

Online Plagiarism Checking Using Data Mining Techniques

Mandar Patil , Prajwal Paulzagade, Varad Andhare,

Guide: Dr Anita Dapke Mahajan

Abstract-

Online Plagiarism Checker Using Data Mining Techniques is a web-based system designed to detect plagiarism in digital documents by applying advanced data mining and text similarity analysis techniques. The rapid growth of online content and digital submissions has increased the need for efficient tools that can identify copied or duplicated content accurately. This project provides an automated solution for detecting textual and document-level similarities among uploaded files.

The system allows users to upload documents in various formats, extracts text using document parsing and Optical Character Recognition (OCR) techniques, and analyzes the content using data mining algorithms such as SimHash and Winnowing. SimHash generates document fingerprints for efficient similarity comparison, while the Winnowing algorithm identifies matching text passages and determines the extent of plagiarism. The system calculates similarity percentages and generates detailed reports highlighting duplicated content.

Developed using modern web technologies, the application features a user-friendly interface, secure authentication, document management, and plagiarism analysis history. By combining data mining techniques with efficient text processing algorithms, the system provides fast, scalable, and accurate plagiarism detection. This solution can be used in educational institutions, research organizations, and content management platforms to ensure originality, maintain academic integrity, and reduce manual effort in plagiarism verification.

Keywords: Plagiarism Detection, Data Mining, SimHash, Winnowing Algorithm, OCR, Text Similarity Analysis, Document Fingerprinting, Academic Integrity.

I. INTRODUCTION

The widespread use of the internet and digital technologies has transformed the way information is created, shared, and accessed. Students, researchers, educators, and content creators can easily obtain vast amounts of information from online sources. While this accessibility promotes learning and knowledge sharing, it has also increased the risk of plagiarism, where individuals intentionally or unintentionally present someone else's work as their own without proper acknowledgment. Plagiarism undermines academic integrity, reduces originality, and can lead to ethical and legal issues.

Traditional methods of plagiarism detection involve manual comparison of documents, which is time-consuming, labor-intensive, and often ineffective when dealing with large volumes of content. As educational institutions and organizations handle thousands of documents regularly, there is a growing need for automated systems capable of detecting similarities quickly and accurately.

The **Online Plagiarism Checker Using Data Mining Techniques** is developed to address this challenge by providing an efficient platform for identifying duplicated or copied content in digital documents. The system utilizes data mining techniques and text analysis algorithms to compare uploaded documents with existing content and determine the degree of similarity. By extracting meaningful patterns from textual data, the system can identify matching phrases, sentences, and document structures that indicate potential plagiarism.

The proposed system incorporates advanced algorithms such as **SimHash** and **Winnowing** to perform document fingerprinting and similarity analysis. SimHash helps in generating compact representations of documents for rapid comparison, while Winnowing identifies exact and near-exact matching text segments. Additionally, Optical Character Recognition (OCR) is employed to extract text from scanned documents and image-based PDFs, ensuring comprehensive plagiarism detection across different document formats.

The application is designed as a web-based platform with features such as user authentication, document upload, plagiarism analysis, report generation, and history management. The system provides users with detailed similarity reports that highlight matched content and present plagiarism percentages, enabling informed decision-making.

By combining data mining methodologies with efficient text processing techniques, the project offers a reliable and scalable solution for maintaining originality in academic, research, and professional environments. The system contributes to improving content authenticity, promoting ethical practices, and reducing the effort required for manual plagiarism checking.

II. SYSTEM ARCHITECTURE

Overview

The **Online Plagiarism Checker Using Data Mining Techniques** follows a multi-layered architecture designed to perform document collection, text extraction, similarity analysis, and report generation efficiently. The system integrates document processing techniques with data mining algorithms to identify duplicated content and provide accurate plagiarism detection results. The architecture consists of five major layers: User Interface Layer, Application Layer, Text Processing Layer, Similarity Analysis Layer, and Data Storage Layer.

