ISSN 2277-2685
IJESR/April-June. 2026/ Vol-16/Issue-2/225-232
Ragi Shailaja *et. al.*, /International Journal of Engineering & Science Research
Full Length Article

Text And Image Plagiarism Detection

G.Srilakshmi[1], T.Kinnera[2], S.Nandini[3], B.Bhavani[4]
[1]Associate Professor; Department Of Electronics And Communication Engineering Bhoj Reddy Engineering College For Women Hyderabad India
[2,3,4]B.Tech Students; Department Of Electronics And Communication Engineering Bhoj Reddy Engineering College For Women Hyderabad India
Mail Id; nandinisama9@gmail.com[3]

Accepted 08-04-2026

Abstract
Plagiarism detection has become an essential requirement in academic institutions, journals, and research organizations to ensure originality in submitted documents and projects. Most existing plagiarism detection systems focus primarily on textual similarity and are effective in identifying copied or paraphrased text. However, these systems often fail to address plagiarism involving images, design files, and graphical content. To overcome this limitation, this paper proposes an adaptive, scalable, and extensible plagiarism detection framework capable of analyzing both textual and image-based content.The proposed model employs a text corpus containing 100 reference documents to identify suspicious textual similarities. For image plagiarism detection, a histogram-based feature extraction technique is used to create feature representations of stored database images. Suspicious images are then compared against the stored histogram models to measure similarity. In addition, clustering algorithms are incorporated to group extracted image features and improve matching robustness without discarding important visual characteristics. A configurable similarity threshold of 40% is used for detection, which can be adjusted according to institutional requirements.Experimental evaluation was conducted using a dataset consisting of 10 training design samples stored in the system database and multiple test samples containing both original and forged designs. The proposed framework achieved a 100% matching rate and an overall detection accuracy of 81%. Results demonstrate that integrating text comparison with intelligent image analysis significantly improves plagiarism identification beyond traditional text-only systems. The developed approach is suitable for educational institutions, publishing agencies, and digital repositories seeking reliable multi-format plagiarism detection solutions.

Keywords: *Plagiarism Detection, Image Plagiarism, Text Similarity, Histogram Features, Clustering Algorithm, Artificial Intelligence, Document Originality, Turnitin Alternative.*

Introduction

In the modern digital age, the availability of online information has increased rapidly. Students, researchers, and professionals can easily access documents, images, reports, and multimedia resources through the internet. While this has improved learning and knowledge sharing, it has also increased the misuse of content through plagiarism. Plagiarism is the act of copying another person's work and presenting it as original without proper acknowledgment. It affects academic honesty, creativity, and intellectual property rights. Therefore, detecting plagiarism has become an important requirement in educational institutions, publishing organizations, and research communities.Most traditional plagiarism detection tools are designed mainly for text comparison. These systems compare documents with online sources or stored databases to identify copied sentences or paragraphs. Although they perform well for direct copying, they often fail when the text is paraphrased or slightly modified. Another major limitation is that many existing tools do not support image plagiarism detection. Images can be copied and changed through resizing, cropping, rotation, brightness adjustment, or minor editing, making detection more difficult.The proposed Text and Image Plagiarism Detection system is developed to overcome these limitations by detecting plagiarism in both textual and visual content. For text analysis, the system uses corpus-based comparison methods, keyword extraction, and similarity analysis. These methods help identify copied content even when the wording has been changed. For image analysis, Histogram and Perceptual Hashing (PHash) techniques are used to compare visual patterns and structures. These methods can detect similarity even when the image is resized, rotated, or slightly modified.The system provides a simple and secure user interface where users can register, log in, upload source files, upload suspicious content, and obtain plagiarism reports. It supports multiple formats such as TXT, PDF, DOCX, JPG, and PNG. The system is designed to reduce search time by using optimized feature extraction techniques instead of comparing complete files directly. It also ensures secure handling of user data during processing.Overall, the proposed system offers an effective solution for detecting plagiarism in academic and creative content. It improves detection accuracy, supports both text and images, reduces

processing time, and helps maintain originality standards in institutions and organizations.

