International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.040 ISSN: 2581-4621

DESIGN AND IMPLEMENTATION OF DOCUMENT SUMMARIZER WITH REALTIME NEWS INTEGRATION

Prof. Suvarna Sujit Wakchaure[1], Mr. Darshan Mangesh Jadhav[2], Mr. Dhanraj Rajendra Dingar[3], Miss. Neha Sushil Gupta[4], Mr. Rehan Abid Tamboli[5]

[1]Assistant professor , Dept.of Computer Engineering ,Sir Visvesvaraya Institute Of Technology, Nashik , Maharashtra,India

[2]UG, Dept.of Computer Engineering ,Sir Visvesvaraya Institute Of Technology, Nashik, Maharashtra , India

[3]UG, Dept.of Computer Engineering ,Sir Visvesvaraya Institute Of Technology, Nashik, Maharashtra , India

[4]UG, Dept.of Computer Engineering ,Sir Visvesvaraya Institute Of Technology, Nashik, Maharashtra , India

[5]UG, Dept.of Computer Engineering ,Sir Visvesvaraya Institute Of Technology, Nashik, Maharashtra , India

Abstract

In today's fast-paced digital era, individuals encounter vast amounts of information through documents, articles, and reports. Reading and comprehending all this content can be time-consuming and challenging. To address this issue, this project presents an Intelligent Document Summarization System that automatically condenses lengthy documents while preserving their key ideas and essential meaning. The system supports multiple file formats, including PDF, TXT, and DOCX, and employs advanced text extraction and summarization techniques to produce concise and coherent summaries. It also offers multilingual support for English, Hindi, and Marathi, enhancing accessibility for diverse users. Additionally, the system integrates real-time related news retrieval, enabling users to stay informed on topics relevant to their documents. Users can easily copy, download, or share summaries, and with secure login and authentication, they can save and manage their summary history. Overall, this project delivers a smart, efficient, and user-friendly solution for managing and understanding large volumes of information, ultimately helping users save time and enhance productivity.

Key Words: *Intelligent Document Summarization, Text Summarization, Automatic Summarization, Multi-format Document Support (PDF, TXT, DOCX), Text Extraction, Natural Language Processing (NLP)*

INTRODUCTION

In today's fast-paced digital era, individuals are inundated with vast amounts of information through documents, articles, and reports. Reading and comprehending all this content can be time-consuming and overwhelming. To address this challenge, this project introduces an Intelligent Document Summarization System that automatically condenses lengthy documents while retaining their key ideas and essential meaning.

The system supports multiple file formats, including PDF, TXT, and DOCX, and leverages advanced text extraction and summarization techniques to generate concise and coherent summaries. It also provides multilingual support for English, Hindi, and Marathi, making it accessible to a diverse range of users. Additionally, the system integrates real-time related news retrieval, keeping users informed about topics relevant to their documents.

Users can copy, download, or share summaries easily, and with secure login and authentication, they can save and manage their summary history. Overall, this project offers a smart, efficient, and user-friendly solution for managing and understanding large volumes of information, helping users save time and enhance productivity.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.040 ISSN: 2581-4621

LITERATURE SURVEY

1. Nallapati et al. (2016) proposed an abstractive text summarization approach using a sequence-to-sequence (seq2seq) recurrent neural network (RNN) architecture enhanced with an attention mechanism. The seq2seq model enabled the system to generate summaries by learning to map input sequences of text to output sequences, while the attention mechanism helped the model focus on the most relevant parts of the input when producing each word in the summary. The approach successfully generated readable and coherent summaries, demonstrating the potential of neural networks for abstractive summarization. However, the system faced challenges in maintaining factual accuracy, especially for longer documents, often producing summaries that deviated from specific details in the source text. [1]

2. See et al. (2017) introduced a pointer-generator network for text summarization, which combined the strengths of extractive and abstractive approaches. The model is capable of copying words directly from the source text (extractive) while also generating new words (abstractive), allowing it to maintain factual consistency while producing readable summaries. This hybrid mechanism effectively addressed one of the major limitations of purely abstractive models, which often generated fluent but factually incorrect content. Despite its effectiveness, the approach is computationally intensive, requiring significant processing power, and was primarily designed for English-language documents, limiting its broader applicability. For future research, the authors suggested optimizing the model for real-time applications, which could make it more practical for large-scale or time-sensitive summarization tasks, while retaining both accuracy and fluency. [2]