System Architecture Description

1. User Interface Layer

This layer serves as the interaction point between users and the system. Users can register, log in, upload documents, initiate plagiarism checks, and view generated reports through a web-based interface. The interface ensures ease of use and secure access to system functionalities.

2. Application Layer

The application layer manages user requests and coordinates communication among different modules. It handles authentication, document upload management, report generation, and workflow control. This layer ensures that documents are properly routed to the processing and analysis components.

3. Text Processing Layer

Once a document is uploaded, the text processing module extracts textual content from various file formats. For scanned documents and image-based PDFs, Optical Character Recognition (OCR) technology converts images into machine-readable text. The extracted text is then cleaned and preprocessed by removing unnecessary symbols, formatting inconsistencies, and redundant spaces.

4. Similarity Analysis Layer

This layer forms the core of the plagiarism detection system. Data mining and document fingerprinting techniques are applied to identify similarities between documents.

- **SimHash Algorithm:** Generates compact fingerprints of documents and performs rapid similarity comparisons.
- **Winnowing Algorithm:** Creates document fingerprints based on text substrings and identifies matching passages.
- **Similarity Calculator:** Computes similarity percentages and determines the extent of plagiarism.

The combined use of these techniques improves both detection accuracy and computational efficiency.

5. Data Storage Layer

The storage layer maintains user information, uploaded documents, extracted text, generated fingerprints, and plagiarism reports. The database supports efficient retrieval and comparison of documents for future plagiarism checks.

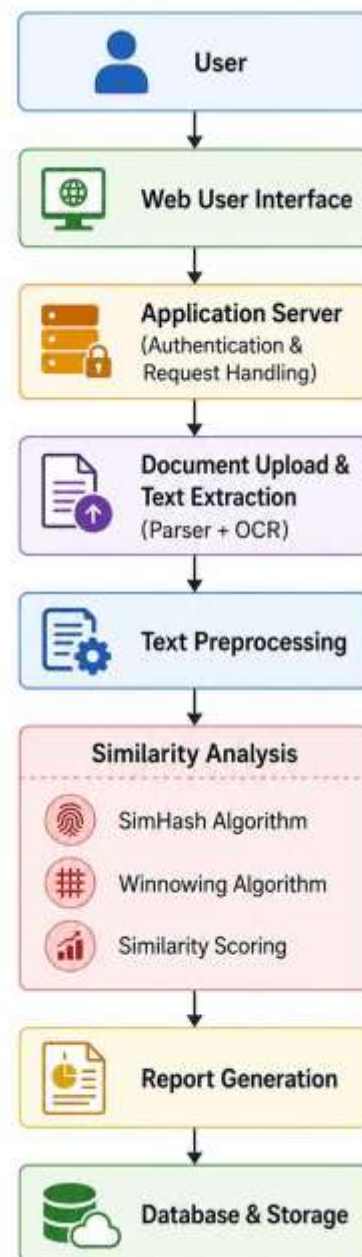
Advantages of the Architecture

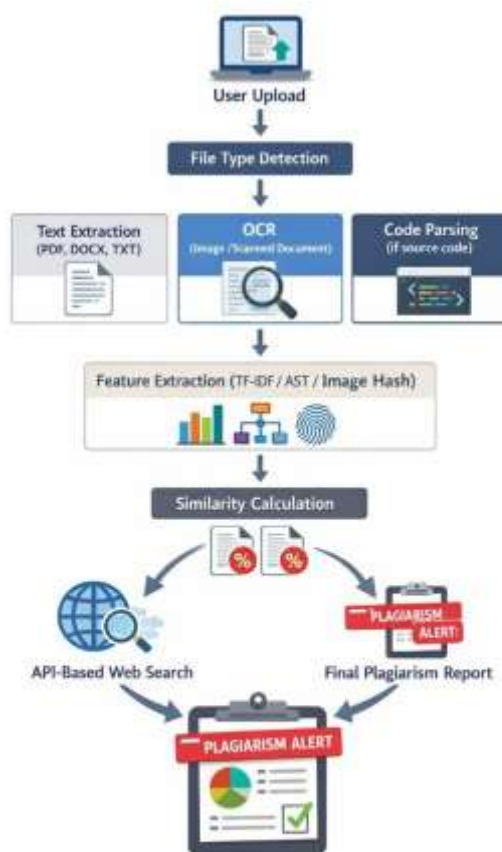
- Modular and scalable design.
- Fast document comparison through fingerprinting techniques.
- Support for scanned and image-based documents.

