e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:08/Issue:04/April-2026 Impact Factor- 8.187 www.irjmets.com

CONTEXT-AWARE PLAGIRASIM DETECTION FOR PHARAPHRASED CONTENT

V.Visalakshi*1 Dr.V.Raghavendran*2

*1III BCA, Student, Vels Institute Of Science, Technology & Advanced Studies (Vistas),

*2Assistant Professor, Vels Institute Of Science, Technology & Advanced Studies (Vistas),

ABSTRACT

Plagiarism detection is an important task in academic and professional fields due to the increasing availability of digital content. Traditional systems mainly detect exact text matching but fail to identify paraphrased content where the meaning is retained but wording is changed.

This project proposes a Context-Aware Plagiarism Detection System that uses Natural Language Processing (NLP) techniques to analyze the semantic similarity between documents. The system performs preprocessing steps such as tokenization, stopword removal, and stemming, followed by feature extraction using TF-IDF. Similarity between texts is calculated using cosine similarity to determine the level of plagiarism.

The system provides a similarity percentage along with highlighted matched content, making it easier for users to understand the results. The proposed approach improves accuracy in detecting paraphrased plagiarism compared to traditional methods. This system can be effectively used in academic institutions and content verification platforms to ensure originality and maintain integrity.

I. INTRODUCTION

Plagiarism is a major problem in academic, research, and professional environments due to the rapid growth of digital content and easy access to online information. Many students and researchers copy content from various sources without proper citation, which affects originality and academic integrity.

Most traditional plagiarism detection systems are designed to detect only exact text matches using keyword comparison and string matching techniques. These systems work well for copy-paste plagiarism but fail to identify paraphrased plagiarism, where the content is rewritten using different words while keeping the same meaning.

To overcome this limitation, this project introduces a Context-Aware Plagiarism Detection System for Paraphrased Content. The system focuses on understanding the meaning and context of the text instead of just comparing words. It uses Natural Language Processing (NLP) techniques to analyze and process the text effectively.

The system includes several steps such as preprocessing, feature extraction, and similarity calculation. In preprocessing, unwanted words are removed and the text is cleaned using tokenization and stemming. Then, TF-IDF is used to convert text into numerical form. Finally, cosine similarity is applied to measure the similarity between documents.

This project aims to improve the accuracy of plagiarism detection by identifying both exact matches and paraphrased content. It can be used in educational institutions, research work, and content verification systems to ensure originality and maintain quality standards.

II. LITERATURE REVIEW

Plagiarism detection has been widely studied in recent years due to the increasing use of digital content. Many researchers and developers have proposed different techniques to identify copied or similar text in documents.

Traditional plagiarism detection systems mainly use string matching and keyword comparison methods. These systems compare the input text with a database and identify exact matches. Tools like Turnitin and other online checkers work based on this approach. While these methods are effective for detecting direct copy-paste plagiarism, they are not suitable for identifying paraphrased content.

Another commonly used approach is the N-gram method, where text is divided into small sequences of words. These sequences are compared between documents to find similarities. Although this method improves detection accuracy compared to simple matching, it still struggles when Sentence structure is changed.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:08/Issue:04/April-2026 Impact Factor- 8.187 www.irjmets.com

Some researchers have used semantic-based techniques such as WordNet to understand the meaning of words and identify similar content. These methods focus on relationships between words, such as synonyms, but they have limitations in handling complex sentence structures and large datasets.

Recent advancements include the use of machine learning and Natural Language Processing (NLP) techniques. These methods convert text into numerical form using models like TF-IDF and word embeddings, and then calculate similarity using cosine similarity. Such approaches provide better results in detecting paraphrased content compared to traditional systems.

However, existing systems still face challenges such as lower accuracy in complex paraphrasing, high computational cost, and difficulty in understanding deep context.

This project aims to overcome these limitations by using a context-aware approach, which focuses on semantic similarity and improves the detection of paraphrased plagiarism.