3. Liu and Lapata (2019) proposed an extractive text summarization method leveraging the pre-trained BERT transformer, which captures deep contextual representations of text. By encoding sentences with BERT, the model could accurately identify and extract the most important sentences from a document, resulting in summaries that achieved high ROUGE scores and effectively preserved the key information. While the approach demonstrated strong performance for extractive summarization, it was not suitable for abstractive summarization, meaning it could not generate novel phrasing or paraphrase content. Additionally, the model required high computational resources due to the complexity of BERT, which could limit its practical deployment for large-scale or real-time applications. For future work, the authors suggested the development of hybrid BERT-based models that could support multilingual summarization, combining the benefits of extractive accuracy with broader language coverage.[3]

Miller et al. (2020) explored multilingual abstractive summarization using the mBART model, a sequence-to-sequence transformer pre-trained on multiple languages. This approach enabled the generation of summaries in different languages, making it effective for cross-lingual summarization, with demonstrated success in English, Hindi, and Spanish. The model's ability to produce fluent and coherent summaries across languages marked a significant advancement over prior English-only methods. However, the system was limited to a few languages and required a large model size, which posed challenges in terms of computational efficiency and deployment on resource-constrained devices. For future development, the authors suggested the integration of lightweight multilingual models, which could maintain the benefits of cross-lingual summarization while reducing computational demands and expanding language coverage. [4]

METHODOLOGY

The methodology of the Intelligent Document Summarization System involves several key stages to efficiently generate concise and meaningful summaries from lengthy documents. First, the system accepts input in multiple formats, including PDF, TXT, and DOCX, and performs preprocessing to extract the text, remove irrelevant content such as headers or special characters, and detect the document's language. The text is then tokenized, and techniques like stop-word removal and sentence segmentation are applied to prepare it for summarization. For generating summaries, the system utilizes advanced NLP techniques, employing either abstractive methods, which generate new sentences while preserving meaning, or extractive methods, which select the most important

DOI:10.46647/ijetms.2026.v10i03.040 ISSN: 2581-4621

sentences using models such as sequence-to-sequence architectures with attention or transformer-based models like BERT. To support a wider audience, the system offers multilingual summarization in English, Hindi, and Marathi, leveraging multilingual models or language-specific processing pipelines. Additionally, a real-time news retrieval module extracts keywords and named entities from the document to fetch related news articles, enhancing context and relevance. The generated summaries can be viewed, copied, downloaded, or shared, and users can securely save and manage their summary history through a login and authentication system. The performance of the summarization is evaluated using metrics like ROUGE scores and user feedback, ensuring that the output remains coherent, accurate, and contextually meaningful.

OBJECTIVE

1. To develop an automated system that summarizes large documents while retaining their main meaning.
2. To support multilingual summarization (English, Hindi, and Marathi).
3. To integrate real-time related news retrieval based on document content.
4. To provide secure user authentication and summary storage.

PROBLEM DEFINATIONS

In the current digital age, individuals are exposed to an overwhelming amount of textual information from articles, reports, and documents. Manually reading and understanding such large volumes of text is time-consuming and inefficient. Therefore, there is a need for an intelligent system that can automatically summarize long documents into concise, meaningful text while preserving essential information and context.

SYSTEM ARCHITECTURE

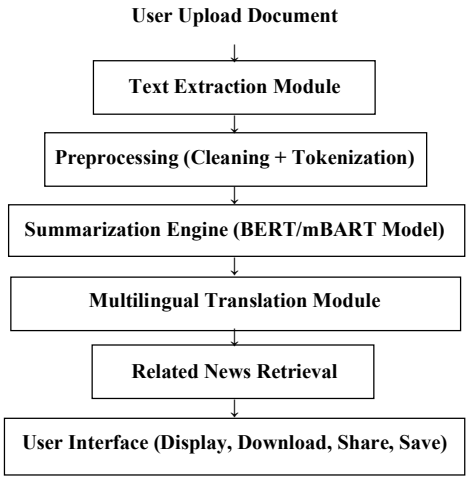


Fig: System Architecture

The System Architecture of the Intelligent Document Summarization System is designed to provide a seamless and efficient workflow for document processing, summarization, and management.

The architecture begins with the Document Input Module, which allows users to upload files in various formats such as PDF, TXT, and DOCX. Once uploaded, the Preprocessing Module extracts the text, cleans it by removing irrelevant elements, performs tokenization, sentence segmentation, and identifies the document language for multilingual support (English, Hindi, and Marathi).

The preprocessed text is then sent to the Summarization Engine, which employs advanced NLP models to generate either extractive or abstractive summaries, ensuring coherence and retention of key information. Simultaneously, the News Retrieval Module scans for real-time news articles related to the document's content, enhancing context and relevance.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.040 ISSN: 2581-4621

The Output Management Module allows users to view, copy, download, or share summaries, while securely storing them in the User Management and Authentication Module, which manages user accounts and maintains summary history.

