



Integrated Plagiarism Detection System For Text And Image Based Content In Documents

¹ Dr Palvadi Srinivas Kumar, ² K.Mohana Preethi, ² V.Praneeth, ² J.Sharon Rose, ² A.Gayathri

¹ Associate Professor, Department of Computer Science and Engineering, SRK Institute of Technology,
Vijayawada, Andhra Pradesh, INDIA

&

Post Doctoral Research Fellow, Institute of Computer and Information Sciences, Srinivas University,
Mangalore, Karnataka, INDIA

² Students, Department of Computer Science and Engineering, SRK Institute of Technology, Vijayawada,
Andhra Pradesh, INDIA

Abstract

The integrated plagiarism detection system for text and image-based content in documents is designed to enhance the accuracy and efficiency of identifying duplicate or copied content across various formats. With the increasing accessibility of digital information, the risk of plagiarism has grown, making it essential to develop robust mechanisms that can analyse both textual and visual data. Traditional plagiarism detection tools focus primarily on textual similarities, leaving a gap in detecting image-based content that may also be duplicated or manipulated. This system addresses this limitation by incorporating advanced text-matching algorithms alongside image recognition techniques to comprehensively analyse documents. The proposed system utilizes natural language processing and deep learning methods to detect similarities in textual content, while computer vision techniques help identify duplicated or modified images. By integrating these approaches, the system ensures a more comprehensive evaluation of plagiarism in academic, research, and professional documents. The implementation of such a system is valuable for universities, publishers, and businesses aiming to maintain content originality and intellectual integrity. The integration of multi-modal detection methods improves the overall reliability of the system, reducing false positives and enhancing the accuracy of plagiarism reports. This approach provides an effective solution for modern plagiarism detection challenges, ensuring fair and ethical use of digital content.

Keywords:

Plagiarism detection, Image plagiarism, Online data sharing, Document sharing, Plagiarism checking software, AI-powered tools, Digital collaboration, Academic integrity, Content attribution, Comparative analysis.

I. Introduction

Plagiarism has become a widespread issue in the digital era due to the easy availability of information and the rapid growth of online resources. The integrity of academic, research, and professional content is constantly at risk due to the unauthorized duplication of text and images. Traditional plagiarism detection methods primarily focus on textual content, relying on keyword-based or semantic analysis techniques to identify similarities between documents. However, with the increasing use of multimedia in documents, there is a growing need for an integrated approach that can detect not only text-based plagiarism but also image-based duplication. This project aims to develop an integrated plagiarism detection system that can efficiently analyze both textual and visual content within documents, ensuring a more comprehensive and accurate assessment of originality.

Importance of Plagiarism Detection

Plagiarism undermines creativity, intellectual honesty, and academic credibility. In academic settings, plagiarism can lead to severe consequences such as rejection of research papers, loss of credibility, and even legal repercussions. In professional environments, copied content can affect brand reputation, lead to copyright violations, and result in financial losses. Detecting plagiarism accurately is crucial for maintaining the originality of content and protecting intellectual property rights. While text-based plagiarism detection has been well-studied, image-based plagiarism remains an area that requires more attention due to the complex nature of visual similarity assessment.

Challenges in Plagiarism Detection

Detecting plagiarism in documents involves various challenges, especially when dealing with multimedia content. Some of the key challenges include:

1. **Textual Similarity Detection:** Traditional text-matching techniques often struggle to identify paraphrased content, synonym substitutions, and slight modifications made to plagiarized text. Advanced approaches such as natural language processing and deep learning are required to improve the accuracy of similarity detection.
2. **Image-Based Plagiarism:** Unlike text, images can be altered in multiple ways, including cropping, filtering, resizing, and modifying specific elements. Identifying plagiarized images requires advanced computer vision techniques that can analyze visual features and detect similarities even when modifications have been applied.
3. **Multi-Modal Integration:** An integrated plagiarism detection system must efficiently combine both text and image analysis techniques to provide a unified result. This requires robust algorithms capable of processing and correlating different types of content.
4. **Scalability and Performance:** The system must be able to handle large volumes of documents efficiently. With increasing data, it is essential to design an optimized architecture that can process content quickly without compromising accuracy.

