Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

# AI-Based Plagiarism Detection System

Faizan Shaikh
Master's Student Department of Computer Science
Abeda Inamdar Senior College
Pune, India

Shakila Siddavatam
Head of Department of Computer Science
Abeda Inamdar Senior College
Pune, India

**Abstract - Plagiarism has become a serious concern in academic, research, and digital content environments due to the rapid growth of online information and the ease with which content can be reused. Most existing plagiarism detection tools rely on traditional techniques such as keyword matching and surface-level text comparison. While these methods are effective in detecting direct copying, they often fail to identify deeper forms of plagiarism where the original meaning remains unchanged despite modified wording. This limitation reduces the reliability of plagiarism evaluation and raises concerns related to academic integrity. This study presents an AI-based plagiarism detection system that focuses on analyzing the meaning of text rather than matching exact words. A design and development research approach was followed to build a system capable of comparing text at the sentence level across multiple languages. The system represents sentences in a shared semantic space, enabling meaningful comparison of content without relying on translation-based techniques. Experimental results show that the proposed system effectively detects paraphrased and cross-language plagiarism that conventional tools are unable to identify. The system highlights similar sentences and generates measurable outputs such as plagiarism percentage and unique content percentage. Performance evaluation indicates that the system operates efficiently on standard computer hardware, making it suitable for use in academic institutions. The study concludes that meaning-based plagiarism detection provides more accurate and reliable results than traditional keyword-based approaches. Future work may include expanding the reference dataset and integrating online source comparison to further improve detection accuracy.**

***Keywords— Semantic Plagiarism Detection, Artificial Intelligence, Text Similarity, Multilingual Text Analysis, Sentence-Level Similarity Analysis***

## INTRODUCTION

In recent years, writing original academic content has become more difficult because most information is easily available online. Students and researchers regularly consult digital sources for understanding concepts, which sometimes results in ideas being reused without proper awareness of originality requirements. This has made plagiarism a frequent concern in academic work. Existing plagiarism detection tools mostly focus on matching words or sentence patterns. Such tools can identify copied text, but they struggle when the same idea is expressed in a different way. Content that is paraphrased, rearranged, or translated often passes undetected, even though the meaning remains unchanged. To handle this limitation, plagiarism detection needs to focus on understanding meaning instead of comparing words. With recent progress in artificial intelligence, it is now possible to analyze text at a semantic level. Based on this idea, the proposed work develops an AI-based plagiarism detection system that identifies semantic and cross lingual plagiarism more reliably.

Addressing the weaknesses of conventional plagiarism detection systems requires a shift from simple text matching toward deeper interpretation of content. Recent progress in artificial intelligence has enabled techniques that examine contextual meaning and relationships between words and sentences. Such approaches make it possible to recognize similarity even when ideas are rewritten, reorganized, or expressed in another language. Inspired by these advancements, this work proposes an AI-driven plagiarism detection system that emphasizes semantic understanding and cross-lingual comparison. The objective is to deliver accurate plagiarism identification while ensuring that the system remains efficient and suitable for everyday academic use on standard computing platforms.

### 1.1 Problem Statement

Checking the originality in today's world has become challenging because the same idea can be expressed in multiple ways. Writers can rewrite information after referring to different sources, which changes the wording but not the underlying meaning. In such cases, plagiarism may still exist even though the text does not appear identical. Most existing detection tools depend on the factors like word matching, Basic Text Comparison techniques, these techniques will fail when content is paraphrased or translated to other languages, As a result, Such tools will often generate inaccurate and unreliable reports. Therefore, there is a need for an efficient tool for plagiarism detection which can focus on meaning-based analysis, can detect multiple languages and also should be lightweight enough for the standard computing machines.

### 1.2 Significance

This study is significant as it addresses the limitation of traditional plagiarism detection tools in identifying paraphrased

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

content where the meaning remains unchanged. By focusing on meaning-based analysis rather than keyword matching, the proposed approach improves plagiarism detection accuracy and supports fair academic evaluation.

### 1.3 Proposed Solution

The proposed solution is an AI-based plagiarism detection system that analyzes text based on semantic meaning rather than exact word matching. The system compares sentences using meaning-level representations, allowing it to detect paraphrased and restructured content accurately. It also supports multilingual text analysis for cross-language plagiarism detection. The approach is designed to be lightweight and efficient, making it suitable for academic use on standard computer systems.