- Accurate identification of exact and near-duplicate content.
- Secure storage of documents and reports.
- Easy integration with educational and research platforms.

This architecture enables the **Online Plagiarism Checker Using Data Mining Techniques** to deliver reliable, efficient, and scalable plagiarism detection for academic, research, and professional applications.

Architectural Diagram





IV. LITERATURE REVIEW

The increasing availability of digital content has made plagiarism detection an important area of research in academic, professional, and publishing environments. Researchers have proposed various approaches ranging from simple string matching techniques to advanced data mining and machine learning methods for identifying copied or duplicated content. This section reviews existing studies and technologies related to plagiarism detection and highlights their contributions and limitations.

Early plagiarism detection systems primarily relied on exact text matching methods. These systems compared documents character by character or word by word to identify duplicated content. Although effective for detecting direct copying, they often failed to recognize paraphrased or modified text. Furthermore, their computational complexity increased significantly when processing large document collections. To overcome these limitations, researchers introduced fingerprinting techniques that generate compact representations of documents. Document fingerprinting reduces storage requirements and improves comparison speed. One widely used fingerprinting method is the **Winnowing algorithm**, which creates fingerprints from selected text fragments and identifies overlapping content between documents. The algorithm has gained popularity due to its ability to detect copied passages while maintaining efficiency in large-scale document repositories. Another significant advancement in plagiarism detection is the use of **SimHash**, a locality-sensitive hashing technique originally developed for detecting near-duplicate web pages.

SimHash converts textual content into fixed-length fingerprints, allowing rapid similarity comparisons between documents. Researchers have demonstrated that SimHash effectively identifies documents with substantial textual overlap while minimizing computational overhead.

Several commercial plagiarism detection tools, such as Turnitin, Grammarly Plagiarism Checker, and Copyscape, utilize large document databases and similarity matching techniques to detect copied content. While these systems provide reliable results, many are subscription-based and do not disclose their internal detection mechanisms. This has encouraged researchers to develop open and customizable plagiarism detection solutions that can be adapted to specific organizational requirements.

Recent studies have explored the application of **data mining techniques** for plagiarism detection. Data mining enables the extraction of meaningful patterns and relationships from large volumes of textual data. By combining text preprocessing, document fingerprinting, and similarity analysis, data mining-based systems can improve detection accuracy and handle extensive document collections efficiently. These approaches are particularly useful in educational institutions where large numbers of assignments and research papers require evaluation.

The integration of **Optical Character Recognition (OCR)** technology has further enhanced plagiarism detection capabilities. OCR enables systems to extract text from scanned documents and image-based files, expanding the range of documents that can be analyzed. This feature addresses a common limitation of traditional systems that process only machine-readable text.

Although existing approaches have significantly improved plagiarism detection, challenges remain in detecting paraphrased content, handling multilingual documents, and maintaining high accuracy with growing datasets. The proposed **Online Plagiarism Checker Using Data Mining Techniques** addresses these challenges by combining OCR-based text extraction, SimHash fingerprinting, Winnowing-based pattern matching, and similarity scoring within a unified web-based platform. This combination provides efficient document processing, accurate plagiarism identification, and scalability for academic and research applications.

The literature indicates that combining document fingerprinting, data mining techniques, and OCR-based text extraction can significantly improve plagiarism detection performance. Therefore, these concepts form the foundation of the proposed **Online Plagiarism Checker Using Data Mining Techniques**.