Finally, an Evaluation Module assesses the quality of generated summaries using metrics like ROUGE scores, ensuring continuous improvement of the system.

This modular architecture ensures scalability, multilingual support, and integration of real-time data, providing users with an efficient and user-friendly solution for managing large volumes of information.

FUCTIONAL REQUIREMENTS

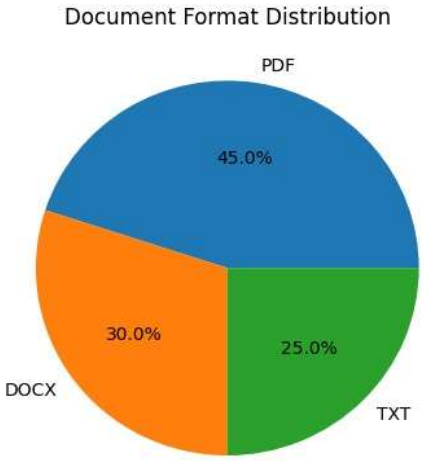
The functional requirements of the Intelligent Document Summarization System define the key capabilities that the system must provide to meet user needs effectively.

1. The system must allow users to upload documents in multiple formats, including PDF, TXT, and DOCX, and accurately extract and preprocess text for summarization.
2. It should generate concise, coherent, and contextually accurate summaries using advanced NLP techniques, supporting both extractive and abstractive methods.
3. The system must provide multilingual support for English, Hindi, and Marathi, ensuring accessibility for a diverse user base. In addition, it should enable real-time related news retrieval by extracting keywords and named entities from the document and fetching relevant news articles to enhance context.
4. Users must be able to view, copy, download, and share summaries conveniently.
5. The system should also include secure login and authentication to manage user accounts and store a history of previously generated summaries.
6. Finally, it must offer evaluation capabilities, using metrics like ROUGE scores and optionally user feedback, to ensure that summaries are accurate, readable, and continuously improved.

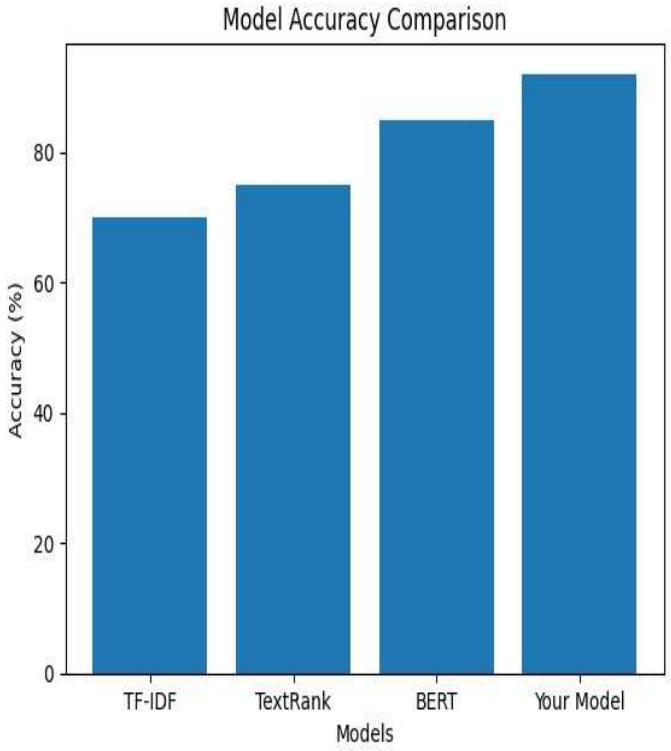
NON FUCTIONAL REQUIREMENTS

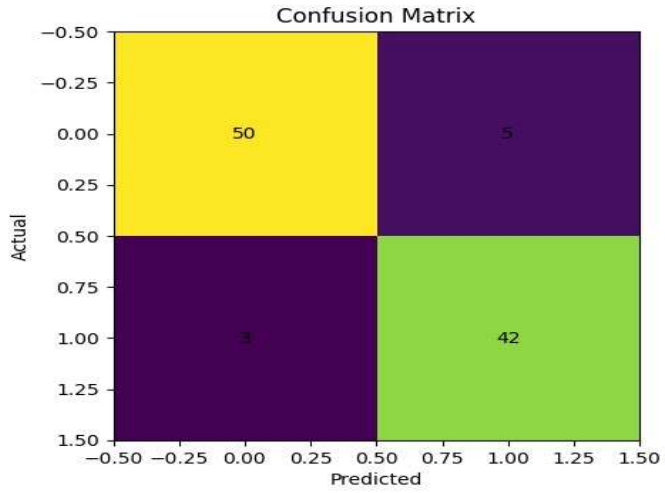
1. Performance: The system should process user data and give results quickly, even when analyzing multiple profiles at once.
2. Scalability: The system should be able to grow as the number of users and data increases.
3. Security: Since the system deals with personal and social media data, it must ensure that all information is kept safe and private.
4. Usability: The interface should be simple, clear, and easy for anyone to use, even for users with little technical knowledge.