False Positives and False Negatives:

Ensuring a high level of accuracy in plagiarism detection is crucial. Many systems struggle with generating false positives, where non-plagiarized content is incorrectly flagged, or false negatives, where actual plagiarism goes undetected. The proposed system must minimize such errors through improved algorithms.

II. LITERATURE WORK

Various methodologies for text detection in images and videos have been developed over time. These methods generally fall into two primary categories: connected component-based techniques and texture-based approaches. Optical Character Recognition (OCR) involves converting scanned images of printed, handwritten, or typewritten text into machine-readable text. This technology enables automatic recognition of characters through optical scanning processes. The evolution of text detection and recognition has transitioned from document analysis to more complex scenarios involving camera-captured images. Early research focused on basic preprocessing, text detection, and OCR technology. [1] These advancements have significantly improved the extraction and interpretation of text from images, which is essential for tasks such as assignment plagiarism checking.

Machine learning algorithms are employed to analyze and compare student submissions against a broad dataset. By training on labeled examples, these systems can identify patterns and anomalies, providing a highly accurate and adaptable solution for detecting plagiarism. This technology supports academic integrity by offering educators a powerful tool to ensure that student work is original. However, challenges include occasional false negatives [2] where subtle plagiarism might go undetected and potential biases arising from the quality of training data. Privacy concerns also need to be addressed, as analyzing student submissions requires careful management of sensitive information.

Integrating data mining and Natural Language Processing (NLP) enhances academic integrity by evaluating the similarity between student submissions and reference corpora. This approach uses advanced algorithms to detect potential plagiarism and provides detailed reports for instructors through a user-friendly interface. The system's accuracy in identifying plagiarism benefits from sophisticated similarity analysis and NLP, [3] which helps in understanding linguistic nuances and detecting subtle academic dishonesty. However, privacy issues must be managed carefully, as the technology involves analyzing student work.

Text extraction from images involves several steps: text detection to identify text-containing areas, text localization to pinpoint exact text positions, text segmentation to separate text from its background, and binarization to convert color images to binary format. OCR systems face challenges such as recognizing handwritten text or complex fonts, which can limit their effectiveness in certain contexts.

Advanced algorithms are used to analyze and compare visual content to ensure the originality of online images. Image recognition and similarity metrics help identify instances of image plagiarism, providing a robust solution for maintaining integrity in visual content [4] and assignments. This technology is crucial for educators, content creators, and image-based platforms in preserving originality and intellectual property rights. Despite its effectiveness, challenges include accurately discerning context and intent, which can lead to potential false positives when images are similar but used legitimately. The system's performance may also be affected by the quality and diversity of the reference image database, influencing its ability to detect nuanced instances of plagiarism.

According to author [5] discuss the prevalence of plagiarism in free text, attributing it to the accessibility of vast information resources. Automated plagiarism detection systems are developed to identify plagiarized content in large repositories. However, the emergence of sophisticated plagiarism techniques, such as paraphrasing and summarizing, poses challenges as they effectively conceal instances of plagiarism.

According to author [6] point out that text-focused plagiarism detection tools often overlook visual elements, such as images and flowcharts, in scholarly works. These visual components are crucial for conveying complex data within research papers. The diverse range of images and their prevalence in computer-generated texts introduce the potential for plagiarism within visual content. Additionally, flowcharts, commonly used for information presentation, are also susceptible to plagiarism.

According to author [7] highlights the increasing concern of plagiarism in digital media due to the ease of copying. Detection methods focus on two forms: 'copy' and 'regeneration'. 'Copy' involves direct

replication, while 'regeneration' reproduces copied content with some creativity. In digital image works, 'copy' plagiarism is prevalent, emphasizing the importance of copy detection. However, current systems primarily address 'copy' plagiarism, leaving 'regeneration' relatively unaddressed.

According to author [8] define plagiarism as the act of presenting another author's work, thoughts, or ideas as one's own without proper citations. While technology advancements have facilitated easier access to knowledge, plagiarism remains a challenge. Despite the evolution of image-based plagiarism detection concepts, existing systems still exhibit shortcomings.