V. MATHEMATICAL MODEL

Overview

The **Online Plagiarism Checker Using Data Mining Techniques** employs a mathematical model to represent documents, analyze textual patterns, and measure similarities between documents. The model transforms unstructured textual data into a structured form that can be efficiently processed by plagiarism detection algorithms. By applying document fingerprinting and similarity analysis techniques, the system identifies duplicated content and generates an overall plagiarism score.

The mathematical model serves as the foundation of the plagiarism detection process and ensures that comparisons are performed systematically and accurately. It provides a framework for document representation, feature extraction, similarity computation, and result generation.

4. MATHEMATICAL MODEL

The Online Plagiarism Checker Using Data Mining Techniques models the plagiarism detection process as a series of mathematical operations on textual data. The model includes text representation, document fingerprinting using Simhash, similarity computation, and passage matching using the Winnowing algorithm.

4.1 Set Definitions

Let the system contain a set of documents against which the uploaded document is compared.

$$\begin{aligned} D &= \{d_1, d_2, \dots, d_n\} & D &= \text{Set of documents in the database} \\ Q &= q & Q &= \text{Query document uploaded by the user} \\ T &= \{t_1, t_2, \dots, t_m\} & T &= \text{Set of all unique terms after preprocessing} \\ W &= \{w_1, w_2, \dots, w_k\} & W &= \text{Set of weighted features (terms)} \end{aligned}$$

4.2 Term Weighting Model

Each term is assigned a weight using the TF-IDF scheme:

$$\begin{aligned} TF(t_i, d_j) &= \frac{f_{i,j}}{\sum_{t \in T} f_{t,j}} & f_{i,j} &= \text{Frequency of term } t_i \text{ in document } d_j \\ IDF(t_i) &= \log\left(\frac{N}{df(t_i)}\right) & N &= \text{Total number of documents} \\ w_{i,j} &= TF(t_i, d_j) \times IDF(t_i) & df(t_i) &= \text{Number of documents containing } t_i \end{aligned}$$

4.3 Document Vector Representation

Each document is represented as a weighted vector:

$$\begin{aligned} \vec{V}_j &= [w_{1,j}, w_{2,j}, \dots, w_{k,j}] & \vec{V}_j &= \text{Weighted vector of document } d_j \\ \text{where } j &= 1, 2, \dots, n & k &= \text{Total number of unique terms} \end{aligned}$$

4.4 Simhash Fingerprint Generation

Simhash generates a fixed-length fingerprint for each document:

$$\begin{aligned} S(d_j) &= \sum_{i=1}^k w_{i,j} \times h(t_i) & h(t_i) &= \text{Hash function for term } t_i, \text{ producing } +1 \text{ or } -1 \\ s_i(d_j) &= \begin{cases} 1, & \text{if } S_i(d_j) \geq 0 \\ 0, & \text{if } S_i(d_j) < 0 \end{cases} & S_i(d_j) &= i^{\text{th}} \text{ component of the summed vector} \\ FP(d_j) &= [s_1, s_2, \dots, s_L] & s_i(d_j) &= i^{\text{th}} \text{ bit of the Simhash fingerprint} \end{aligned}$$

4.5 Document Similarity using Simhash

Similarity between query document Q and document d_j is computed using Hamming distance:

$$\begin{aligned} HD(Q, d_j) &= \sum_{i=1}^L |FP(Q)_i \oplus FP(d_j)_i| & HD(Q, d_j) &= \text{Hamming distance between fingerprints} \\ \text{Sim}_{sim}(Q, d_j) &= 1 - \frac{HD(Q, d_j)}{L} & FP(Q) &= i^{\text{th}} \text{ bit of query and document fingerprints} \end{aligned}$$

4.6 Winnowing for Passage Matching

Used to find similar passages between documents:

$$\begin{aligned} g_i &= \text{hash}(w_1, w_2, \dots, w_{k-i+1}) & k &= \text{Window size (number of terms in a fingerprint)} \\ G_i &= \min\{g_i, g_{i+1}, \dots, g_{i+\beta-1}\} & \beta &= \text{Window step} \\ FP_{win}(d_j) &= \{G_1, G_2, \dots, G_n\} & G_i &= \text{Minimum hash value in the current window} \end{aligned}$$

