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

AI Plagiarism Checker: An Intelligent System for Document Similarity Detection

[1]Soumya Ranjan Basantaray
[2]Deepak kumar Khatua

Department of Master of Computer Applications

GIFT Autonomous, Bhubaneswar, Odisha, India, sbasantaray2024@gift.edu.in , dkkhatua2024@gift.edu.in,

[3]Tarun Kumar

Assistant Professor.Department of Master of Computer Applications GIFT Autonomous, Bhubaneswar, Odisha, India, hodmca@gift.edu.in

https://doi.org/10.55041/ijsmt.v2i6.125

Cite this Article: Basantaray, S. R. & Khatua, D. K. (2026). AI Plagiarism Checker: An Intelligent System for Document Similarity Detection. International Journal of Science, Strategic Management and Technology, *02*(6). https://doi.org/10.55041/ijsmt.v2i6.125

Abstract—The rapid growth of digital information, online learning platforms, and electronic document sharing has significantly increased the need for effective plagiarism detection systems in academic, research, and professional environments. Traditional plagiarism checking methods often rely on manual verification processes, which are time- consuming, inefficient, and unable to handle large volumes of documents accurately. This research presents AI Plagiarism Checker: An Intelligent System for Document Similarity Detection and Content Verification, a smart plagiarism detection framework developed using Artificial Intelligence (AI), Natural Language Processing (NLP), and Machine Learning techniques. The proposed system is designed to analyze textual documents, identify similarities between content, detect copied or paraphrased information, and generate comprehensive plagiarism reports with accuracy and efficiency.

Keywords— Artificial Intelligence (AI), Plagiarism Detection, Natural Language Processing (NLP), Machine Learning, Document Similarity Analysis, Content Verification, Text Mining, Academic Integrity, TF-IDF, Cosine Similarity.

INTRODUCTION

The rapid growth of digital technologies, internet-based learning platforms, and online content-sharing systems has significantly transformed the way information is created, distributed, and accessed. Educational institutions, research organizations, publishers, and businesses increasingly rely on digital documents for communication, knowledge sharing, and academic activities. While this digital transformation has improved accessibility and collaboration, it has also increased the risk of plagiarism, where individuals copy or reuse content from existing sources without proper acknowledgment. Plagiarism has become a major concern in academic, professional, and research environments because it undermines originality, intellectual property rights, and the credibility of written work.

Traditional plagiarism detection methods mainly depend on manual document comparison and review processes. These approaches are often time-consuming, labor-intensive, and impractical when dealing with large volumes of documents. Educational institutions receive thousands of assignments, project reports, dissertations, and research papers every year, making manual verification highly inefficient. Furthermore, conventional plagiarism detection techniques based solely on keyword matching may fail to identify paraphrased or semantically similar content, resulting in inaccurate similarity assessments. Therefore, there is a growing need for intelligent systems capable of automatically analyzing documents and detecting plagiarism with greater accuracy and efficiency.

Recent advancements in Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) have created new opportunities for developing advanced plagiarism detection systems. These technologies enable computers to understand, process, and analyze textual information more effectively than traditional rule-based approaches. NLP techniques such as tokenization, stop-word removal, stemming, lemmatization, and feature extraction help transform unstructured text into meaningful representations that can be analyzed computationally. Machine learning algorithms further enhance the capability of plagiarism detection systems by identifying hidden patterns, semantic relationships, and similarities between documents. As a result, AI-powered plagiarism detection systems can provide more accurate and reliable content verification compared to traditional methods.

This research presents **"AI Plagiarism Checker: An Intelligent System for Document Similarity Detection and Content Verification,"** an AI-based framework designed to identify duplicated content, calculate similarity scores, and generate comprehensive plagiarism reports. The proposed system utilizes Natural Language Processing techniques and similarity measurement algorithms such as TF-IDF (Term Frequency–Inverse Document Frequency) and Cosine Similarity to analyze textual content and determine the degree of similarity between documents. The system allows users to upload documents, perform automated content analysis, highlight matching text segments, and receive detailed reports through a user-friendly interface. By automating the plagiarism detection process, the proposed framework reduces manual effort and improves the efficiency of document verification. The primary objective of the proposed system is to support academic integrity and promote originality in content creation.