According to author [9] introduce the Flowchart Plagiarism Detection approach, employing Canny edge detection to identify plagiarism. The system processes flowchart images, detecting edges using Canny edge detection. Centroids and borders of shapes are identified, and their distances are calculated for comparison with an original image graph. This method is particularly effective for detecting plagiarism in flowchart images.

According to author [10] discuss perceptual image hashing, which generates content-based image hashes resilient to minor modifications like compression or color correction. These hashes are valuable for tasks such as duplicate image detection and reverse image searches. However, their effectiveness in detecting content-changing manipulations, like object addition/removal and copy-move, is underexplored. Recent research emphasizes the need for improved detection methods in this domain.

III. EXISTING AND PROPOSED WORK

Most existing plagiarism detection tools focus solely on textual content. These tools use keyword-based matching, fingerprinting, or sequence-matching techniques to identify similarities. Some of the well-known plagiarism detection systems include Turnitin, Grammarly, and Copyscape, which specialize in text-based plagiarism detection. However, these tools have several limitations:

- **Lack of Image Analysis:** Existing tools do not effectively detect plagiarized images or modified visual content.
- **Limited Contextual Understanding:** Many tools fail to detect sophisticated paraphrasing and rely heavily on exact word matching.
- **Dependence on Online Databases:** Some plagiarism checkers only compare content against a limited set of online sources, making them ineffective for detecting plagiarism from unpublished or offline sources.

To overcome these challenges, this project introduces an integrated system that can analyze both text and images, providing a more holistic approach to plagiarism detection.

Proposed Solution

The integrated plagiarism detection system combines text analysis and image processing techniques to enhance the detection of duplicated content. The system consists of two major components:

1. **Text-Based Plagiarism Detection:**
 - Utilizes natural language processing and machine learning algorithms to analyze textual content.
 - Compares text against multiple sources, including academic papers, research articles, online content, and proprietary databases.
 - Identifies paraphrased content and subtle textual modifications.
2. **Image-Based Plagiarism Detection:**
 - Employs deep learning-based computer vision models to detect duplicate images.

- Analyzes image metadata, structural features, and content similarity to identify plagiarized images.
- Detects alterations such as cropping, resizing, and watermark removal.

By integrating these two modules, the system ensures that documents are comprehensively analyzed for both textual and visual plagiarism. This improves accuracy and reduces the chances of undetected plagiarism.

IV. RESULTS

To develop an effective plagiarism detection system, a combination of advanced technologies and algorithms is required. Here we used following technologies like Natural Language Processing, Machine Learning and Deep Learning, Computer Vision, Database Management and Cloud Computing.