III. OBJECTIVE

The main objective of this project is to develop a system that can effectively detect plagiarism, including paraphrased content, by analyzing the meaning of the text.

The specific objectives are:

- To design and develop a context-aware plagiarism detection system
- To detect both exact and paraphrased plagiarism
- To use Natural Language Processing (NLP) techniques for text analysis
- To preprocess text using methods like tokenization, stopword removal, and stemming
- To convert text into numerical form using TF-IDF
- To calculate similarity between documents using cosine similarity
- To improve accuracy and reduce false positives and false negatives
- To provide a user-friendly interface with similarity percentage output
- To highlight plagiarized or similar content for easy understanding

This project aims to ensure originality and maintain academic integrity by Providing an efficient and reliable plagiarism detection system.

IV. METHODOLOGY

The proposed system follows a step-by-step process to detect plagiarism, especially paraphrased content, using context-aware techniques.

Data Collection

The system takes input documents from the user and compares them with reference documents stored in a dataset or database.

Text Preprocessing

In this stage, the input text is cleaned and prepared for analysis:

- Convert text into lowercase
- Remove punctuation and special characters
- Remove stopwords (common words like “the”, “is”)
- Perform tokenization (split text into words)
- Apply stemming or lemmatization to reduce words to their root form

Feature Extraction

The processed text is converted into numerical form using TF-IDF (Term Frequency–Inverse Document Frequency).

This helps in identifying important words in the document.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:08/Issue:04/April-2026 Impact Factor- 8.187 www.irjmets.com

Similarity Calculation

The system compares the documents using cosine similarity.

This method calculates how similar two texts are based on their vector representations.

Plagiarism Detection

A threshold value is set:

- If similarity is above the threshold → content is considered plagiarized
- If similarity is below the threshold → content is considered original

Result Generation

The system displays:

- Similarity percentage
- Highlighted plagiarized or paraphrased text
- Visual representation (like percentage graph)

V. ALGORITHM

Context-Aware Plagiarism Detection Algorithm

Input: Document A, Document B

Output: Similarity Percentage and Plagiarism Result

Step 1: Input Documents

- Read the input document (Document A)
- Read the reference document (Document B)

Step 2: Preprocessing

- Convert text to lowercase
- Remove punctuation and special characters
- Remove stopwords
- Perform tokenization (split into words)
- Apply stemming or lemmatization

Step 3: Feature Extraction

- Convert processed text into numerical vectors using TF-IDF
- Represent both documents as vector form

Step 4: Similarity Calculation

- Compute cosine similarity between the two vectors
- Obtain a similarity score between 0 and 1

Step 5: Decision Making

- Set a threshold value (e.g., 0.7 or 70%)
- If similarity score ≥ threshold → Mark as Plagiarized
- Else → Mark as Original

Step 6: Output Result

- Display similarity percentage
- Highlight similar or paraphrased text
- Show result in graphical format (percentage chart)

End

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:08/Issue:04/April-2026 Impact Factor- 8.187 www.irjmets.com

VI. HARDWARE AND SOFTWARE SPECIFICATIONS

Hardware Requirements

The system requires basic hardware components to run the application smoothly:

- Processor: Intel Core i3 or above
- RAM: Minimum 4 GB (8 GB recommended for better performance)
- Storage: 256 GB HDD/SSD
- System Type: Laptop or Desktop computer
- Input Devices: Keyboard and Mouse
- Output Devices: Monitor