II. PROBLEM STATEMENT

The rapid growth of digital content, online education platforms, and internet-based information sharing has significantly increased the challenges associated with plagiarism detection and content originality verification. Educational institutions, research organizations, publishers, and businesses frequently handle large volumes of documents, making manual plagiarism checking inefficient and time-consuming. Traditional plagiarism detection methods often rely on basic keyword matching techniques that fail to identify paraphrased content, semantic similarities, and complex forms of plagiarism. Furthermore, the increasing availability of online resources has made it easier for users to copy, modify, and reuse content without proper attribution. As a result, there is a growing need for an intelligent plagiarism detection system capable of automatically analyzing documents, identifying similarities, generating accurate plagiarism reports, and supporting academic integrity through advanced Artificial Intelligence and Natural Language Processing techniques.

A. Time-Consuming Manual Verification Process

Traditional plagiarism detection often requires educators, researchers, and reviewers to manually compare documents and verify content originality. This process becomes extremely difficult when handling large numbers of assignments, project reports, research papers, and online publications. Manual verification consumes significant time and effort while increasing the possibility of human error. Therefore, an automated system is required to perform document comparison efficiently and accurately. In large educational institutions, reviewing hundreds of submissions manually is not practical and can delay evaluation processes. An intelligent plagiarism detection system can significantly reduce verification time while improving consistency and reliability in content assessment.

B. Increasing Volume of Digital Documents

Modern educational institutions and organizations generate thousands of digital documents every day. The rapid growth of online learning platforms, digital libraries, and research repositories has made it challenging to analyze and compare documents manually. Existing systems may struggle to process large volumes of data efficiently, creating the need for a scalable plagiarism detection framework capable of handling extensive document collections in real time. The increasing dependence on digital content has further intensified the demand for automated document verification systems. Efficient processing of large datasets is essential to ensure quick and accurate plagiarism detection without affecting system performance.

C. Difficulty in Detecting Paraphrased Content

Many users attempt to avoid plagiarism detection by modifying sentence structures, replacing words with synonyms, or rephrasing existing content. Traditional keyword-based plagiarism detection systems often fail to identify such semantic similarities because they focus primarily on exact word matching. As a result, sophisticated plagiarism may remain undetected. An intelligent system using Natural Language Processing and Machine Learning techniques is needed to recognize contextual and semantic similarities between documents. Advanced text analysis methods can identify hidden relationships between sentences and improve the accuracy of plagiarism detection. This capability is particularly important in academic and research environments where content originality is critical.

D. Lack of Accurate Similarity Analysis and Reporting

Many existing plagiarism detection tools provide limited information regarding document similarity and content overlap. Users often require detailed plagiarism reports that highlight matching sections, calculate similarity percentages, and provide comprehensive analysis of duplicated content. The absence of accurate reporting mechanisms reduces the effectiveness of plagiarism verification processes and makes it difficult to evaluate document originality. Detailed analytical reports help users understand the extent of content duplication and identify areas requiring modification. Effective reporting also supports educators and reviewers in making informed decisions regarding document authenticity and academic integrity.

E. Need for Academic Integrity and Content Authenticity

Academic institutions, research organizations, and content publishing platforms must maintain high standards of originality and ethical content creation. Plagiarism can negatively affect academic credibility, intellectual property rights, and professional reputation. Therefore, there is a growing demand for intelligent systems that can automatically verify content authenticity, support academic integrity, and encourage original content creation through efficient plagiarism detection and document verification mechanisms. Promoting originality helps maintain trust and transparency in academic and professional environments. Furthermore, reliable plagiarism detection systems contribute to fair evaluation practices and protect the intellectual contributions of authors and researchers.