Literature Survey

Many researchers have worked on plagiarism detection and fake content identification using machine learning and image processing techniques. Earlier systems mainly focused on text plagiarism using string matching, keyword comparison, and similarity measurement methods. These systems were useful for direct copying but were less effective for paraphrased content.Researchers also studied the detection of false news and manipulated online content by combining text and image information. Hybrid models such as Text-Image Convolutional Neural Networks (TI-CNN) were developed to analyze both textual and visual data together. These models showed improved accuracy compared with text-only systems.Machine learning algorithms such as Support Vector Machine (SVM) and Complement Naïve Bayes (CNB) have also been used for text classification tasks. In some studies, these methods were applied to detect fake accounts and harmful content on social media platforms. The results demonstrated that machine learning can effectively analyze patterns in text data.For image forgery detection, researchers proposed methods based on resampling features, local descriptors, and deep learning networks. Some systems used Long Short-Term Memory (LSTM) models to identify manipulated image regions.

Machine Learning

Machine Learning is a branch of Artificial Intelligence that enables computers to learn from data and make decisions without being explicitly programmed for every task. Instead of following fixed instructions, machine learning systems analyze patterns in data and improve their performance over time.Machine learning is widely used in image recognition, speech processing, recommendation systems, medical diagnosis, fraud detection, and language translation. It is especially useful when handling large volumes of data where manual analysis is difficult.

In plagiarism detection systems, machine learning helps analyze similarities in text and images. It can identify hidden patterns, repeated structures, and suspicious similarities more accurately than traditional rule-based systems.

Features of Machine Learning

Machine learning has several important features:

Data Driven: It learns patterns directly from datasets.

Automation: It reduces manual effort in analysis and decision-making.

Prediction: It can predict future outcomes or classify unknown data.

Adaptability: Models improve when more training data becomes available.

Scalability: It can process very large datasets efficiently.

Generalization: It applies learned knowledge to new unseen data.

Need for Machine Learning

The need for machine learning has increased due to the growth of digital data. Traditional programming methods cannot efficiently process complex tasks such as text understanding, image recognition, or recommendation systems. Machine learning solves these problems by learning from examples.In plagiarism detection, machine learning improves the ability to detect copied, paraphrased, or manipulated content. It increases accuracy, saves time, and supports large-scale analysis for universities, publishers, and organizations.