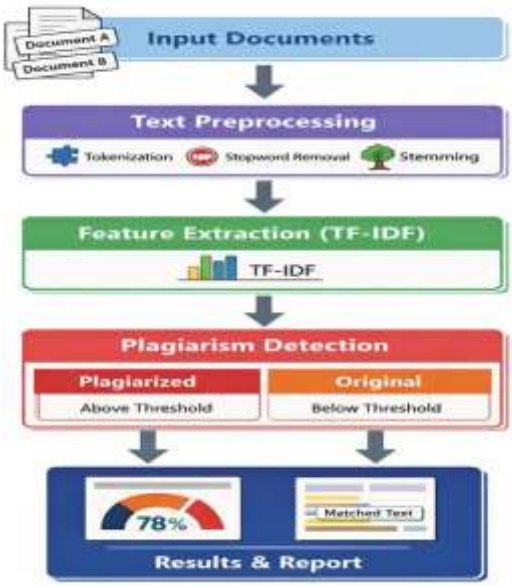
These hardware requirements are sufficient for running the plagiarism detection system and processing text data efficiently.

Software Requirements

The system is developed using the following software tools and technologies:

- Operating System: Windows / Linux / macOS
- Programming Language: Python
- Framework: Flask (for web-based interface)
- Libraries Used:
- NLTK (Natural Language Processing)
- Scikit-learn (TF-IDF and similarity calculation)
- NumPy (numerical operations)
- Pandas (data handling)
- Development Tools:
- Visual Studio Code / PyCharm (IDE)
- Database (Optional):
- SQLite or MongoDB (for storing documents)

VII. FLOWCHART

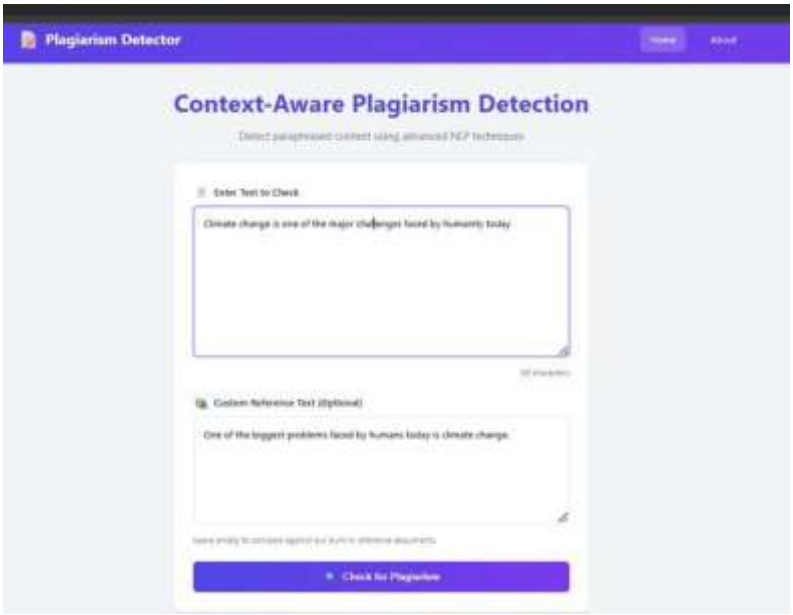


e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:08/Issue:04/April-2026 Impact Factor- 8.187 www.irjmets.com

The flowchart shows the step-by-step process of the plagiarism detection system. First, the user provides input documents. Then, the text is preprocessed by removing unwanted words and converting it into a clean format. Next, feature extraction is done using TF-IDF to convert text into numerical form. After that, cosine similarity is used to compare the documents. Based on the similarity score, the system decides whether the content is plagiarized or original. Finally, the system displays the result with a similarity percentage and highlighted text.