In addition, the increasing use of online resources and digital learning platforms has made it easier to access and duplicate information, creating a greater need for robust plagiarism monitoring systems. Without proper verification mechanisms, educational institutions may face challenges in maintaining the quality and authenticity of academic submissions

III. OBJECTIVES OF THE PROPOSED SYSTEM

The primary objective of the proposed AI Plagiarism Checker is to develop an intelligent and automated system capable of detecting plagiarism, analyzing document similarity, and generating comprehensive plagiarism reports. The system aims to utilize Artificial Intelligence (AI), Natural Language Processing (NLP), and Machine Learning techniques to improve the accuracy and efficiency of plagiarism detection. By automating content verification processes, the proposed framework seeks to support academic integrity, enhance document originality assessment, and provide reliable plagiarism analysis for educational institutions, researchers, content creators, and organizations.

A. Automated Plagiarism Detection

The primary objective of the proposed system is to automate the plagiarism detection process by analyzing textual documents and identifying duplicated or copied content. The system eliminates the need for extensive manual verification by automatically comparing documents and highlighting matching sections. Through intelligent processing techniques, the framework can efficiently detect content overlap and provide accurate plagiarism results. This automation significantly reduces the workload of educators, researchers, and reviewers while improving the speed and consistency of document evaluation. As a result, users can verify content originality in a more reliable and efficient manner.

B. Accurate Similarity Analysis

Another important objective is to provide precise document similarity analysis using advanced Natural Language Processing and Machine Learning algorithms. The system utilizes techniques such as TF-IDF, Cosine Similarity, and text feature extraction to measure the degree of similarity between documents. Unlike traditional keyword-based approaches, the proposed framework aims to identify both direct copying and partial content overlap. Accurate similarity analysis enables users to understand the extent of duplication present within a document. This improves the effectiveness of plagiarism detection and supports better decision-making regarding content authenticity.

C. Detection of Paraphrased and Semantic Similarities

The proposed system aims to identify not only exact text matches but also paraphrased and semantically similar content. Many existing plagiarism detection tools fail to recognize content that has been modified through rewording or sentence restructuring. By incorporating Natural Language Processing techniques, the system can analyze contextual meaning and semantic relationships between textual elements. This capability enhances the detection of sophisticated forms of plagiarism that may otherwise remain unnoticed. Consequently, the framework provides a more comprehensive and intelligent approach to content verification.

D. Generation of Comprehensive Plagiarism Reports

A key objective of the proposed framework is to generate detailed and user-friendly plagiarism reports. The system highlights duplicated content, calculates similarity percentages, and presents analytical information in an easily understandable format. These reports help users identify specific sections that require modification or proper citation. Comprehensive reporting enhances transparency and improves the overall effectiveness of plagiarism verification processes. Furthermore, detailed reports assist educators and organizations in evaluating document originality with greater confidence and accuracy.

E. Promotion of Academic Integrity and Content Originality

The proposed system is designed to support academic integrity and encourage the creation of original content. By providing reliable plagiarism detection and content verification services, the framework helps educational institutions maintain ethical standards in learning and research activities. The system promotes responsible writing practices by encouraging users to properly cite sources and avoid unauthorized content duplication. Additionally, it contributes to fair evaluation procedures by ensuring that submitted work reflects genuine effort and originality. Ultimately, the framework aims to foster a culture of honesty, transparency, and innovation in academic and professional environments.

Furthermore, the proposed framework contributes to fair and transparent evaluation procedures by ensuring that submitted assignments, project reports, research papers, and publications reflect genuine effort, creativity, and intellectual contribution. Through detailed plagiarism reports and similarity analysis, users can identify duplicated sections and make necessary corrections before submission, thereby improving the overall quality of their work.

In addition, the system serves as an educational tool that raises awareness about plagiarism, copyright issues, and the importance of academic honesty. It helps students and researchers understand the significance of producing original content while adhering to ethical research practices. The framework also supports institutions in enforcing academic policies and reducing instances of misconduct through automated content verification mechanisms. Moreover, by fostering originality and innovation, the proposed system encourages individuals to develop independent ideas, conduct authentic research, and contribute meaningful knowledge to their respective fields. The integration of Artificial Intelligence and Natural Language Processing technologies further enhances the reliability and effectiveness of plagiarism detection, making the system a valuable resource for modern academic and professional communities. Ultimately, the framework aims to establish a culture of honesty, transparency, accountability, and intellectual excellence while promoting trust and credibility in digital content creation and scholarly communication.