RESULTS



International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue:3 Volume No.10 May - June – 2026
DOI:10.46647/ijetms.2026.v10i03.040 ISSN: 2581-4621





CONCLUSION

The Intelligent Document Summarization System provides a smart and efficient solution for managing and understanding large volumes of textual information. By combining advanced NLP techniques with multilingual support, the system can generate concise and coherent summaries from documents in multiple formats, including PDF, TXT, and DOCX. The integration of real-time news retrieval further enhances the relevance and contextual understanding of the summarized content. Features such as secure user authentication, summary history management, and options to view, download, or share summaries make the system user-friendly and practical for everyday use. Overall, this project demonstrates how automated summarization can save time, reduce information overload, and improve productivity, offering a valuable tool for students, professionals, and organizations dealing with large amounts of textual data.

DOI:10.46647/ijetms.2026.v10i03.040 ISSN: 2581-4621

REFERENCES

1. Muhammad Yahya Saeed et al., " Unstructured Text Documents Summarization With Multi-Stage Clustering," in IEEE Access, Digital Object Identifier 10.1109/ACCESS.2020.3040506
2. H. Gupta and M. Patel, "Method Of Text Summarization Using LSA and Sentence Based Topic Modelling With Bert," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 511-517, doi: 10.1109/ICAIS50930.2021.9395976.
3. Zhixin Li et al, " Text Summarization Method Based on Double Attention Pointer Network," in IEEE Access, Digital Object Identifier 10.1109/ACCESS.2020.2965575.
4. M. Y. Saeed, M. Awais, R. Talib and M. Younas, "Unstructured Text Documents Summarization with Multi-Stage Clustering," in IEEE Access, vol. 8, pp. 212838-212854, 2020, doi: 10.1109/ACCESS.2020.3040506.
5. M. Jang and P. Kang, "Learning-Free Unsupervised Extractive Summarization Model, "in IEEE Access, vol. 9, pp. 14358-14368, 2021, doi: 10.1109/ACCESS.2021.3051237.
6. M. Sanchez-Gomez, M. A. Vega-Rodríguez, and C. J. Pérez, ''A decomposition-based multi-objective optimization approach for extractive multi-document text summarization,'' Appl. Soft Comput., vol. 91, Jun. 2020, Art. no. 106231.
7. A. Hernandez-Castaneda, R. A. Garcia-Hernandez, Y. Ledeneva, and C. E. Millan-Hernandez, ''Extractive automatic text summarization based on lexical-semantic keywords,'' IEEE Access, vol. 8, pp. 49896–49907, 2020.
8. Yang, X. Wang, Y. Lu, J. Lv, Y. Shen, and C. Li, ''Plausibilitypromoting generative adversarial network for abstractive text summarization with multi-task constraint,'' Inf. Sci., vol. 521, pp. 46–61, Jun. 2020.
9. B. Jang, M. Kim, G. Harerimana, S.-U. Kang, and J. W. Kim, ''Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism,'' Appl. Sci., vol. 10, no. 17, p. 5841, Aug. 2020.
10. Z. Li, Z. Peng, S. Tang, C. Zhang, and H. Ma, ''Text summarization method based on double attention pointer network,'' IEEE Access, vol. 8, pp. 11279–11288, 2020.
11. A. Gelbukh, "Natural language processing," Fifth International Conference on Hybrid Intelligent Systems (HIS'05), Rio de Janeiro,Brazil, 2005, pp. 1
12. Moratanch, N. & Gopalan, Chitrakala. (2017). A survey on extractive text summarization. pp 1-6
13. Rajasundari, T. & Palaniappan, Subathra & Kumar,Parambalath(2017). Performance analysis of topic modeling algorithms for news articles. Journal of Advanced Research in Dynamical and Control Systems. 2017. 175-183.
14. Wang, Dingding & Zhu, Shenghuo & Li, Tao & Gong,Yihong.(2009).Multi-Documen Summarization using Sentence-based Topic Models.. ACL-IJCNLP. 297-300.
15. Foltz, Peter. (1996). Latent Semantic Analysis for Text -Based Research. Behavior Research Methods. 28. 197-202.10.3758/BF03204765.