4.7 Overall Similarity Score

Final plagiarism score is computed by combining Simhash similarity and Winnowing similarity:

$$\begin{aligned} \text{Sim}_{total}(Q, d_j) &= \alpha \times \text{Sim}_{sim}(Q, d_j) + \beta \times \text{Sim}_{win}(Q, d_j) & \text{Sim}_{win}(Q, d_j) &= \text{Similarity using Winnowing} \\ \text{Plagiarism (\%)} &= \text{Sim}_{total}(Q, d_j) \times 100 & \alpha, \beta &= \text{Weighting factors, } \alpha + \beta = 1 \end{aligned}$$

Input Representation

The system accepts digital documents uploaded by users. Each uploaded document is treated as a collection of textual information that must be analyzed. Before comparison, the document undergoes preprocessing operations such as text extraction, tokenization, normalization, and removal of irrelevant characters.

The processed text is then converted into a set of meaningful terms that represent the content of the document. These terms form the basis for further analysis and comparison.

Feature Extraction

Feature extraction is performed to identify important textual components within a document. The system analyzes the occurrence and significance of words and phrases to generate a representative feature set.

This stage reduces the complexity of the original document while preserving its essential information. The extracted features are used to create document fingerprints and support efficient similarity measurement.

Document Fingerprinting

Document fingerprinting is a technique used to create a compact representation of a document. Instead of comparing

entire documents directly, the system generates unique fingerprints that capture the characteristics of the content.

The proposed system utilizes fingerprinting methods to reduce computational overhead and improve processing speed. Fingerprints enable rapid identification of documents that share similar textual patterns and provide an efficient mechanism for large-scale document comparison.

Similarity Measurement

The similarity measurement component evaluates the degree of resemblance between the uploaded document and documents stored in the database.

The system compares generated fingerprints and textual patterns to determine whether significant overlap exists between documents. Similarity values are calculated and converted into percentage scores that indicate the likelihood of plagiarism.

Higher similarity scores suggest a greater amount of shared content, while lower scores indicate a higher degree of originality.

Passage-Level Matching

In addition to overall document comparison, the system performs detailed passage-level analysis. This process identifies specific sections of text that appear similar across documents.

By examining textual fragments and matching patterns, the system highlights potentially copied content. This allows users to locate duplicated passages and understand the source of detected similarities.

Passage-level matching improves transparency and provides more meaningful plagiarism reports.

Plagiarism Score Generation

The final plagiarism score is obtained by combining results from multiple analysis stages. Document-level similarity and passage-level similarity are integrated to produce a comprehensive assessment of content originality.

The generated score represents the percentage of content that matches existing documents within the system. Based on this score, users can evaluate whether a document contains significant duplication or maintains acceptable originality.

Output of the Mathematical Model

The mathematical model produces the following outputs:

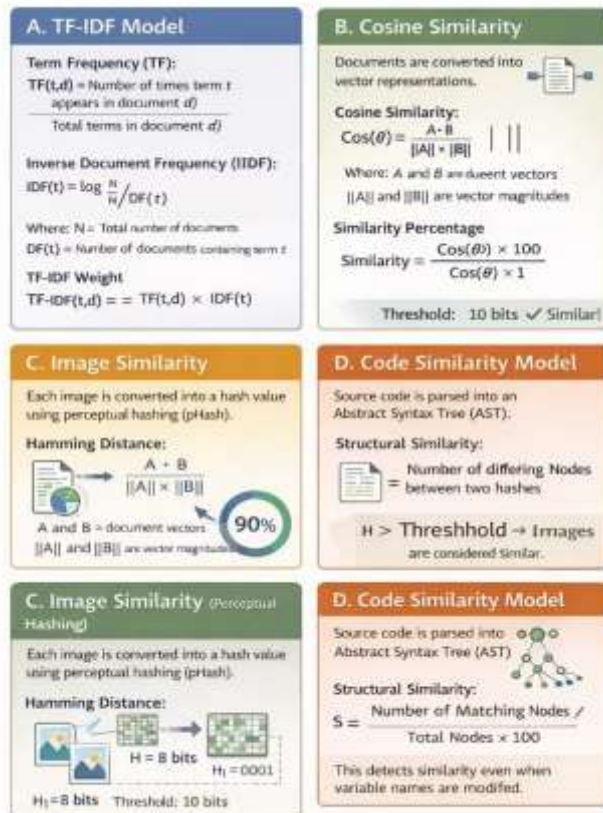
- Similarity percentage between documents.
- Identification of matching passages.
- Plagiarism assessment report.
- Document fingerprint information.
- Ranked list of similar documents.
- Visual representation of detected similarities.