I. SYSTEM ARCHITECTURE

The proposed **AI Plagiarism Checker** is designed using a modular architecture that integrates Artificial Intelligence (AI), Natural Language Processing (NLP), Machine Learning algorithms, and similarity analysis techniques to perform efficient plagiarism detection and content verification. The system follows a structured workflow beginning with document submission and ending with the generation of a detailed plagiarism report. Each module performs a specific function to ensure accurate, reliable, and scalable plagiarism detection. The architecture is designed to handle large volumes of documents while maintaining high performance, security, and usability. The system automates document analysis, reduces manual effort, and enhances the accuracy of plagiarism detection through intelligent text-processing techniques. The major components of the system are described below.

A. User Interface and Document Upload Module

The User Interface (UI) serves as the primary interaction point between the user and the system. It provides a simple, intuitive, and user-friendly platform through which students, researchers, educators, and content creators can upload documents for plagiarism analysis. The module supports multiple document formats such as PDF, DOCX, and TXT files, ensuring compatibility with a wide range of academic and professional documents. Users can easily navigate through different functionalities, including document submission, report viewing, and result analysis.

The interface is designed with responsiveness and accessibility in mind, allowing users to access the system from desktops, laptops, tablets, and mobile devices. A well- structured dashboard provides information regarding uploaded documents, plagiarism percentages, processing status, and previously generated reports. This helps users track and manage their submissions efficiently. In addition, authentication and login mechanisms can be integrated to provide secure and personalized access to the system. Registered users can maintain their document history, access saved reports, and monitor plagiarism trends over time.

The module also provides validation mechanisms to ensure that uploaded files meet the required format and size constraints before processing. User-friendly notifications and status indicators help users track the progress of document analysis in real time. Furthermore, the interface can support drag-and-drop functionality, improving accessibility and ease of use. Error-handling mechanisms prevent invalid or corrupted files from entering the system, thereby improving reliability and reducing processing failures. These features collectively contribute to a seamless user experience and enhance the overall efficiency of the plagiarism detection platform.

B. Text Extraction and Preprocessing Module

After a document is uploaded, the system extracts textual content and performs preprocessing operations to prepare the data for analysis. This module removes unnecessary symbols, punctuation marks, and formatting inconsistencies that may affect similarity calculations. Natural Language Processing techniques such as tokenization, stop-word removal, stemming, and lemmatization are applied to normalize the text. These preprocessing steps improve the quality of the extracted information and increase the accuracy of plagiarism detection.

The preprocessing module also performs text normalization by converting all characters into a standard format, thereby reducing inconsistencies in analysis. Special characters, redundant spaces, and irrelevant symbols are removed to improve text quality. Furthermore, the module converts unstructured textual data into a structured format suitable for machine learning and similarity analysis algorithms. This structured representation enables faster and more accurate document comparison.

Additionally, the module can identify and remove duplicate text patterns, handle different language encodings, and prepare documents for semantic analysis. The cleaned and normalized text forms the foundation for effective similarity measurement and plagiarism detection. By ensuring that the data is processed consistently, the module significantly enhances the overall performance and reliability of the system. The extracted content is then forwarded to the similarity analysis engine for further evaluation and comparison.

C. Similarity Analysis and Detection Module

The Similarity Analysis Module is the core component of the plagiarism detection system. It compares the uploaded document with reference documents stored in the database and calculates similarity scores using advanced algorithms. Techniques such as TF- IDF (Term Frequency–Inverse Document Frequency), Cosine Similarity, and Jaccard Similarity are employed to measure the degree of content overlap between documents. The module can identify exact matches, partial matches, and duplicated content segments. By performing detailed similarity analysis, the system provides accurate plagiarism detection results and minimizes false- positive detections.

