International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 3 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

MULTI-DOCUMENT COMPARATOR(FIND DIFFERENCES ACROSS REPORTS)

[1] R Uma, [2] K Sruthi, [3] G Ashish Kumar, [4] P Jaasiel Mark, [5] B Shiva Pandu
[1]Assistant Professor, [2345]Students
Department of CSE(AIML)
Siddhartha Institute of Technology & Sciences, Narapally
umakola_cse@siddhartha.co.in, 23TQ1A6680@siddhartha.co.in, 23TQ1A66B5@siddhartha.co.in, 23TQ1A66B2@siddhartha.co.in, 23TQ1A66C0@siddhartha.co.in

Abstract

The **Multi-Document Comparator** is an intelligent document analysis system designed to compare multiple reports and identify differences, similarities, additions, deletions, and modifications between them. Comparing documents manually can be difficult and time-consuming, especially when the reports contain hundreds of pages or similar information presented in different formats. The proposed system uses Artificial Intelligence and Natural Language Processing to automate the comparison process.

The system allows users to upload two or more documents such as PDF files, Word documents, text reports, research papers, business reports, or project documents. The uploaded documents are first processed to extract their text and important structural information. The system then analyzes the content and compares corresponding sections, paragraphs, sentences, tables, and key information across the documents.

Natural Language Processing techniques are used to understand the meaning of the content rather than depending only on exact word matching. The system can identify whether information has been added, removed, changed, or rewritten using different words. It can also detect similar statements, contradictory information, changes in numerical values, and differences in important sections. This makes the comparison more meaningful than a basic text-difference tool.

The system presents the comparison results through an easy-to-understand interface. Differences can be highlighted using categories such as **Added, Removed, Modified, Similar, and Conflicting**. Users can view side-by-side document comparisons, section-level differences, and an overall summary of the changes. The system can also generate a comparison report that provides important findings without requiring the user to read every document again.

Overall, the **Multi-Document Comparator** reduces the manual effort required for document comparison and improves the accuracy and speed of identifying changes. It can be useful for students, researchers, organizations, legal teams, auditors, content reviewers, and business professionals. The system can be further enhanced with multilingual comparison, table comparison, image comparison, advanced contradiction detection, and integration with cloud document management platforms.

I. INTRODUCTION

Documents are widely used to store and communicate information in education, business, research, government, and many other fields. Often, multiple versions of the same report are created as information is updated or corrected. Users may need to compare these versions to understand what has changed. When documents are short, manual comparison may be possible, but comparing large reports with hundreds of pages becomes difficult and inefficient.

Traditional document comparison tools generally identify differences by comparing words or characters between two files. Although these tools are useful for basic editing tasks, they may not understand the actual meaning of the content. For example, two sentences can express the same idea using completely different words, while two similar-looking sentences may contain different numerical values or meanings. Therefore, semantic understanding is important for effective document comparison.

The **Multi-Document Comparator** uses Artificial Intelligence and Natural Language Processing to provide a more intelligent approach to document comparison. The system extracts content from multiple documents and divides it into meaningful units such as headings, paragraphs, sentences, and tables. It then analyzes these units and determines their similarities and differences. This allows the system to identify meaningful changes instead of simply highlighting every changed word.

The system can compare multiple versions of reports and classify detected changes into different categories. It can identify newly added information, deleted information, modified content, similar content, and potentially conflicting statements. Numerical values, dates, names, headings, and important keywords can also be compared. A summary of major changes helps users quickly understand the differences between documents.

The proposed system aims to make document comparison faster, more accurate, and easier to understand. It reduces repetitive manual work and allows users to focus on important changes rather than spending hours reading multiple versions of the same report. The system can support academic research, document review, auditing, business analysis, legal document checking, and other applications where accurate comparison of multiple documents is required.

II. LITERATURE SURVEY

Traditional document comparison techniques mainly use text-matching and string-difference algorithms to identify changes between documents. Algorithms such as character-level and word-level comparison can detect inserted, deleted, or modified text. These methods are useful for identifying exact changes but may produce many unnecessary differences when formatting or sentence structure changes. They also have limited ability to understand the meaning of the content.

Document analysis systems have introduced Natural Language Processing techniques to improve the understanding of textual information. NLP methods such as tokenization, sentence segmentation, keyword extraction, named entity recognition, and text similarity can be used to analyze document content. These techniques help systems identify important concepts and relationships within reports. However, traditional NLP

approaches may have difficulty understanding complex context and semantic relationships.