Text and Image Plagiarism Detection

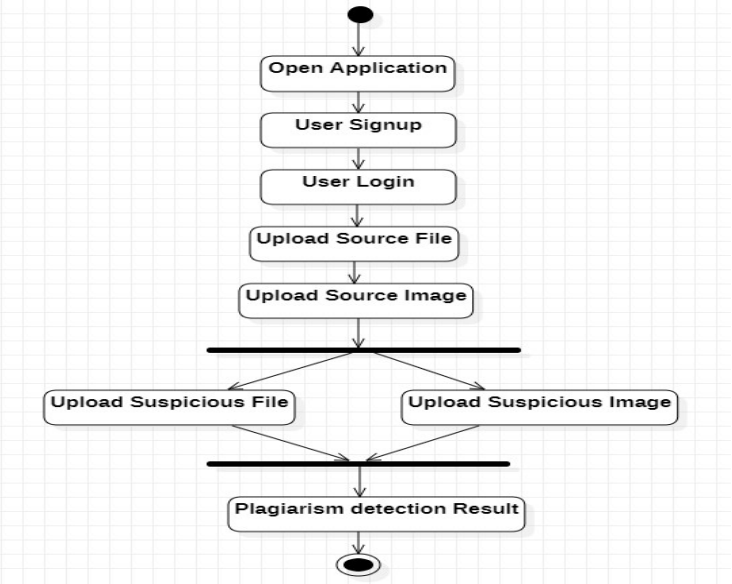
Existing System

Existing plagiarism detection systems mainly focus on text comparison. They use basic matching techniques to compare documents and identify copied sentences or paragraphs. These systems work effectively only when the copied text is directly duplicated or slightly modified.However, they fail to detect paraphrased content accurately. Most of these systems are not designed for image plagiarism detection. They cannot identify copied images that are resized, rotated, cropped, or edited. The comparison process is often slow when dealing with large databases. Accuracy is also low when images are transformed using brightness or contrast changes.Because of these limitations, existing systems are not fully reliable for academic and creative applications.

Proposed System

The proposed system is developed to detect plagiarism in both text and images with better accuracy and faster performance. Users can upload suspicious text files or images and compare them with original content stored in the database.For text detection, the system uses corpus-based algorithms to analyze keywords, sentence structure, and similarity scores. This helps detect copied or paraphrased text.For image detection, Histogram and Perceptual Hashing (PHash) methods are used to extract visual features and compare images. These techniques help identify similarity even when the image is rotated, resized, cropped, or slightly edited.The system supports multiple file formats and automatically updates the database to improve future detection quality. A user-friendly interface allows users to register, log in, upload files, and view plagiarism reports with percentage results.The proposed system reduces search time through optimized feature extraction and provides secure temporary storage of uploaded files. It is scalable for large datasets and suitable for universities, publishers, and research organizations.

3.4 Block Diagram



Block Diagram of Text and Image Plagiarism Detection

Methodology

The system uses separate methodologies for text and image plagiarism detection.

For text plagiarism, Natural Language Processing techniques are used to compare uploaded text with a reference corpus. The system analyzes keywords, sentence similarity, and matching patterns.For image plagiarism, Histogram comparison and Perceptual Hashing methods are used. These techniques generate image signatures based on visual structure and compare them with stored images. Similarity is identified even if the image has minor modifications.Feature extraction is used to convert raw text and images into compact numerical representations. This reduces comparison time and improves system efficiency. A similarity threshold is applied to determine whether plagiarism exists.Thus, the methodology combines text analytics and image processing to provide an accurate and reliable plagiarism detection system.

Software Design

Computational Resource Requirements

Text plagiarism detection generally requires preprocessing tasks such as tokenization, stop-word removal, stemming, and vector generation. Low-complexity methods such as fingerprinting or direct matching need relatively low CPU and memory resources. Moderate-complexity methods such as TF-IDF and Longest Common Subsequence require higher processing capability. Advanced semantic methods using transformer-based language models require significant computational resources, including GPU acceleration.Image plagiarism detection is comparatively more resource intensive because visual data must be processed at pixel or feature level. Feature-based approaches such as SIFT and SURF demand high computational power for keypoint extraction and matching. Perceptual hashing methods are more efficient and suitable for real-time applications. Deep learning models such as Convolutional Neural Networks require high-end GPUs, large RAM, and optimized storage systems, especially for large-scale image databases.

Software Requirements

The development environment can be implemented on Windows 10 or later operating systems. Python is used as the primary programming language because of its simplicity, extensive libraries, and strong support for machine learning and image processing. Django or Flask frameworks can be used for backend development to manage user requests, authentication, database operations, and application logic.The frontend interface can be developed using HTML, CSS, and JavaScript to provide an interactive web environment. MySQL is suitable for storing user data, uploaded files, metadata, and extracted features. Additional Python libraries such as OpenCV, NumPy, Pandas, Scikit-learn, and NLTK may be used for text and image processing.

Hardware Requirements

The system can operate on a standard personal computer with an Intel Core i5 processor or equivalent. A minimum of 4 GB RAM is required, while 8 GB or higher is recommended for smoother performance when processing multiple files. Storage of at least 250 GB is sufficient for datasets, software tools, and result logs. For advanced machine learning models, higher RAM and GPU support are desirable.