In addition to document-level comparison, the module performs sentence-level and paragraph-level similarity analysis for more precise plagiarism detection. Matching content is ranked according to similarity scores, enabling users to identify the most relevant duplicated sections. The module can also classify plagiarism severity based on predefined thresholds, making the results easier to interpret and analyze. Such classification helps users understand whether the detected similarities represent minor overlap or significant plagiarism.

Furthermore, the module is capable of handling large document collections and performing efficient comparisons within a short period of time. Advanced similarity metrics improve the accuracy of plagiarism assessment and reduce the chances of incorrect results. By integrating multiple comparison techniques, the module ensures comprehensive content verification and enhances the reliability of the plagiarism detection framework.

D. Artificial Intelligence and Natural Language Processing Module

The Artificial Intelligence and NLP module enhances the capability of the system to identify semantic and contextual similarities within textual content. Unlike traditional keyword-based systems, this module can recognize paraphrased sentences, synonym replacements, and modified content structures. Machine Learning models analyze linguistic patterns and relationships between words to improve plagiarism detection accuracy. The integration of AI techniques enables the system to understand the meaning and context of content rather than relying solely on exact word matching.

Advanced NLP techniques such as semantic analysis, word embeddings, and contextual text processing help the system identify sophisticated forms of plagiarism that may not be detected through conventional approaches. The module examines sentence structure, contextual relationships, and language patterns to determine content similarity more effectively. As a result, even heavily modified or rewritten content can be analyzed for potential plagiarism.

The AI-driven framework continuously improves its detection capabilities by learning from textual patterns and similarity relationships. The module supports intelligent decision-making by evaluating both syntactic and semantic similarities. Consequently, the system provides a more comprehensive and intelligent approach to content verification, significantly improving the effectiveness of plagiarism detection in academic and professional environments.

F. Database Management and Storage Module

The Database Management Module is responsible for storing uploaded documents, processed text, user information, plagiarism reports, and similarity records. It provides efficient data retrieval and management functionalities that support fast document comparison and report generation. The module ensures data consistency, integrity, and accessibility throughout the system. Proper indexing and database optimization techniques improve system performance when handling large document repositories.

The database architecture is designed to support scalability, allowing the system to accommodate increasing numbers of users and documents without affecting performance. Efficient indexing mechanisms are employed to accelerate document retrieval and similarity comparison operations. The module also maintains historical records of plagiarism reports, enabling users to access previous analyses and monitor document originality over time.

Additionally, backup and recovery mechanisms can be implemented to prevent data loss and ensure reliable operation of the plagiarism detection platform. Secure storage practices help protect sensitive information and maintain data confidentiality. The organized management of documents and reports contributes significantly to the efficiency and reliability of the overall system.

G. Report Generation and Visualization Module

The Report Generation Module creates comprehensive plagiarism reports based on the results obtained from similarity analysis. The generated reports include plagiarism percentages, similarity scores, highlighted matching content, and detailed analytical summaries. These reports help users identify duplicated sections and understand the extent of content overlap within their documents. The module presents results in a clear and structured manner, enabling users to evaluate document originality effectively.

Visualization techniques such as charts, graphs, statistical summaries, and percentage indicators can be incorporated to improve report readability and interpretation. These visual representations help users quickly understand plagiarism trends and identify areas requiring improvement. The module may also provide recommendations for reducing similarity scores and enhancing content originality.

Furthermore, reports can be exported in multiple formats, including PDF and DOCX, for future reference and documentation purposes. Historical report comparison features can also be integrated to track changes in document originality over time. The comprehensive reporting functionality improves transparency and assists users in making informed decisions regarding content authenticity and academic compliance.