These outputs help users understand the extent of content duplication and support informed decision-making regarding document authenticity.

Conclusion

The mathematical model provides a structured approach for detecting plagiarism through document representation, feature extraction, fingerprint generation, similarity measurement, and plagiarism scoring. By integrating data mining techniques with efficient document comparison methods, the model enables accurate, scalable, and reliable

plagiarism detection. This framework forms the analytical core of the **Online Plagiarism Checker Using Data Mining Techniques** and supports effective evaluation of document originality in academic, research, and professional environments.



Multilingual content presents an additional challenge. Documents written in different languages or containing translated text may exhibit similarities that are difficult to detect using conventional text comparison methods. Developing techniques that can analyze semantic meaning across multiple languages remains an important area of research.

Another concern is balancing sensitivity and accuracy. A highly sensitive system may identify common phrases, references, or technical terminology as plagiarism, resulting in false positives. Conversely, a less sensitive system may fail to detect subtle forms of content duplication. Achieving an appropriate balance is essential for generating reliable plagiarism reports.

Ethical Considerations

The development and use of plagiarism detection systems involve several ethical responsibilities. One important consideration is the protection of user privacy. Documents submitted for analysis may contain confidential academic, research, or organizational information. Therefore, the system must ensure secure storage, controlled access, and proper handling of user data to prevent unauthorized disclosure.

Transparency is another critical ethical aspect. Users should be informed about how plagiarism scores are generated and how the system performs similarity analysis. Providing clear reports and understandable explanations helps users interpret results correctly and reduces the risk of misunderstanding or misuse.

The system must also ensure fairness in plagiarism assessment. Similarity scores should not be treated as definitive proof of misconduct. Certain similarities may occur due to commonly used terminology, properly cited references, or standardized expressions within a subject area. Human review and contextual evaluation remain important before making academic or professional decisions based on system-generated reports.

Intellectual property rights must also be respected. The system should be designed to detect unauthorized content duplication while ensuring that original authors receive proper recognition for their work. At the same time, uploaded documents should not be distributed, copied, or used beyond the intended purpose of plagiarism analysis without permission.

Finally, developers must ensure that the system is used responsibly and ethically. The primary objective of plagiarism detection technology is to promote originality, academic honesty, and ethical content creation rather than to punish users unfairly. By maintaining transparency, privacy, fairness, and accountability, the **Online Plagiarism Checker Using Data Mining Techniques** can serve as a trustworthy tool for supporting academic and professional integrity.

VII. RESULTS AND DISCUSSION

Results

The **Online Plagiarism Checker Using Data Mining Techniques** was successfully developed and tested using various types of digital documents. The system was able to extract textual content from uploaded files, process the

VI. CHALLENGES AND ETHICAL CONSIDERATIONS

Challenges

One of the major challenges in plagiarism detection is accurately identifying content that has been paraphrased or modified. While exact text matching can detect directly copied material, detecting similarities in reworded sentences requires more advanced analysis techniques. Users may change sentence structures, replace words with synonyms, or rearrange content while retaining the original meaning, making detection more difficult.

