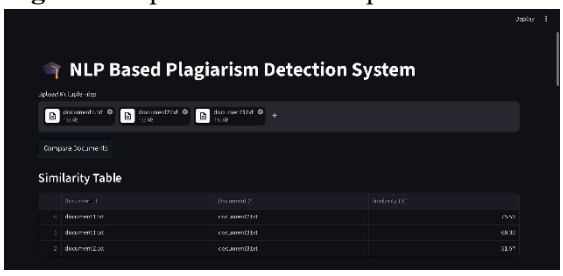
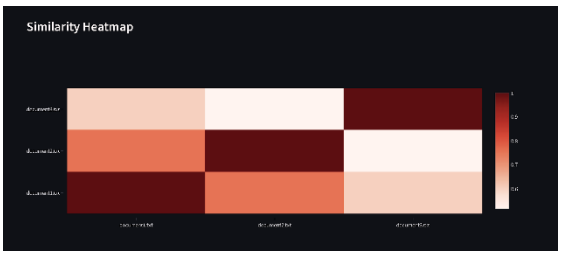



**Fig 3:** Multiple document comparison




## Heatmap


## Similarity Ranking

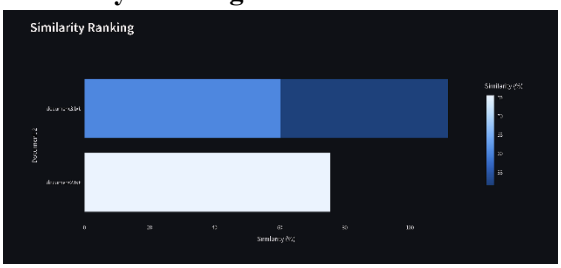
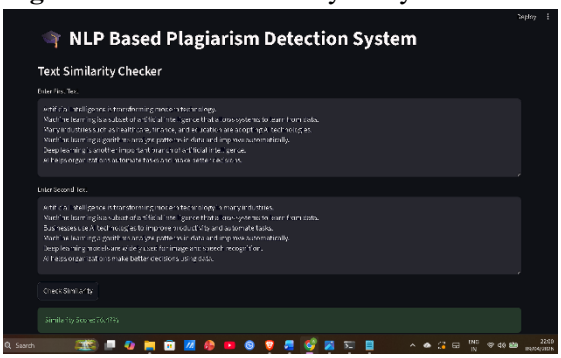




**Fig 4:** Text-to-text similarity analysis



## V. CONCLUSION

In this paper, an efficient and practical NLP-based plagiarism detection system has been proposed and implemented using TF-IDF vectorization and cosine similarity techniques. The system successfully addresses the growing need for automated plagiarism detection by providing a fast, reliable, and user-friendly solution for analyzing textual similarity.

The proposed approach effectively preprocesses textual data through tokenization, stopword removal, and lemmatization, ensuring that only meaningful information contributes to similarity computation. By representing documents as TF-IDF vectors, the system reduces the impact of commonly occurring words while emphasizing unique and informative terms. The use of cosine similarity further enables accurate measurement of similarity between documents, even when exact word matches are limited.

One of the key strengths of the system is its versatility. It supports multiple document formats such as TXT, PDF, and DOCX, and offers three different modes of analysis: single document plagiarism detection, multiple document comparison, and direct text-to-text similarity evaluation. This flexibility makes the system suitable for a wide range of applications, including academic submissions, research validation, and content originality checking.

Additionally, the integration of interactive visualizations such as heatmaps, similarity graphs, and gauge indicators enhances the interpretability of results. Users can easily understand the level of similarity and identify potentially plagiarized sections without requiring technical expertise.

Experimental results demonstrate that the system performs efficiently for small to medium-sized datasets, providing accurate similarity scores and meaningful insights. However, the system has certain limitations, particularly in detecting deeply paraphrased content and understanding semantic context beyond surface-level similarity.

Overall, the proposed system serves as a cost-effective and lightweight alternative to existing plagiarism detection tools. It demonstrates how classical NLP techniques can be effectively applied to real-world problems with minimal computational resources. Future enhancements involving semantic analysis, deep learning models, and web-based data integration can further improve the system's accuracy and scalability, making it more robust for large-scale and real-time applications.

In conclusion, this work contributes to the field of text mining and natural language processing by providing a simple yet powerful framework for plagiarism detection, with potential for further research and development in advanced similarity analysis techniques.

## REFERENCES

[1] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, 1988.

[2] Manning, C. D., Raghavan, P., & Schütze, H., "Introduction to Information Retrieval," Cambridge University Press, 2008.

[3] Bird, S., Klein, E., & Loper, E., "Natural Language Processing with Python," O'Reilly Media, 2009.

[4] Scikit-learn Documentation, https://scikit-learn.org

[5] NLTK Documentation, https://www.nltk.org