Semantic similarity techniques have been widely explored for comparing texts based on meaning. Word embeddings and sentence embeddings allow text units with similar meanings to be represented in a common semantic space. Similarity measurements can then be used to identify sentences or paragraphs that express related information even when different words are used. Such approaches are useful for detecting paraphrased content that basic text comparison systems may fail to identify.

Recent developments in Transformer-based language models and Large Language Models have significantly improved document understanding. These models can analyze long passages, summarize information, identify relationships, and compare statements based on their contextual meaning. They can also identify potential contradictions between different pieces of information. However, processing large documents with AI models can require significant computational resources and appropriate techniques for handling long contexts.

Research in document intelligence has also explored information extraction from PDFs, scanned documents, tables, and structured reports. Optical Character Recognition can convert scanned pages into machine-readable text, while document layout analysis can identify headings, tables, images, and other structural elements. Combining document extraction with semantic comparison can provide more complete analysis. Nevertheless, differences in document formats and layouts continue to create challenges.

Existing research demonstrates that combining traditional text comparison, NLP, semantic similarity, and Generative AI can provide a powerful solution for multi-document analysis. However, challenges remain in areas such as accuracy, contradiction detection, numerical comparison, long-document processing, and result interpretation. The proposed **Multi-Document Comparator** addresses these challenges by integrating document extraction, structural analysis, semantic comparison, change classification, conflict detection, and visual result presentation into a single system.

III. SYSTEM ANALYSIS

The **Multi-Document Comparator** is designed to analyze multiple reports and identify meaningful differences between them. The user first uploads two or more documents through the application interface. The system validates the files and extracts their textual and structural information. The preprocessing module removes unnecessary formatting and organizes the extracted content into headings, paragraphs, sentences, and tables. NLP techniques are then used to identify keywords, entities, dates, numbers, and important concepts. A document alignment module finds corresponding sections and content across the uploaded reports. The comparison engine performs both exact text comparison and semantic similarity analysis. Changes are classified as added, deleted, modified, similar, or conflicting information. The system also checks important numerical values, dates, and statements for inconsistencies. A result-generation module prepares a summary of the major differences and similarities. Users can view detailed changes through side-by-side or highlighted comparison views. Finally, the system

generates an overall comparison report that can be saved or exported for future reference.

Existing System

Existing document comparison systems mainly depend on traditional text-based comparison methods. Users generally upload two documents and the system compares their characters, words, or sentences to identify changes. These systems are effective when documents have only small textual modifications. However, they often struggle when the same information is rewritten using different words or sentence structures. Most basic comparison tools also focus primarily on comparing two documents rather than analyzing multiple reports together. Users may need to manually review large numbers of highlighted changes to determine which modifications are actually important. Differences in document formatting, headings, tables, and layouts can also make the comparison difficult. Some systems do not provide meaningful summaries of the detected changes. They may also fail to identify contradictions in numerical values, dates, or statements. Manual interpretation is therefore required after the comparison process. When reports are large, reviewing all the detected differences becomes time-consuming. As a result, traditional systems provide limited semantic understanding and are not ideal for complex multi-document analysis.

Disadvantages of Existing System

- Mainly depends on exact word or character matching.
- Limited understanding of the meaning of text.
- Difficulty detecting paraphrased information.
- Usually designed for comparing only two documents.
- Cannot effectively identify semantic contradictions.
- Limited comparison of tables and structured information.
- Large numbers of insignificant differences may be displayed.
- Requires manual interpretation of comparison results.
- Difficult to compare very large reports efficiently.
- Limited automatic summarization of important changes.
- Formatting and layout differences can affect comparison quality.
- Limited support for intelligent classification of changes.

Proposed System

The proposed **Multi-Document Comparator** uses Artificial Intelligence and Natural Language Processing to intelligently compare multiple reports. Users can upload two or more documents in supported formats such as PDF, DOCX, and TXT. The system extracts text and structural information while preserving important headings, sections, paragraphs, and tables. A preprocessing module cleans the extracted information and prepares it for analysis. NLP techniques identify important keywords, entities, dates, numbers, and concepts from each document. A semantic analysis module converts text into meaningful representations to identify content with similar meanings. The document alignment engine matches related sections and paragraphs across different reports. The comparison engine identifies additions, deletions, modifications, similarities, and potential conflicts. Special analysis is performed on important values such as dates, percentages, quantities, and names. The system generates a visual

comparison showing the detected changes in an understandable format. An AI summary module explains the major differences and provides key findings. Users can review, filter, export, and save the comparison results for future use.