Another challenge is handling documents from different sources and formats. Research papers, assignments, scanned documents, and image-based files often contain varying layouts and structures. Extracting accurate text from these documents, particularly from low-quality scans, can affect the reliability of plagiarism analysis. Errors introduced during text extraction may lead to inaccurate similarity results.

The performance of plagiarism detection systems can also be affected by the increasing volume of digital content. As the database grows, the system must compare uploaded documents against a larger collection of existing files. Ensuring fast processing and maintaining high accuracy while handling large datasets requires efficient algorithms and optimized storage mechanisms.

extracted information, and perform similarity analysis using data mining-based plagiarism detection techniques. The implemented approach demonstrated the capability to identify duplicated content and generate detailed plagiarism reports within a reasonable processing time.

During testing, the system effectively detected similarities between documents containing directly copied text as well as documents with partially modified content. The combination of document fingerprinting and passage-level analysis enabled the identification of matching sections while providing an overall similarity percentage. The generated reports clearly highlighted duplicated passages, making it easier for users to understand the source and extent of detected similarities.

The integration of Optical Character Recognition (OCR) further improved the functionality of the system by enabling the analysis of scanned documents and image-based PDFs. This feature expanded the range of supported document formats and ensured that plagiarism detection was not limited to machine-readable text files. The system also maintained records of analyzed documents, allowing future comparisons against previously uploaded content.

The web-based interface provided a user-friendly environment for document submission, plagiarism checking, and report viewing. Users were able to upload documents, initiate analysis, and obtain results without requiring technical expertise. The overall implementation demonstrated the practicality of applying data mining techniques to automate plagiarism detection in academic and professional environments.

Discussion

The results indicate that the proposed system provides an effective solution for identifying textual similarities in digital documents. By combining multiple analysis techniques, the system achieves better detection performance than approaches that rely solely on direct text matching. Document fingerprinting reduces comparison time, while passage-level matching improves the precision of plagiarism identification.

One of the significant observations during testing was the importance of preprocessing in improving detection accuracy. Proper text extraction, normalization, and cleaning helped reduce inconsistencies and enabled more reliable similarity measurements. The quality of extracted text directly influenced the effectiveness of the plagiarism detection process.

The incorporation of OCR technology contributed positively to the system's performance by allowing analysis of scanned documents. However, the accuracy of OCR-based extraction depended on the quality and clarity of the uploaded files. Poorly scanned documents occasionally produced extraction errors that could affect similarity calculations. Despite this limitation, the system successfully processed most scanned documents with satisfactory accuracy.

The study also highlighted the benefits of using data mining techniques for handling large document collections. Instead of performing exhaustive comparisons between complete documents, the system utilized compact representations and efficient matching methods, resulting in improved scalability and faster response times.

Although the system performed well in detecting direct and near-duplicate content, identifying heavily paraphrased text remains a challenging task. Future enhancements involving semantic analysis, natural language processing, and machine learning techniques could further improve the system's ability to detect complex forms of plagiarism.

Overall, the obtained results demonstrate that the **Online Plagiarism Checker Using Data Mining Techniques** is capable of providing accurate, efficient, and scalable plagiarism detection. The findings validate the effectiveness of the proposed approach and support its application in educational institutions, research organizations, and content management systems where originality verification is essential.

VIII. CONCLUSION AND FUTURE SCOPE

Conclusion

The **Online Plagiarism Checker Using Data Mining Techniques** was developed to provide an efficient and reliable solution for detecting plagiarism in digital documents. The system combines document processing, text extraction, data mining techniques, and similarity analysis methods to identify duplicated content and evaluate document originality. By automating the plagiarism detection process, the proposed solution reduces the time and effort required for manual document comparison while improving the accuracy of content verification.

The implementation successfully integrates document fingerprinting, passage-level matching, and OCR-based text extraction to analyze both machine-readable and scanned documents. The system generates detailed plagiarism reports that include similarity percentages and highlighted matching sections, enabling users to understand the extent of content duplication. The web-based architecture further enhances accessibility and usability, making the system suitable for academic institutions, research organizations, and professional environments.