## 2.LITERATURE REVIEW

Roșu et al. proposed a plagiarism detection approach based on transformer-based deep learning models such as BERT and RoBERTa, focusing on sentence-level semantic representations to identify paraphrased plagiarism more accurately than traditional keyword-based methods [1]. Their results demonstrated improved detection performance; however, the approach required high computational resources, limiting its applicability in lightweight academic environments.

Bidgoli, Abdous, and Piroozfar introduced a hybrid method for cross-lingual semantic textual similarity using multilingual transformer models without relying on machine translation [2]. The proposed approach enabled effective similarity detection between sentences written in different languages and achieved high accuracy. Despite its effectiveness, the use of multiple transformer models increased system complexity, making real-time implementation challenging.

Another study by Vyas *et al.* presented an integrated semantic similarity and plagiarism checking system that combined semantic analysis, real plagiarism detection, online plagiarism detection, and textual entailment within a single framework [3]. The system improved overall plagiarism detection accuracy and provided detailed similarity analysis. However, the reliance on multiple models and features increased computational overhead, highlighting the need for a more efficient and simplified plagiarism detection solution.

### 2.1 Research Gap

Existing plagiarism detection studies demonstrate improved accuracy using semantic and transformer-based approaches, including cross-lingual analysis. However, most proposed systems are computationally complex and difficult to deploy in lightweight academic environments. There remains a research gap in developing an efficient, meaning-based plagiarism detection system that balances accuracy with practical usability on standard computing systems.

## 3. METHODOLOGY (DEVELOPMENT PROCESS)

### 3.1 Methodology

This study adopts a development-oriented research methodology to design and implement an AI-based plagiarism detection system. The proposed system focuses on identifying plagiarism based on semantic similarity rather than direct word matching. The overall process is organized into sequential stages, including text acquisition, preprocessing, semantic representation, similarity measurement, and result generation. This structured workflow ensures clarity, efficiency, and suitability for academic plagiarism detection tasks.

### 3.2 Input and Text Processing

The process begins with accepting textual input from the user for analysis. The input text undergoes preprocessing to eliminate unnecessary characters, redundant spacing, and formatting inconsistencies. Following preprocessing, the cleaned text is segmented into individual sentences to enable fine-grained analysis. Additionally, the language of the input text is automatically identified, allowing the system to handle multilingual content effectively.

### 3.3 Semantic Analysis

Once preprocessing is completed, each sentence is transformed into a vector-based representation that reflects its semantic meaning. This transformation is achieved using transformer-based sentence embedding techniques. By encoding contextual information, the system can recognize similarities between sentences even when different vocabulary or sentence structures are used, which is essential for detecting paraphrased plagiarism.

### 3.4 Similarity Evaluation and Result Generation

The generated semantic vectors are compared with reference content using cosine similarity to quantify the degree of similarity between sentences. A predefined similarity threshold is used to classify content as plagiarized or original. Based on these comparisons, the system computes plagiarism and originality percentages and generates a comprehensive plagiarism report that is presented to the user.

### 3.5 Database Methodology

The system includes a database component to support structured data management during plagiarism analysis. The database is used to store reference documents required for similarity comparison, as well as generated plagiarism reports and analysis history. This storage mechanism enables efficient retrieval of reference content and maintains result consistency

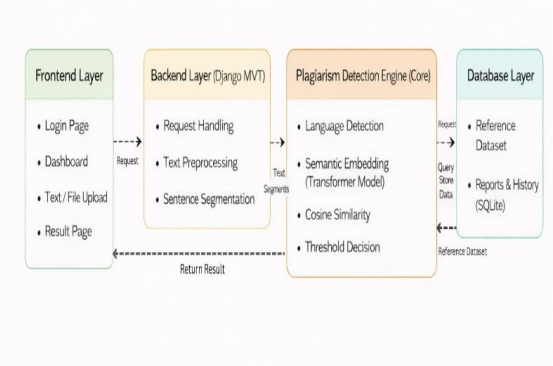
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

across multiple analyses. A lightweight database design is adopted to minimize system overhead and ensure smooth operation in standard academic computing environments.