System Design

System Architecture

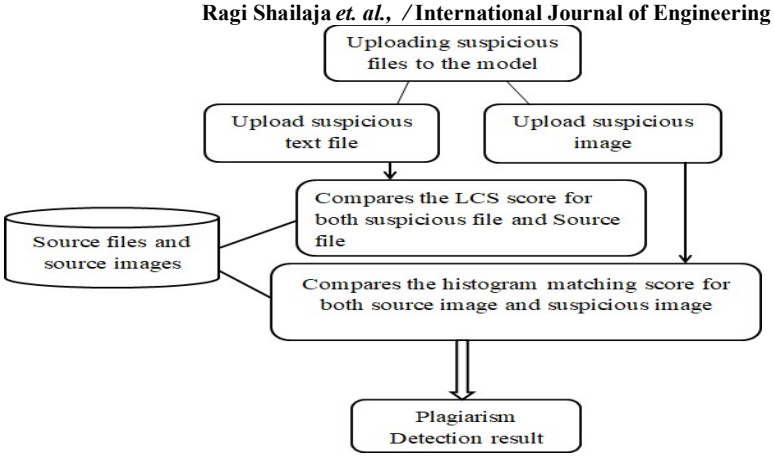


Figure.1 System Architecture

The system architecture of the Text and Image Plagiarism Detection project is designed as a multi-layer framework consisting of user interface, processing modules, database storage, and reporting components. Users interact with the system through a web-based interface where they register, log in, and upload documents or images.Once content is submitted, it is forwarded to the processing layer. The text analysis module compares uploaded textual data with the reference corpus using similarity measures. The image analysis module extracts visual features using histogram analysis and perceptual hashing techniques. The resulting similarity scores are sent to the reporting module, which generates plagiarism results. All source files, extracted features, and user records are managed in the database layer.

Technical Architecture

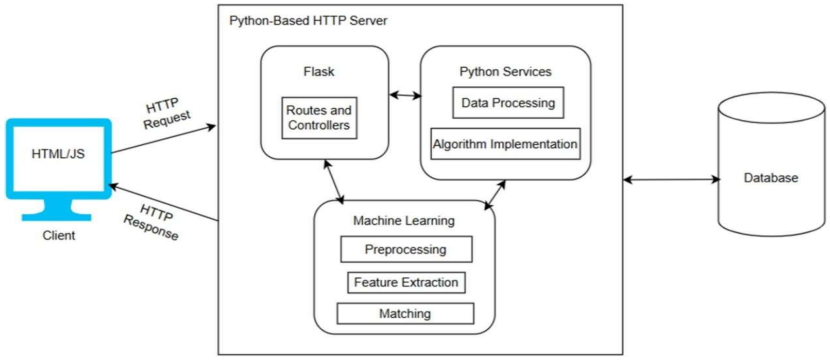


Figure : 2 Technical Architecture

The technical architecture follows a modular design. The frontend interface is built using HTML, CSS, and JavaScript, enabling users to upload files and view results. The backend server, implemented using Django or Flask, receives user requests and coordinates all processing tasks.Two independent pipelines operate in the backend. The first is the Text Processing Module, which performs preprocessing, tokenization, corpus comparison, and similarity calculation. The second is the Image Processing Module, which applies histogram extraction, perceptual hashing, and feature comparison.The database layer stores user profiles, source datasets, suspicious uploads, extracted metadata, and result logs. This architecture supports scalability, maintainability, and future integration of advanced machine learning algorithms.