The results obtained during testing demonstrate that the proposed approach can effectively detect direct and near-duplicate textual content while maintaining efficient processing performance. The use of data mining techniques contributes to faster document comparison and improved scalability when handling large collections of documents. Overall, the project achieves its objective of providing an automated platform for plagiarism detection and promoting originality in digital content creation.

Future Scope

Although the developed system provides effective plagiarism detection capabilities, several enhancements can be incorporated in future versions to improve its performance and functionality.

One potential improvement is the integration of **Artificial Intelligence (AI)** and **Machine Learning (ML)** techniques to detect complex forms of plagiarism, including paraphrased and semantically similar content. Such techniques can analyze contextual meaning rather than relying solely on textual similarity, resulting in more accurate detection.

Another area for future development is **multilingual plagiarism detection**. Extending the system to support multiple languages would allow the comparison of documents written in different languages and help identify translated forms of plagiarism.

The system can also be enhanced through the incorporation of **Natural Language Processing (NLP)** methods. NLP-based approaches can improve text understanding, semantic analysis, and contextual similarity measurement, thereby increasing detection effectiveness.

Future versions may include integration with larger academic databases, digital libraries, and online repositories. Access to broader document collections would improve comparison coverage and increase the likelihood of identifying copied content from external sources.

Additional improvements can focus on real-time plagiarism monitoring for educational platforms, learning management systems, and research submission portals. Such integration would allow plagiarism checks to be performed automatically during document submission.

The development of advanced visualization and reporting tools represents another promising direction. Interactive dashboards, similarity graphs, and detailed analytics could provide users with deeper insights into plagiarism patterns and document originality.

Furthermore, cloud-based deployment and distributed processing techniques can be adopted to support large-scale plagiarism detection and improve system scalability for institutions handling substantial volumes of documents.

In conclusion, the future expansion of the **Online Plagiarism Checker Using Data Mining Techniques** through AI, NLP, multilingual support, larger datasets, and cloud technologies can significantly enhance its accuracy, scalability, and applicability, making it a more comprehensive solution for maintaining academic and professional integrity.

REFERENCES

- [1] S. Schleimer, D. S. Wilkerson, and A. Aiken, "Winnowing: Local Algorithms for Document Fingerprinting," in *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, San Diego, CA, USA, 2003, pp. 76–85.
- [2] M. S. Charikar, "Similarity Estimation Techniques from Rounding Algorithms," in *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, Montreal, Canada, 2002, pp. 380–388.
- [3] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge University Press, 2008.
- [4] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2012.
- [5] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington, MA, USA: Morgan Kaufmann, 2016.

- [6] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology Behind Search*, 2nd ed. Harlow, U.K.: Addison-Wesley, 2011.
- [7] A. Jain, M. N. Murty, and P. J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, Sept. 1999.
- [8] R. Smith, "An Overview of the Tesseract OCR Engine," in *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR)*, Curitiba, Brazil, 2007, pp. 629–633.
- [9] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, 2023.
- [10] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.
- [11] A. Hotho, A. Nürnberger, and G. Paaß, "A Brief Survey of Text Mining," *LDV Forum*, vol. 20, no. 1, pp. 19–62, 2005.
- [12] B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. Boston, MA, USA: Addison-Wesley, 2010.
- [13] Turnitin, "Turnitin Similarity and Originality Checking," Turnitin LLC. [Online]. Available: <https://www.turnitin.com>. [Accessed: Jun. 15, 2026].
- [14] Copyscape, "Copyscape Plagiarism Detection Service," Indigo Stream Technologies Ltd. [Online]. Available: <https://www.copyscape.com>. [Accessed: Jun. 15, 2026].
- [15] Grammarly Inc., "Grammarly Plagiarism Checker," Grammarly. [Online]. Available: <https://www.grammarly.com/plagiarism-checker>. [Accessed: Jun. 15, 2026].