Fig. 1. System Architecture of the Proposed Plagiarism Detection System

## 4 TECHNOLOGIES USED

The proposed plagiarism detection system is developed using a combination of web technologies and machine learning techniques to support semantic text analysis and system




functionality. The selected technologies facilitate text input handling, backend processing, semantic similarity computation, and data storage within a unified framework. These technologies are chosen to maintain system simplicity while enabling accurate meaning-based plagiarism detection. Table 1 presents the primary technologies employed at different levels of the proposed system.

Table 1. Technology Stack for AI Plagiarism Detection

| Category | Technology Used |
|---|---|
| Frontend Interface | HTML, CSS, JavaScript |
| Backend Framework | Django |
| Programming Language | Python |
| Semantic Modeling | Lightweight Transformer-based Sentence Embedding Model (MiniLM-based) |
| Similarity Measure | Cosine Similarity |
| Database | SQLite |

### 4.1 User Interface

The proposed plagiarism detection system provides a simple and user-friendly web-based interface to ensure ease of use for academic users. The interface is designed to operate efficiently on desktop and laptop systems, allowing users to perform plagiarism checks without requiring technical expertise. Emphasis is placed on simplicity, clarity, and smooth navigation so that users can easily submit content and view analysis results.

### 4.2 User Interface Overview

The system consists of the following interfaces:
**Registration Page**— Allows new users to create their account.
**Login Page--** Allows registered users to log in securely to access the plagiarism detection system.
**Dashboard Page—** Acts as the main control panel where users can navigate to different features of the system, such as text submission, File Upload, History, Statistics, Help.
**Text / File Upload Page—** Enables users to either paste text directly or upload supported files for plagiarism checking. This page collects the input data and forwards it to the backend for processing.
**Result Page—** Displays the plagiarism analysis results, including plagiarism percentage, unique content percentage, detected language, and highlighted plagiarized text for better understanding.