Results and Discussion

System Modules and Output Analysis

The first module is the User Signup Module, which enables new users to create accounts by entering personal details such as name, email address, and password. Input validation ensures that invalid data formats are rejected. Successful registrations are securely stored in the database and confirmation messages are displayed.The Login Module authenticates registered users by verifying credentials against stored records. Authorized users are redirected to the dashboard, while invalid attempts are denied with suitable error messages. This module strengthens security by restricting unauthorized access.The Upload Source File Module allows users to submit original text documents in formats such as TXT, PDF, and DOCX. Uploaded files are validated for format and size, then text content is extracted and stored for future plagiarism comparison.The Upload Suspicious File Module is used to submit text

documents that need to be checked for plagiarism. The system preprocesses the text, extracts features, and compares the content with stored source files using similarity analysis methods.The Upload Source Image Module accepts original reference images in standard formats such as JPG and PNG. Images are normalized and processed using histogram extraction and perceptual hashing techniques before storage in the database.The Upload Suspicious Image Module enables users to upload images for plagiarism verification. The system extracts image signatures using pHash or other visual descriptors and compares them with stored image datasets to determine duplication or modification levels.The Logout Module securely terminates user sessions, clears active session variables, and redirects users to the login page. This prevents unauthorized access after session completion.

Test Cases and Results

Functional testing was performed on all modules of the application. New user registration was successful when valid details were entered, while incorrect email or mobile formats were rejected. Login functionality correctly accepted valid credentials and denied invalid usernames or passwords.For text plagiarism detection, suspicious documents uploaded to the system were successfully analyzed, and similarity reports were generated. The framework correctly identified copied or partially modified text passages.For image plagiarism detection, suspicious images were compared with the source image database. The system successfully detected duplicated images and visually similar images even after resizing or minor edits. Histogram graphs and similarity outputs were displayed as part of the result report.Overall testing indicated that all major modules performed as expected, with consistent outputs and successful execution of validation, detection, and reporting processes.

Discussion

The results indicate that integrating text analysis and image comparison in a single platform significantly improves plagiarism detection capability. Traditional systems mainly focus on text, whereas the proposed model extends verification to graphical and visual content.Text detection performance was effective for exact matches, partial copying, and moderate paraphrasing. Image detection performance was satisfactory for duplicate images, resized copies, and slightly modified versions. The use of histogram comparison and perceptual hashing reduced computational complexity while maintaining acceptable accuracy.The system also demonstrated strong usability through a simple interface and fast response times for moderate datasets. Therefore, the developed framework can be considered practical for institutions, publishers, and organizations requiring multimodal plagiarism verification.