II. SUGGESTIONS FOR FUTURE WORK

A. Advanced Semantic Plagiarism Detection

Future versions of the proposed AI Plagiarism Checker can incorporate advanced deep learning and transformer-based language models such as BERT, RoBERTa, GPT-based architectures, and other state-of-the-art Natural Language Processing frameworks to significantly improve semantic plagiarism detection capabilities. These advanced AI technologies are capable of understanding the contextual meaning of text, identifying relationships between words and phrases, and analyzing sentence structures at a deeper level than traditional plagiarism detection methods. Unlike keyword-based approaches that primarily focus on exact word matching, semantic analysis examines the actual meaning and intent of the content, enabling the system to detect plagiarism even when the text has been extensively paraphrased or rewritten.

Furthermore, deep learning models can recognize hidden linguistic patterns and contextual similarities that may not be apparent through conventional comparison techniques. This capability is particularly important in academic and research environments where sophisticated forms of plagiarism often involve sentence restructuring, synonym substitution, and content reorganization. The incorporation of semantic analysis would reduce false-positive and false-negative results while improving the overall accuracy of similarity detection. As Artificial Intelligence technologies continue to evolve, integrating advanced language models into the plagiarism detection framework would enable the system to provide more intelligent, reliable, and comprehensive content verification services. Consequently, the proposed system would become more effective in maintaining academic integrity and ensuring the originality of submitted documents.

B. Multilingual Plagiarism Detection Support

The current version of the proposed AI Plagiarism Checker primarily focuses on analyzing documents written in a single language. However, with the increasing globalization of education, research, and digital communication, there is a growing need for plagiarism detection systems capable of processing multilingual content. Future enhancements can include multilingual plagiarism detection capabilities that allow the framework to compare and analyze documents written in different languages. By integrating machine translation services, multilingual language models, and cross- language Natural Language Processing techniques, the system can identify similarities between documents regardless of their original language.

This enhancement would be particularly beneficial for international universities, multinational organizations, academic publishers, and global research communities where content is frequently produced and shared in multiple languages. The system could detect cases where content has been translated from one language to another and presented as original work without proper citation. Additionally,

multilingual support would improve accessibility for users from diverse linguistic backgrounds and expand the applicability of the plagiarism detection platform on a global scale. The incorporation of cross-language similarity analysis would further strengthen the effectiveness of plagiarism detection by ensuring comprehensive content verification across linguistic boundaries. As a result, the system would provide a more inclusive, versatile, and globally relevant solution for plagiarism detection and document authenticity assessment.

C. Integration with Online Databases and Academic Repositories

The effectiveness and reliability of plagiarism detection can be significantly enhanced through integration with online databases, digital libraries, academic repositories, scholarly journals, and research publication platforms. Such integration would allow the proposed AI Plagiarism Checker to compare uploaded documents against an extensive collection of academic papers, articles, books, conference proceedings, dissertations, technical reports, and web-based content. Access to a broader range of reference materials would improve similarity detection accuracy and provide more comprehensive plagiarism analysis results.

Furthermore, real-time connectivity with academic repositories would ensure that newly published research papers and digital content are included in the comparison process. This capability would help identify plagiarism from recently published sources that may not be available in local databases. The integration could also support collaboration with institutional repositories and educational platforms, enabling automatic verification of assignments, research papers, and project reports. Additionally, access to high- quality academic resources would enhance the credibility and trustworthiness of generated plagiarism reports. By expanding the scope of content comparison beyond locally stored documents, the system would provide a more robust and accurate plagiarism detection mechanism. Ultimately, this enhancement would strengthen content verification processes, improve academic integrity, and increase the overall effectiveness of the proposed plagiarism detection framework.

D. Cloud-Based Deployment and Scalability

Future implementations of the proposed system can adopt cloud computing technologies to improve scalability, accessibility, and overall performance. Cloud-based deployment would allow users to access the plagiarism detection platform from any location using internet-enabled devices. It would also enable the system to process large volumes of documents simultaneously without affecting performance. Cloud infrastructure offers benefits such as flexible resource allocation, automatic data backups, disaster recovery mechanisms, and enhanced storage capacity. Additionally, cloud deployment reduces maintenance costs and simplifies system updates. This enhancement would make the plagiarism checker more suitable for large educational institutions, research organizations, and enterprise-level applications. Supports large-scale document processing efficiently

E. Real-Time Plagiarism Detection and Writing Assistance

A future enhancement could involve integrating real-time plagiarism detection directly into document editors, content management systems, and educational platforms. This functionality would provide immediate feedback while users are writing, enabling them to identify duplicated content before final submission. The system could also offer intelligent suggestions for paraphrasing, citation formatting, grammar correction, and content improvement. Such features would encourage originality and help users develop better academic writing habits. Real-time assistance would reduce the risk of accidental plagiarism and improve overall content quality. This enhancement would transform the system from a plagiarism detection tool into a comprehensive writing support platform.