### 4.3 User InterFace And Screens

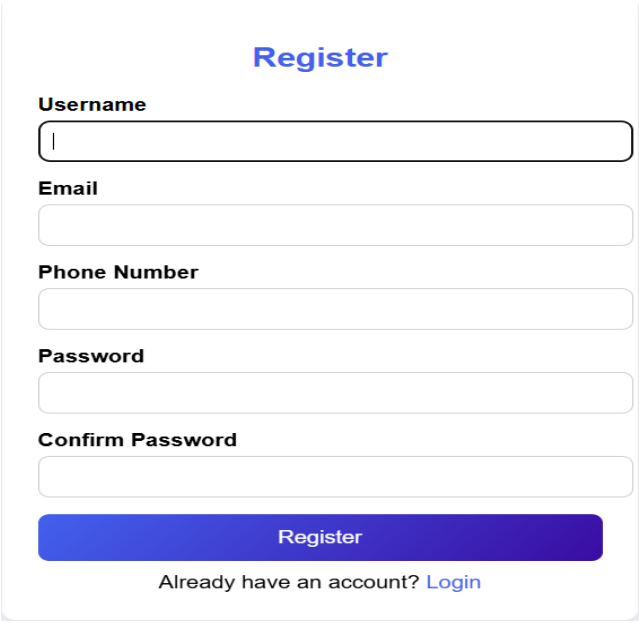



**Fig. 2. Registration Page**

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

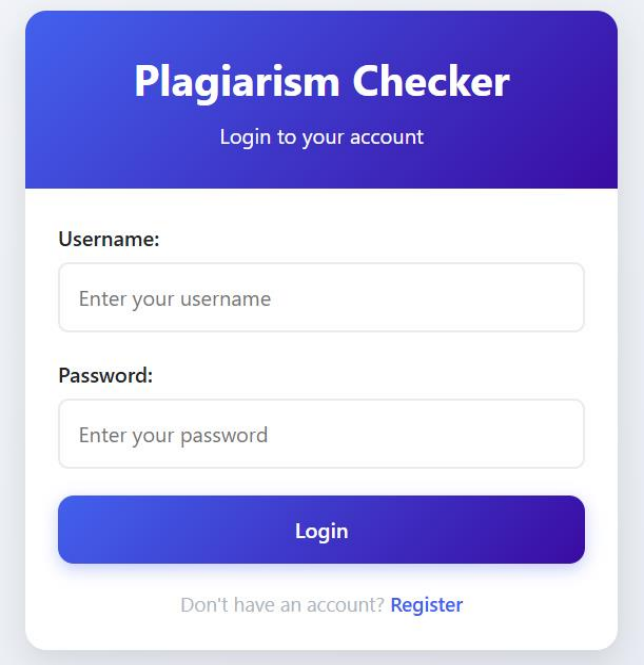



**Fig. 3. Login Page**

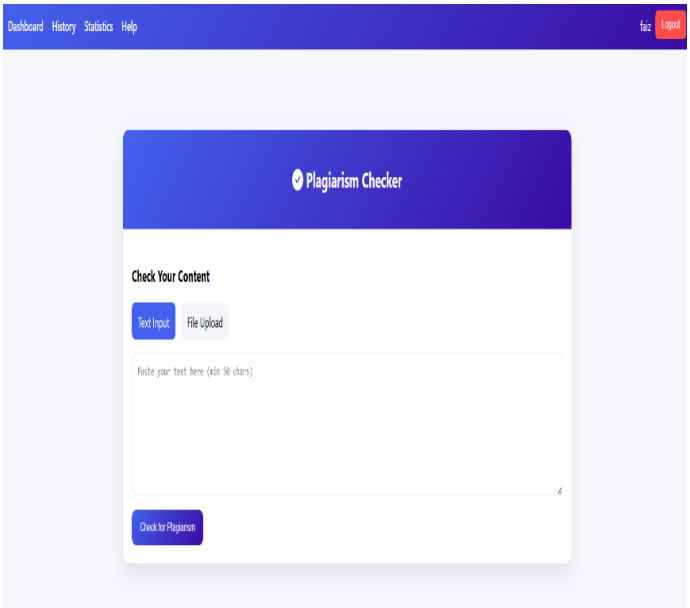



**Fig. 4. Dashboard**

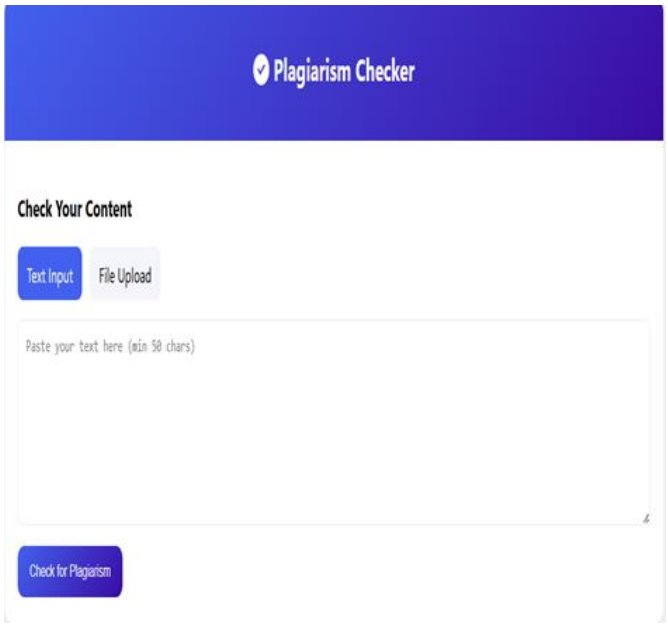



**Fig. 5. Upload Page**

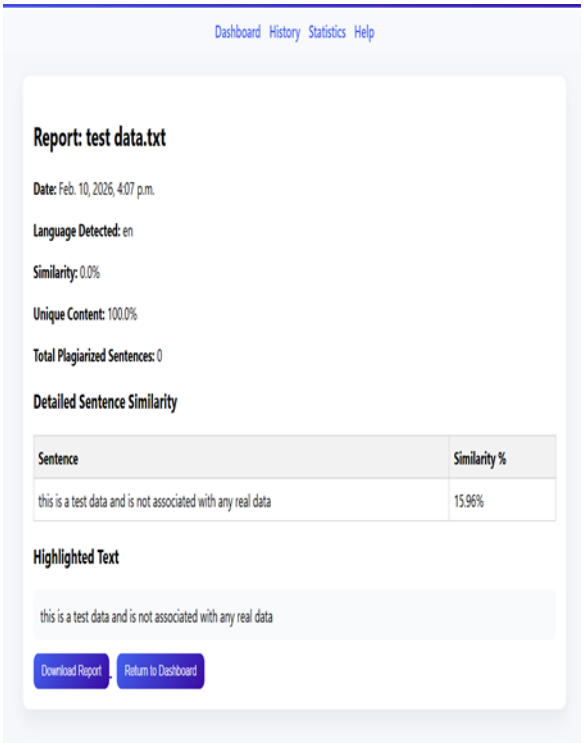



**Fig. 6. Result Page**

## 5. DISCUSSION

### 5.1 Strengths of the System

The proposed system checks plagiarism by understanding the meaning of sentences instead of only matching words. This

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

helps in identifying paraphrased content where the sentence structure is changed but the idea remains the same. The system is capable of handling text written in different languages, which makes it useful for detecting plagiarism in translated content that is usually ignored by traditional plagiarism detection tools. The plagiarism detection system is lightweight and can run on normal computer systems without requiring high processing power, making it suitable for use in colleges and academic institutions. The system analyzes text at the sentence level and highlights similar sentences, which makes the plagiarism results easy to understand and more transparent for users. Additionally, the simple user interface allows users to submit text or upload files easily and view plagiarism results without technical difficulty.

### 5.2 Challenges and Limitations

The effectiveness of the system depends on the availability and quality of reference text used for comparison. If relevant source content is not available, some cases of plagiarism may remain undetected. The plagiarism detection results also depend on the predefined similarity threshold, and selecting a very high or very low threshold may affect accuracy by leading to false positives or missed plagiarism cases. The current system mainly supports text-based inputs and converted documents, and direct handling of complex file formats may require additional preprocessing steps. Furthermore, the system does not perform real-time comparisons with online sources, as plagiarism detection is limited to the provided reference datasets or uploaded documents**.**