Figure 1: Project / Portal Home Page

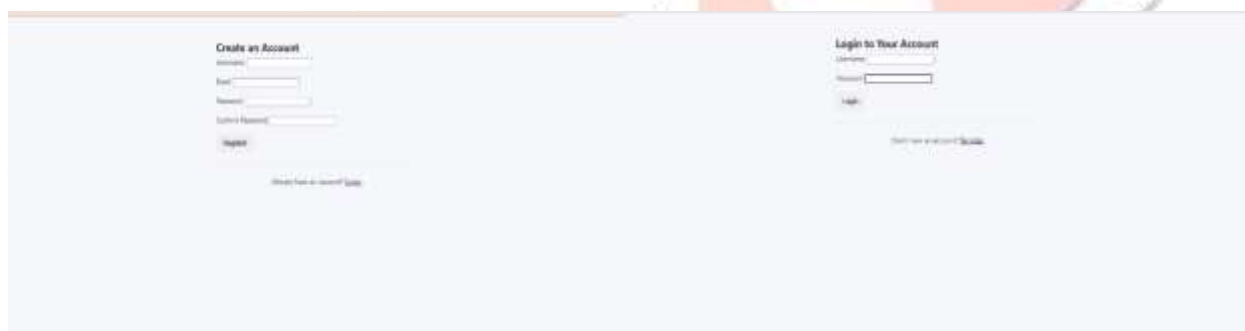


Figure 2: User Registration and Login Page

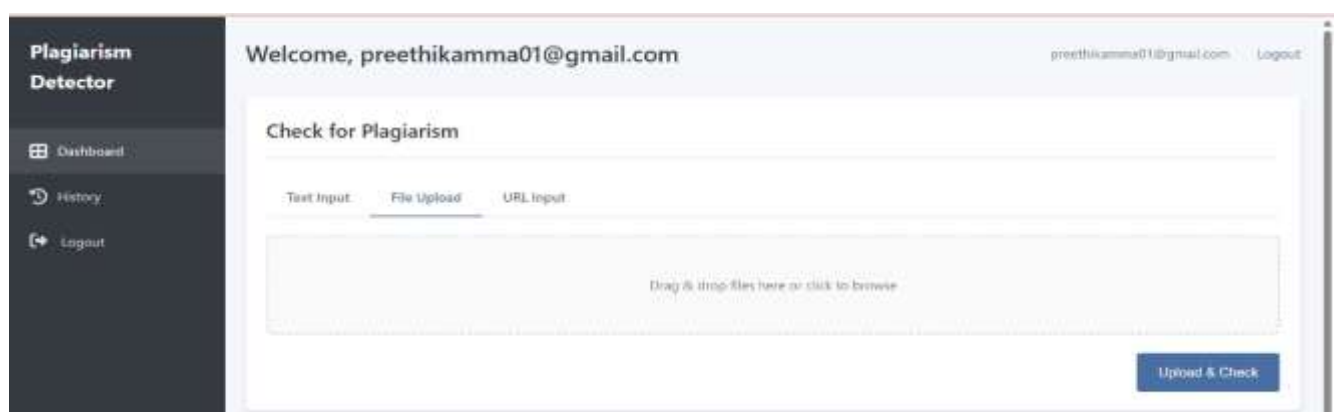


Figure 3: Portal View once the user gets login

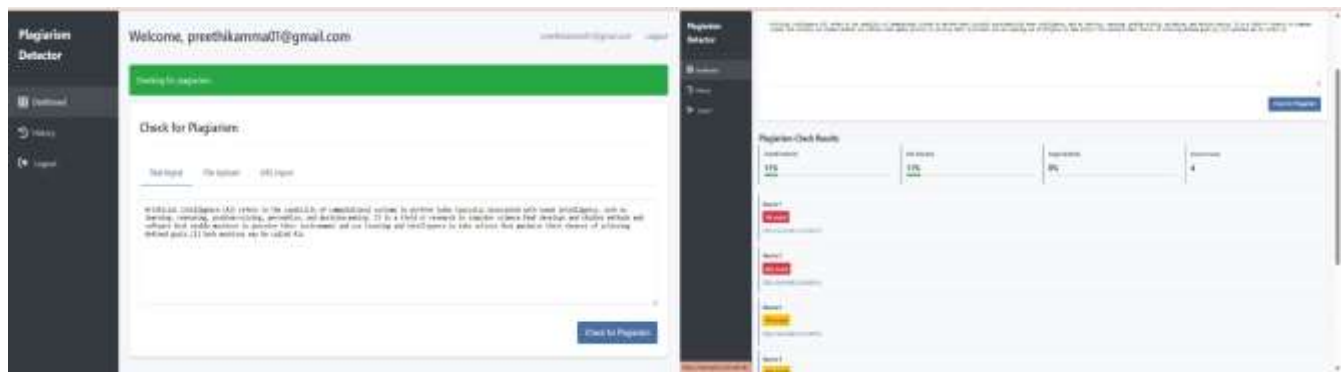


Figure 4: Checking Plagiarism for Text Data with its plagiarism Percentage

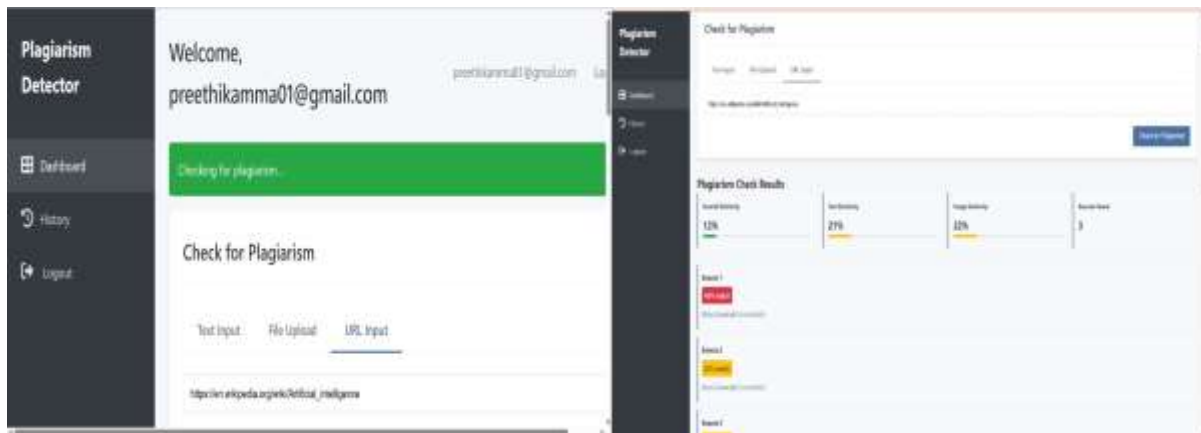


Figure 5: Checking Plagiarism for URL Based Content like Blogs etc.,

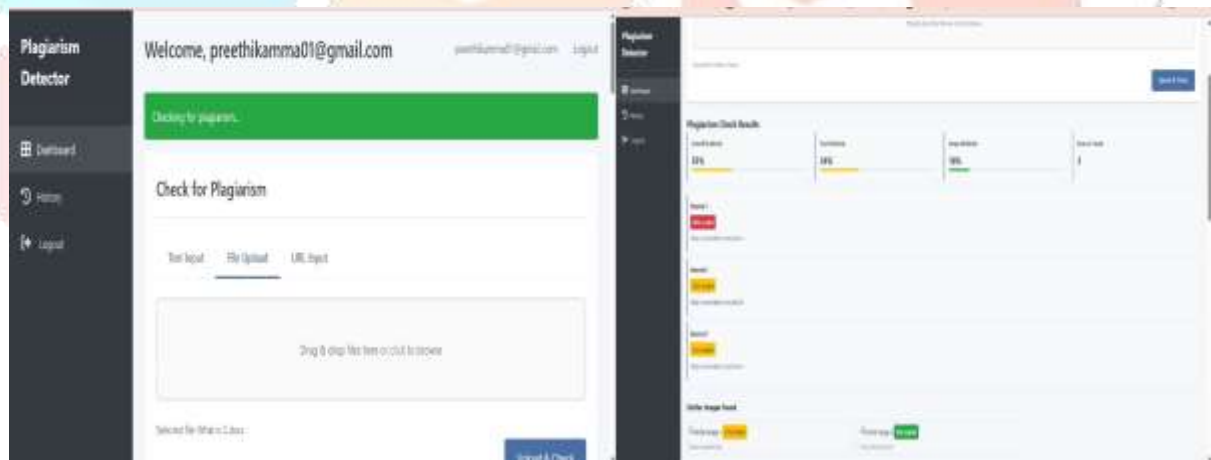


Figure 6: Checking Plagiarism for Image Based Content in the file

V. APPLICATIONS AND ETHICAL CONSIDERATIONS

Applications of the System

The proposed integrated plagiarism detection system can be applied in various domains like Academic Institutions, Publishing Industry, Legal and Copyright Protection, Corporate Sector

Ethical Considerations

While developing a plagiarism detection system, ethical considerations must be taken into account like Privacy Protection, Fair Use and Attribution, Bias in Detection Algorithms.

VI. CONCLUSION & FUTURE WORK

CONCLUSION

This integrated plagiarism detection system addresses the shortcomings of traditional tools by combining natural language processing for text analysis and deep learning-based computer vision for image similarity detection, crucial in today's digital content-heavy environment across academia, research, and professional fields; it enhances accuracy by identifying direct textual matches and modified images, providing a holistic authenticity evaluation, overcoming limitations of existing systems that often miss paraphrased content and image duplication through advanced techniques like semantic analysis, word embeddings, transformer models, convolutional neural networks, and feature extraction. Designed for scalability, efficiency, and user-friendliness, it leverages cloud processing, distributed computing, and optimized indexing for fast, accurate detection, even with large document volumes, and features an intuitive interface for seamless uploads and detailed reports, while ensuring robust security and privacy compliance. Its adaptability across education, publishing, corporate, and legal domains allow for verification of research papers, protection of proprietary materials, and copyright enforcement. By integrating text and image analysis, it significantly improves detection accuracy, minimizing false positives and negatives, and provides comprehensive reports with clear source references. Ultimately, this system, powered by AI, machine learning, and cloud computing, offers a valuable solution for maintaining content originality and intellectual integrity, promoting ethical writing practices and protecting intellectual property rights.