Output

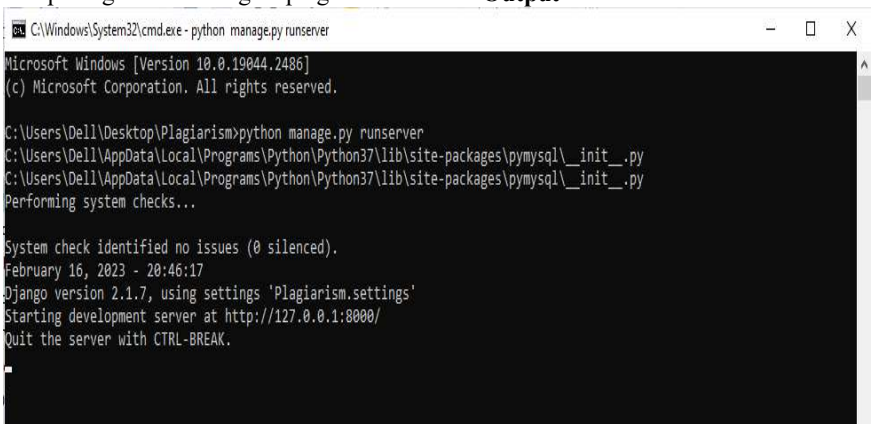


Figure : 1 manage.py command prompt image

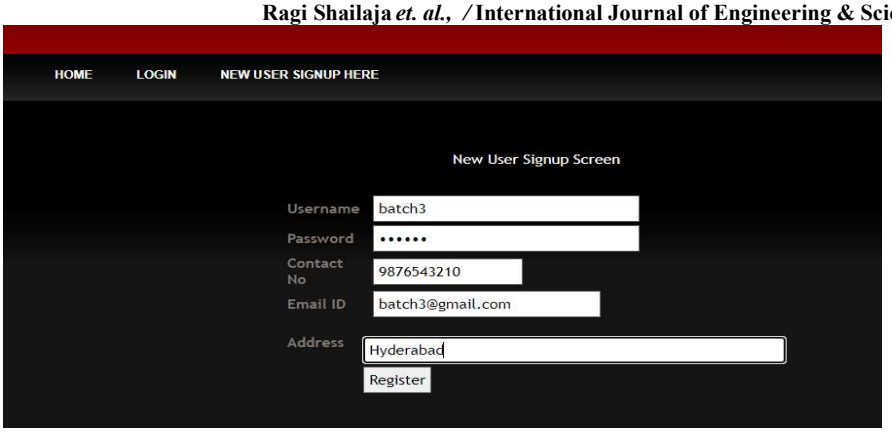


Figure : 2 User SignUp

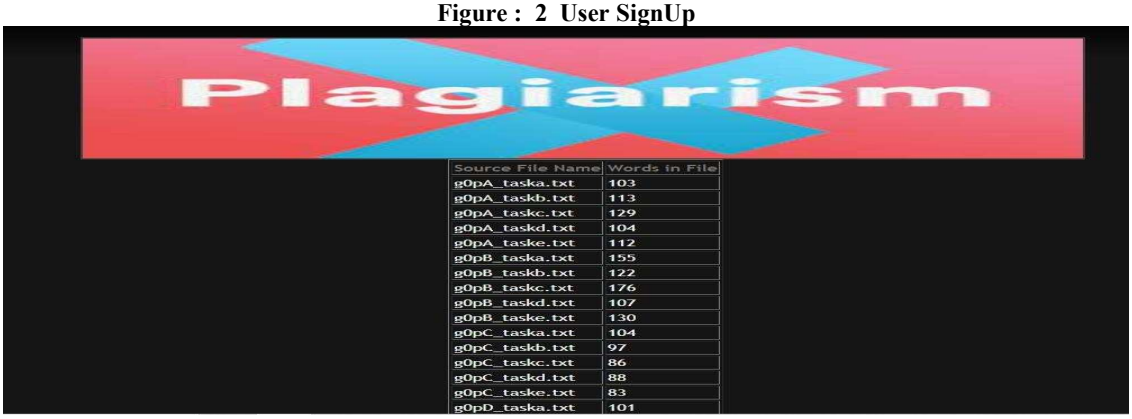


Figure : 3 uploaded source files

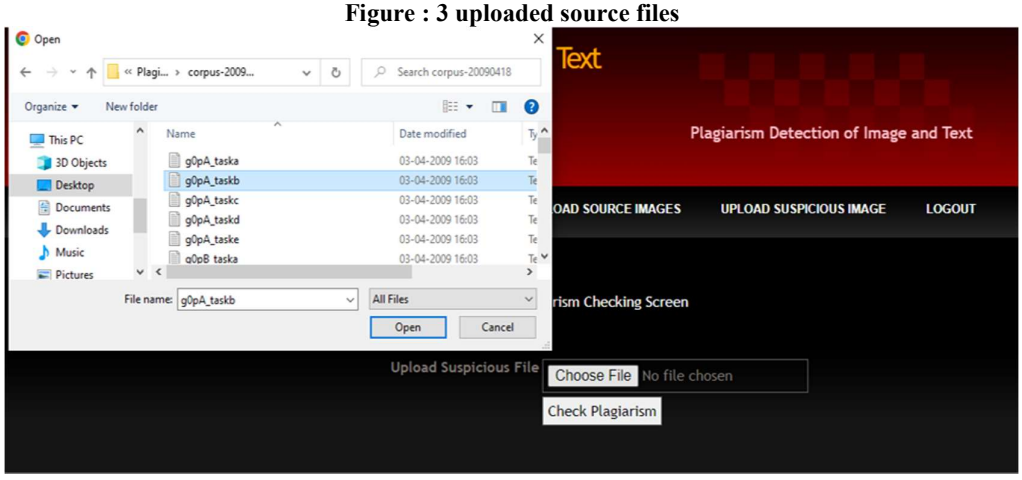


Figure : 4 Upload source file

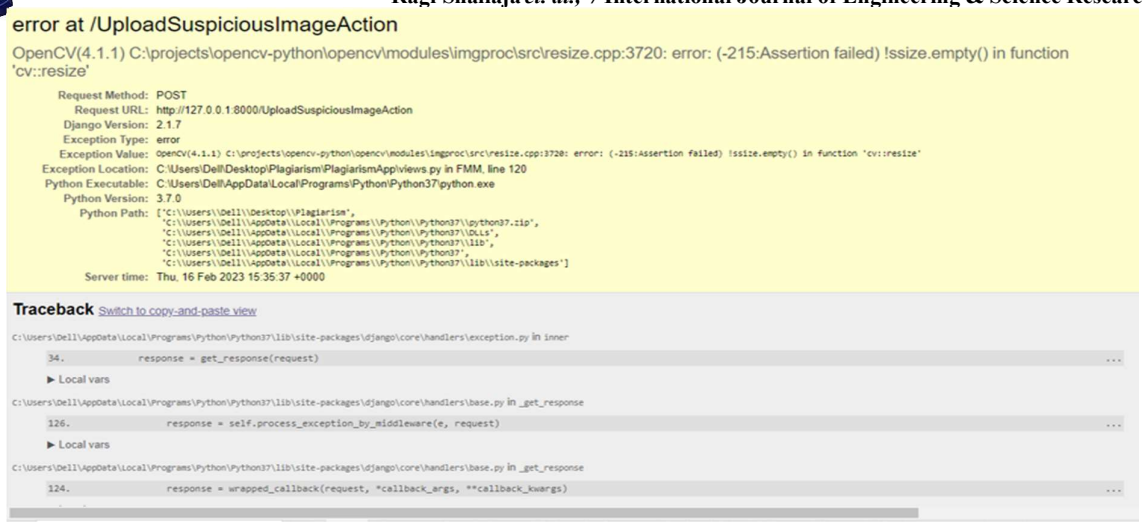


Figure : 5 Error due to uploading text file instead of image

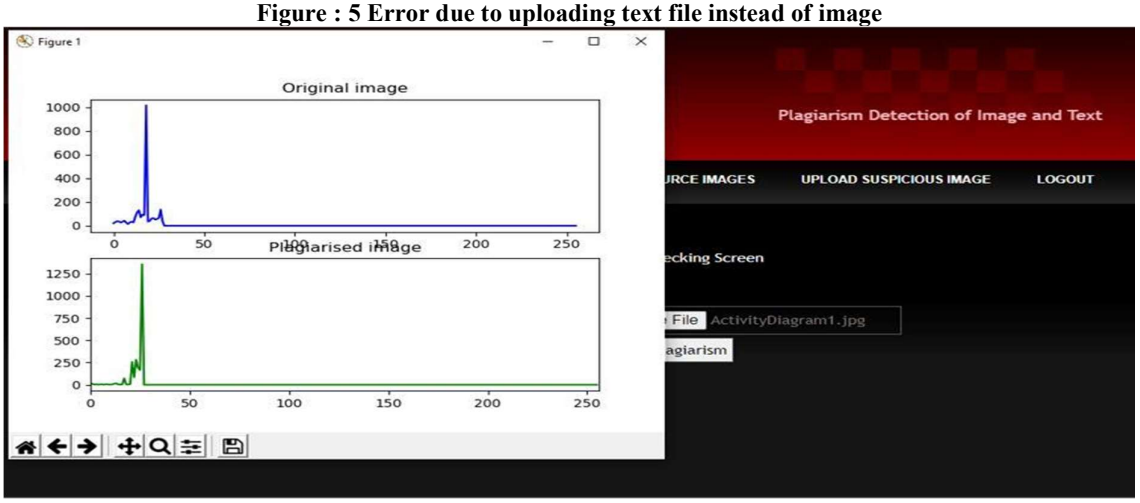


Figure : 6 Histogram values

Figure : 7 plagiarism detection image

Applications

The proposed plagiarism detection system can be applied in many domains. In academic institutions, it can verify originality in assignments, theses, dissertations, and research papers. Publishing houses can use it to detect unauthorized reuse of written or visual content before publication.Online platforms can apply the system to identify duplicate website content and maintain quality standards. Designers, photographers, and content creators can use image plagiarism detection features to protect digital assets from misuse.The system can also be extended for multimedia document auditing, where both embedded text and images in reports or presentations are analyzed together. In computer science education, similar concepts may be adapted for source code plagiarism detection.

Conclusion

The proposed Text and Image Plagiarism Detection system provides an efficient and practical solution for identifying duplicated content in multiple formats. Existing plagiarism tools often concentrate only on text matching and fail to address image-based plagiarism. The developed framework