### 5.3 Future Scope

**Integration with Online and Academic Sources—** In future, the system can be extended to compare input text with online sources and academic repositories to improve plagiarism detection beyond locally stored reference content.

**Support for Advanced Models and File Formats—**
The system can be enhanced by using more advanced transformer models and by supporting additional file formats such as PDF and DOCX for wider academic use.

**Scalability and Platform Integration—**
Future versions may focus on cloud deployment and integration with learning management systems to support multi-user access and large-scale academic environments.

## 6. CONCLUSION

This research presented the design and implementation of an AI-based plagiarism detection system that focuses on identifying plagiarism through semantic similarity rather than exact word matching. In modern academic writing, plagiarism frequently appears in the form of paraphrasing or translated content, which reduces the effectiveness of traditional keyword-based detection techniques [1]. The proposed system addresses this challenge by analyzing text at the sentence level and comparing semantic meaning using transformer-based representations.

The system follows a structured workflow that includes text preprocessing, language identification, semantic embedding generation, and similarity evaluation. By adopting a meaning-based approach, the system is capable of detecting similarity even when sentence structure or wording is modified, thereby improving detection accuracy compared to conventional methods [2].

Another important contribution of this work is its lightweight and practical design, which allows the system to operate efficiently on standard computing environments. The simple web-based interface further supports ease of use in academic settings. Although the system currently relies on local reference datasets and fixed similarity thresholds, it demonstrates effective performance for academic plagiarism analysis. Overall, this research highlights the potential of semantic techniques to enhance plagiarism detection and support academic integrity [3].

## 7. REFERENCES

[1] R. Roșu, A. S. Stoica, P. S. Popescu, and C. M. Mihăescu, “NLP based deep learning approach for plagiarism detection,” *International Journal of User-System Interaction*, vol. 13, no. 1, pp. 48–60, Jan. 2020.

[2] B. Minaei Bidgoli, M. Abdous, and P. Piroozfar, “A hybrid method for cross-lingual semantic textual similarity,” *Research Square*, Jul. 2023.

[3] K. Vyas, A. Kumar, R. Banerjee, and D. B. Chakraborty, “Semantic similarity and plagiarism checker,” *Research Square*, Nov. 2023.