FUTURE WORK

The integrated plagiarism detection system, which combines natural language processing for text and deep learning-based computer vision for image analysis, holds significant potential for ensuring content originality across diverse sectors like academia, publishing, and legal frameworks, but it can be substantially enhanced for future applications; improvements include leveraging advanced deep learning models, enabling cross-language plagiarism detection, and incorporating handwritten text recognition and context-aware paraphrase detection for text; for images, more robust modification detection, video and 3D model plagiarism detection, and Content-Based Image Retrieval (CBIR) integration are vital; scalability and performance can be boosted through edge computing, blockchain for secure verification, federated learning for privacy, and cloud-based AI service integration; user experience can be improved with voice and speech plagiarism detection, AI-assisted interfaces, customizable rules, and interactive reports; ethical and legal compliance necessitates AI ethics and bias mitigation, copyright protection agency integration, social media content detection, and real-time alerts; future integration with educational and corporate systems, such as Learning Management Systems (LMS), research databases, and enterprise-level monitoring, alongside collaboration with publishers, will further solidify its utility; by focusing on automation, reduced computational costs, and high accuracy, while maintaining compliance with global standards, this system will evolve into a crucial tool for upholding academic integrity, protecting intellectual property, and promoting ethical content creation.

REFERENCES

- [1] P. S. Kumar and K. Prasad, "A Comprehensive Survey of Advanced Image Processing and OCR Techniques for Enhanced Image Plagiarism Detection," J. Eng. Sci., vol. 12, no. 2, pp. 137-146, 2021, doi: 10.52783/jes.4488.
- [2] P. S. Kumar and K. Prasad, "Integrating OCR and NLP Techniques for Accurate Text Extraction and Plagiarism Detection in Image-Based Content," Bapas, vol. 44, no. 2, pp. 1- , 2024, doi: 10.48165/bapas.2024.44.2.1.
- [3] Srinivas, P. K. & Krishna Prasad, K. (2024). Integrating Advanced OCR and NLP Techniques for Enhanced Text Extraction and Image Plagiarism Detection. International Journal of Applied Engineering and Management Letters (IJAEML), 8(2), 198-207.DOI: <https://doi.org/10.5281/zenodo.14292114>
- [4] D. S. K. Palvadi and D. K. . Prasad, "A Unified Framework for Text Extraction and Plagiarism Detection in Image-Based Content Using OCR and NLP", CDF, vol. 54, no. 1, pp. 132–141, Jan. 2025, doi: 10.48047/CU/54/01/132-141.
- [5] Dr Palvadi Srinvas Kumar, Dr PRaveen B M "Integrated Plagiarism Detection System for Text and Image Based Content in Document" which is published in "Journal of Information Systems Engineering and Management", Vol. 10 (24s) 2025, page number 177 – 183.
- [6] "Extracting text from image document and displaying its related information", K.N. Natei journal of Engineering Research and Application (ISSN : 2248-9622, Vol. 8, Issue5 (Part -V) May2018.
- [7] .J. Pradeep, E. Srinivasan and S. Himavathi, "Diagonal Based Feature Extraction For Handwritten Alphabets Recognition System Using Neural Network", International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.
- [8] "Text Recognition using image processing", International journal of Advanced Research in Computer Science by Chowdhury Md Mizan, Tridib Chakraborty and Suparna Karmakar (Vol-8,No.5,MayJune 2017).
- [9] A. Chitra et al., "Plagiarism Detection Using Machine Learning-Based Paraphrase Recognizer," Journal of Intelligent Systems, October 2014.
- [10] Sk. Mahaboob Basha et al., "Text and Image Plagiarism Detection," 2022.
- [11] Senosy Arrish et al., "Shape-Based Plagiarism Detection for Flowchart Figures in Texts," International Journal of Computer Science & Information Technology (IJCSIT), vol. 6, no. 1, February 2014.