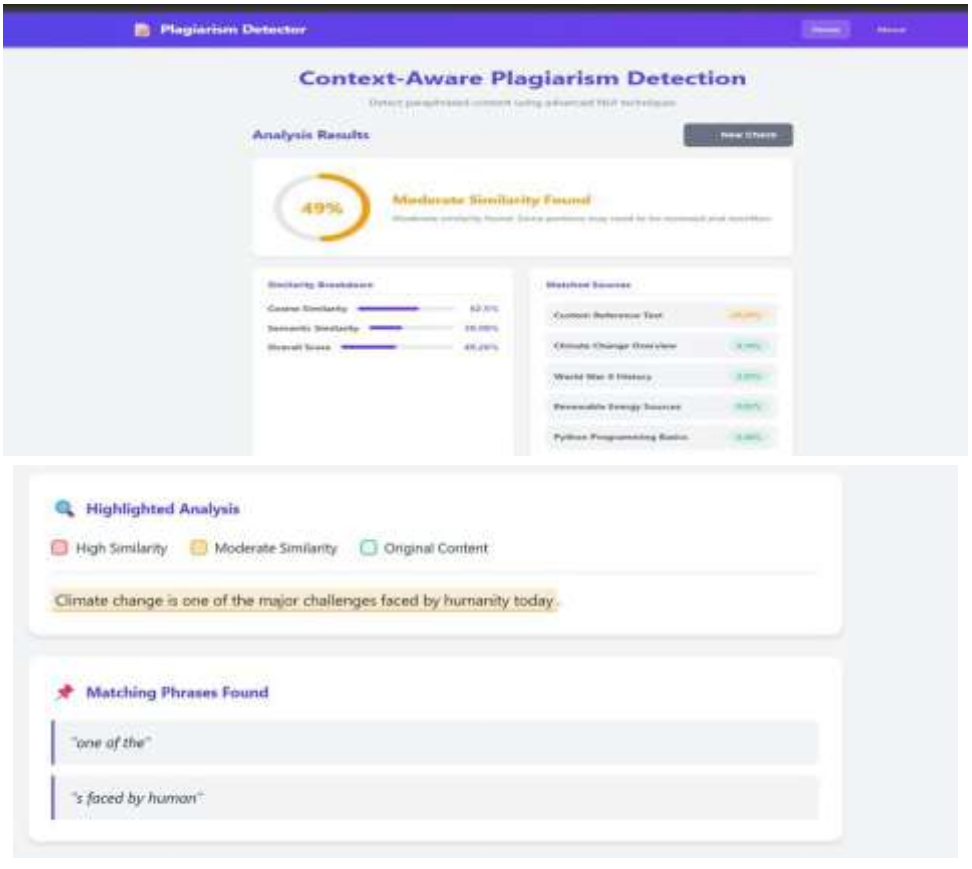
VIII. SYSTEM IMPLEMENTATION

8.1 USER INTERFACE DESIGN



This page represents the working interface of the plagiarism detection system

8.2 OUTPUT INTERFACE



IX. RESULTS

The proposed system was tested with various documents containing original, copied, and paraphrased content. The system successfully detected similarities and provided accurate results.

The output is shown as a similarity percentage, which indicates the level of plagiarism between documents. A higher percentage means more similarity, while a lower percentage indicates originality.

The system effectively identifies both exact matches and paraphrased content. It also highlights the similar portions of text, making it easier for users to understand the results.

The overall performance of the system shows improved accuracy and efficiency compared to traditional plagiarism detection methods.

X. CONCLUSION

The proposed Context-Aware Plagiarism Detection System successfully identifies both exact and paraphrased plagiarism by analyzing the meaning of the text. Unlike traditional systems, it focuses on semantic similarity using Natural Language Processing techniques.

The system uses preprocessing, TF-IDF, and cosine similarity to compare documents and generate accurate results. It provides a similarity percentage and highlights matched content, making it easy for users to understand the output.

Overall, the system improves accuracy, reduces errors, and offers a reliable solution for plagiarism detection. It can be effectively used in academic institutions and content verification systems to maintain originality and integrity.

XI. REFERENCES

[1] Salton Gerard, G., & McGill, M. J. (1986). Introduction to Modern Information Retrieval. McGraw-Hill.

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

[3] Jurafsky Daniel, D., & Martin James H., J. H. (2009). Speech and Language Processing. Pearson.

[4] Stein Benno, B., Lipka, N., & Prettenhofer, P. (2011). Intrinsic plagiarism analysis. Language Resources and Evaluation.