Advantages of Proposed System

- Compares multiple documents efficiently.
- Provides semantic as well as exact text comparison.
- Detects paraphrased information with similar meanings.
- Identifies added, deleted, and modified content.
- Detects potentially conflicting information.
- Compares important numbers, dates, names, and values.
- Provides section-level and paragraph-level comparison.
- Automatically summarizes major differences.
- Reduces manual document-review effort.
- Provides easy-to-understand visual comparison results.
- Supports different document formats.
- Allows comparison results to be saved or exported.

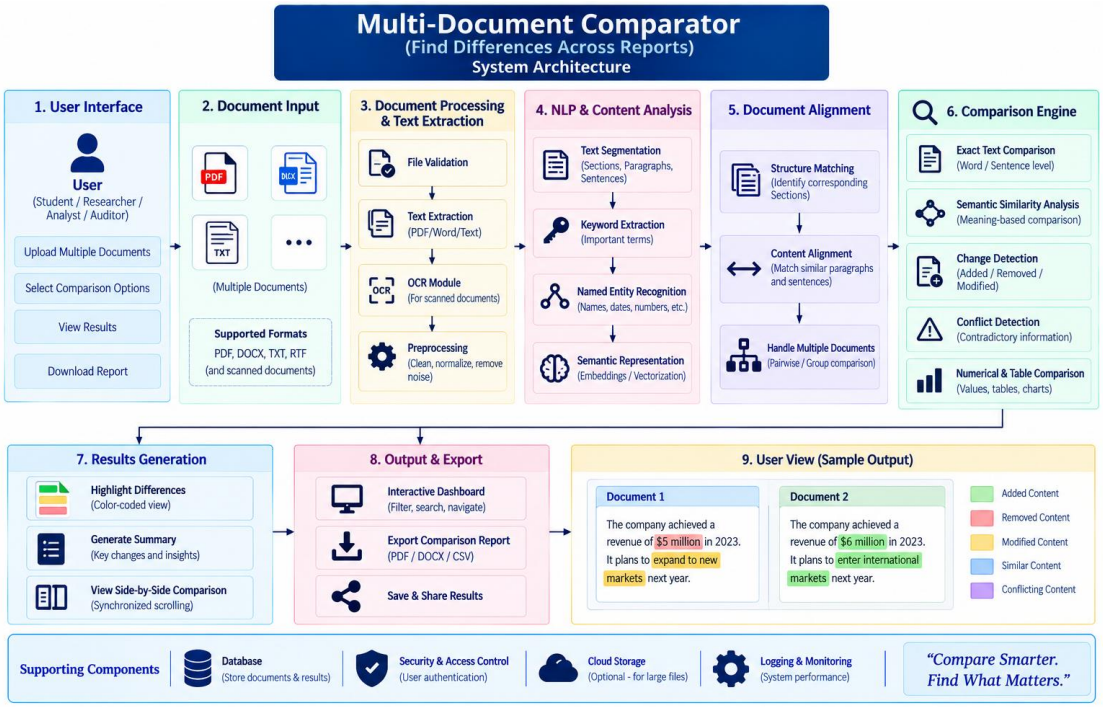
IV. METHODOLOGY

The methodology begins when the user uploads two or more reports through the Multi-Document Comparator interface. The system first validates the uploaded files and identifies their document formats. A document extraction module retrieves text, headings, paragraphs, tables, and other useful structural information. If required, OCR techniques are applied to extract text from scanned documents. The preprocessing module removes unnecessary formatting, normalizes text, and divides the documents into meaningful sections. NLP techniques are then applied to identify keywords, entities, dates, numbers, and important concepts. A semantic representation module generates embeddings for sentences and paragraphs to understand their meanings. The alignment module matches related sections and content across different documents. The comparison engine performs exact matching and semantic comparison to identify changes and similarities. A classification module categorizes each detected change as added, removed, modified, similar, or conflicting. An AI summarization module generates a concise explanation of the major differences. Finally, the results are displayed through a comparison dashboard and can be exported as a detailed comparison report.