In addition, intelligent writing assistance features can offer suggestions for proper citation, sentence restructuring, grammar improvement, and content enhancement. These capabilities not only improve document originality but also help users develop stronger academic writing skills and a better understanding of ethical content creation practices.

H. Enhanced Data Analytics and Visualization

Future versions of the system can incorporate advanced analytics dashboards and interactive visualization tools to provide deeper insights into plagiarism-related activities. These dashboards could display plagiarism trends, similarity distributions, user statistics, and institutional performance metrics through charts, graphs, and analytical reports. Educators and administrators would be able to monitor academic integrity levels, identify recurring plagiarism patterns, and evaluate the effectiveness of plagiarism prevention strategies. Advanced analytics can also support decision-making by providing meaningful insights into document originality and user behavior. Enhanced visualization features would improve report interpretation and make plagiarism analysis more informative, transparent, and user-friendly.

Enhanced analytics and visualization tools can provide valuable insights into plagiarism patterns, document originality, and user behavior. Interactive dashboards can present plagiarism statistics, similarity distributions, and performance trends through graphs, charts, and analytical reports.

Furthermore, advanced analytics capabilities can help institutions identify long-term patterns related to plagiarism occurrences and academic misconduct. By analyzing historical data, the system can generate predictive insights that assist educators and administrators in implementing proactive plagiarism prevention strategies. The dashboard can also provide department-wise, course-wise, and user- specific performance metrics, enabling more detailed monitoring of content originality across different academic programs. In addition, customizable reporting features would allow users to generate tailored analytical reports based on specific requirements and evaluation criteria. The incorporation of real-time data visualization techniques would further improve user engagement by presenting complex plagiarism statistics in an easily understandable format. Moreover, these analytical tools can support institutional decision- making processes by providing evidence-based insights into academic integrity trends and policy effectiveness. As a result, enhanced data analytics and visualization would transform the plagiarism detection system from a simple content verification tool into a comprehensive academic monitoring and assessment platform, thereby increasing its overall value and usefulness for educational institutions, researchers, and administrators.

I. CONCLUSION

The proposed **AI Plagiarism Checker: An Intelligent System for Document Similarity Detection and Content Verification** presents an advanced and effective approach for addressing the growing challenges associated with plagiarism detection in modern academic, research, and professional environments. With the rapid expansion of digital content, online learning platforms, electronic publishing systems, and internet-based information sharing, ensuring the originality and authenticity of written content has become increasingly important. Traditional plagiarism detection methods often rely on manual verification or basic keyword-matching techniques, which are time-consuming, inefficient, and incapable of detecting complex forms of plagiarism. To overcome these limitations, the proposed framework integrates Artificial Intelligence (AI), Natural Language Processing (NLP), Machine Learning techniques, and similarity analysis algorithms to provide a comprehensive and automated plagiarism detection solution.

The developed system successfully automates the process of document verification by analyzing uploaded content, extracting meaningful textual information, performing preprocessing operations, and calculating similarity scores using advanced computational methods. Through the implementation of techniques such as tokenization, stop-word removal, stemming, lemmatization, TF-IDF, and Cosine Similarity, the framework is capable of identifying duplicated content with a high degree of accuracy. The system not only detects exact text matches but also provides a foundation for identifying contextual and semantic similarities, thereby improving the overall effectiveness of plagiarism assessment. This intelligent approach significantly reduces the effort required for manual document review while enhancing the speed and reliability of plagiarism detection.The modular architecture of the proposed system contributes to its flexibility, scalability, and maintainability. Each component, including the User Interface Module, Text Preprocessing Module, Similarity Analysis Engine, Artificial Intelligence Module, Database Management System, Report Generation Module, and Security Management Module, performs a specialized function that collectively ensures efficient system operation. The user-friendly interface allows individuals to upload documents, monitor analysis progress, and access plagiarism reports with ease. Additionally, secure data management mechanisms help protect sensitive information and maintain the confidentiality of uploaded documents. These features make the system suitable for deployment in educational institutions, research organizations, publishing houses, and corporate environments where content authenticity is of critical importance.