overcomes this limitation by combining corpus-based text analysis with visual similarity techniques such as histogram comparison and perceptual hashing.The system successfully supports user registration, secure login, source dataset management, suspicious file uploads, plagiarism analysis, and result reporting. Experimental testing confirmed that the framework can accurately detect copied text, paraphrased content to a reasonable extent, and visually similar images even after transformations such as resizing or cropping.By integrating textual and visual verification in a unified platform, the project improves academic integrity, protects intellectual property, and promotes originality in digital content creation. Therefore, the proposed model is suitable for educational institutions, publishers, and organizations that require reliable plagiarism monitoring.

Future Scope

The system can be further improved by expanding the source database with larger collections of documents and images from diverse domains. A larger reference dataset would increase detection coverage and improve reliability.Future versions may integrate advanced Natural Language Processing models such as BERT or transformer-based embeddings for deeper semantic plagiarism detection and improved paraphrase recognition. For image analysis, convolutional neural networks can be introduced to detect complex manipulations, partial copying, and synthetic image generation.Cross-lingual plagiarism detection is another promising enhancement, enabling the system to identify copied material translated from one language to another. Cloud deployment and distributed databases can improve scalability for universities or enterprise-level usage.Additional features such as citation checking, automated report generation, API integration with learning management systems, and mobile application support can further enhance usability. With these improvements, the framework can evolve into a comprehensive next-generation plagiarism detection platform.

References

1. Clough, P., “Plagiarism in Natural and Programming Languages: Current Tools and Technologies,” University of Sheffield, 2000.
2. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., and Rosso, P., “Overview of the First International Competition on Plagiarism Detection,” CLEF, 2009.
3. Barrón-Cedeño, A., and Rosso, P., “Automatic Plagiarism Detection Based on N-gram Comparison,” ECIR, 2009.
4. Kakkonen, T., and Sutinen, E., “Automatic Detection of Plagiarism in Text Documents,” ACM Computing Surveys, 2010.
5. Swaminathan, A., Mao, Y., and Wu, M., “Robust and Secure Image Hashing,” IEEE Transactions on Information Forensics and Security, 2006.
6. Zauner, C., “Implementation and Benchmarking of Perceptual Image Hash Functions,” Master’s Thesis, 2010.
7. Mihcak, M. K., and Venkatesan, R., “Iterative Geometric Methods for Robust Perceptual Image Hashing,” ICIP, 2004.
8. Kang, X., Wei, S., and Zeng, X., “Survey on Image Copy Detection,” Journal of Visual Communication and Image Representation, 2016.
9. Manning, C. D., Raghavan, P., and Schütze, H., *Introduction to Information Retrieval*, Cambridge University Press, 2008.
10. Goodfellow, I., Bengio, Y., and Courville, A., *Deep Learning*, MIT Press, 2016.