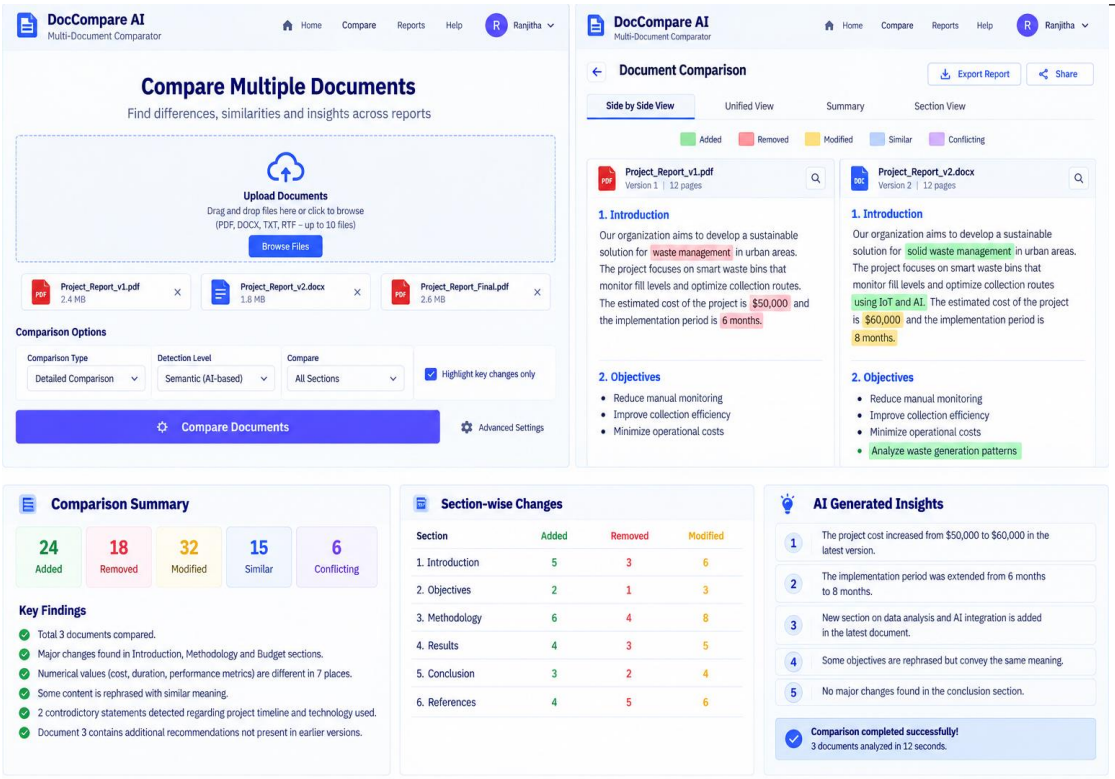
SYSTEM ARCHITECTURE

The system architecture consists of multiple modules that work together to analyze and compare reports intelligently. The User Interface allows users to upload multiple documents and select comparison options. The Document Upload and Validation Module checks file types, size, and document availability. The Document Extraction Module extracts text, headings, paragraphs, tables, and other structural information from the uploaded files. The OCR Module handles scanned or image-based documents when necessary. The Preprocessing Module cleans and normalizes the extracted content. The NLP Analysis Module identifies keywords, entities, concepts, dates, and numerical information. The Semantic Analysis Engine generates meaningful representations of text for semantic comparison. The Document Alignment Module

identifies corresponding sections and paragraphs between reports. The Comparison Engine detects additions, deletions, modifications, similarities, and conflicts. The AI Summary Module generates important findings and an easy-to-read comparison summary. Finally, the Comparison Dashboard and Export Module presents highlighted differences and allows users to save or export the final comparison report.



V. RESULT AND OUTPUT



VI. CONCLUSION

The Multi-Document Comparator provides an intelligent and efficient solution for comparing multiple reports and identifying meaningful differences between them. The system reduces the time and manual effort required to read and compare large documents individually.

By combining Natural Language Processing, semantic analysis, text comparison, and Artificial Intelligence, the system can identify added, removed, modified, similar, and conflicting information. It can also compare important values such as dates, numbers, names, and other key information across different versions of a report.

The system presents the detected changes through a user-friendly dashboard with highlighted differences, section-wise comparisons, summaries, and AI-generated insights. This makes it easier for users to quickly understand the important changes without reviewing every page manually. Users can also export and save the comparison results for future reference.

Overall, the proposed system makes document comparison faster, smarter, and more reliable. It can be useful for students, researchers, businesses, auditors, legal professionals, and organizations that frequently work with multiple versions of reports. In the future, the system can be enhanced with multilingual document comparison, advanced table and image comparison, improved contradiction detection, real-time collaboration, and integration with cloud-based document management systems.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.
[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.
[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.