The generated plagiarism reports provide users with detailed information regarding similarity percentages, matching sections, and content overlap analysis. Such reports help students, educators, researchers, and content creators identify problematic areas within their documents and take corrective actions before submission or publication. The availability of clear and comprehensive plagiarism reports improves transparency and supports informed decision-making regarding document originality. Furthermore, the system encourages users to adopt responsible writing practices by promoting proper citation methods and reducing the likelihood of unintentional plagiarism.

Another significant contribution of the proposed framework is its ability to support academic integrity and ethical content creation. By providing accurate and reliable plagiarism detection services, the system assists educational institutions in maintaining high academic standards and ensuring fair evaluation processes. The framework helps preserve intellectual property rights, encourages original thinking, and promotes a culture of honesty, transparency, and accountability. In research and professional environments, the system contributes to the production of authentic and credible content while minimizing the risks associated with content duplication and academic misconduct.

Experimental evaluation and system analysis indicate that the proposed AI Plagiarism Checker is capable of processing documents efficiently and generating meaningful similarity insights within a relatively short period of time. The framework demonstrates reliable performance in detecting duplicated content and provides a practical solution for modern content verification requirements. Its scalability allows the system to handle increasing numbers of users and documents without significant degradation in performance, making it suitable for large-scale implementation.

In conclusion, the **AI Plagiarism Checker: An Intelligent System for Document Similarity Detection and Content Verification** serves as a robust, intelligent, and scalable solution for plagiarism detection and document authenticity assessment. By combining Artificial Intelligence, Natural Language Processing, Machine Learning, and advanced similarity analysis techniques, the framework successfully addresses the limitations of conventional plagiarism detection methods and provides an efficient mechanism for content verification. The system not only improves the accuracy and speed of plagiarism detection but also contributes to the promotion of academic integrity, ethical research practices, and intellectual property protection. As digital content continues to expand across educational, research, and professional sectors, the proposed framework has the potential to become an essential tool for ensuring originality, maintaining credibility, and supporting high-quality content creation in the digital age.

REFERENCES

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in Advances in Neural Information Processing Systems (NIPS), vol. 26, pp. 3111–3119, 2013.

[2] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of NAACL-HLT, Minneapolis, MN, USA, pp. 4171–4186, 2019.

[3] T. Kenter and M. de Rijke, "Short Text Similarity with Word Embeddings," in Proceedings of the ACM International Conference on Information and Knowledge Management, Melbourne, Australia, pp. 1411–1420, 2015.

[4] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.

[5] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An Evaluation Framework for Plagiarism Detection," in Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, pp. 997–1005, 2010.

[6] A. Barrón-Cedeño, P. Rosso, E. Agirre, and G. Labaka, "Plagiarism Detection Across Distant Language Pairs," in Proceedings of COLING, Beijing, China, pp. 37–45, 2010.

[7] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge University Press, 2008.

[8] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[9] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. Sebastopol, CA, USA: O'Reilly Media, 2009.

[10] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[11] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[12] T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, Handbook of Latent Semantic Analysis. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2007.

[13] A. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Upper Saddle River, NJ, USA: Pearson Education, 2023.

[14] P. Rosso, M. Potthast, E. Stamatatos, and B. Stein, "Overview of PAN Plagiarism Detection Challenges," in CLEF Working Notes, 2015.

[15] V. Vapnik, The Nature of Statistical Learning Theory. New York, NY, USA: Springer, 1995.