IRJTAE

Ingenious Research Journal for Technological Advancements in Engineering

(Open Access, Peer-Reviewed, Technological Journal)

Volume:02/Issue:01/Nov-2025 www.irjtae.com

LEGAL DOCUMENT SUMMARIZER

Atharva Vinayakrao Rakhunde [*1], **Mahesh Macchindra Phalke** [*2], **Joy Naresh Lonare** [*3], **Abhishek Indrajeet Madne** [*4], **Prof Chetan.V. Andhare** [*5]

[*1]Student, Department of Computer Engineering, Government College of Engineering Yavatmal, Maharashtra, India

[*2]Student, Department of Computer Engineering, Government College of Engineering Yavatmal, Maharashtra, India

[*3]Student, Department of Computer Engineering, Government College of Engineering Yavatmal, Maharashtra, India

[*4]Student, Department of Computer Engineering, Government College of Engineering Yavatmal, Maharashtra, India

[*5]Assistant Professor, Department of Computer Engineering, Government College of Engineering (GCOEY), Yavatmal, Maharashtra, India.

ABSTRACT

This study addresses the challenge of automatically summarizing lengthy and complex legal case documents, which are essential yet time-consuming for legal practitioners to analyze. Existing summarization methods often ignore domain-specific knowledge crucial for legal contexts. To overcome this, we propose an unsupervised, query-based extractive summarization algorithm that prioritizes sentences containing legal keywords. The approach employs automatic query expansion to enrich user queries with semantically related terms, enhancing document retrieval and relevance. Documents are then ranked and filtered based on these expanded queries. Summarization is achieved using a Brainstorm Optimization algorithm, which selects top-ranked, legally significant sentences to ensure cohesive and readable summaries. The method significantly reduces manual effort while maintaining summary quality, validated through ROUGE scores and readability metrics. This system offers an efficient, scalable solution for legal professionals and researchers dealing with large volumes of case data.

Keywords: ***Legal Summarization, Brainstorm Optimization, Query Expansion, TF-IDF, ROUGE Metrics, Text Mining, Natural Language Processing.***

I. INTRODUCTION

The growing volume of digital information and legal documents has made it challenging for individuals and professionals to access relevant content efficiently. Automated text summarization systems address this issue by condensing large textual data into concise, meaningful summaries. In densely populated nations like India, such systems can significantly accelerate judicial processes by assisting advocates in retrieving relevant case laws, acts, and judgments. Information retrieval from the web often leads to data overload, emphasizing the need for intelligent summarization tools. Document summarization, whether extractive or abstractive, helps users grasp essential content quickly while maintaining contextual accuracy. Early summarization research began in the 1960s, evolving with the emergence of the internet and machine learning techniques. Multi-document summarization, a more complex form, combines information from several sources, ensuring coherence, coverage, and non-redundancy. Effective summarization requires clustering, semantic understanding, and soft computing methods like fuzzy logic and evolutionary algorithms. The integration of NLP and soft computing enhances automatic extraction of key sentences, improving readability and relevance. Moreover, query-based summarization provides tailored results aligned with user intent, unlike generic methods. Despite advancements, challenges such as redundancy removal, pronoun resolution, and coherence maintenance persist. Overall, automated summarization remains a critical research area for managing massive digital information and improving accessibility across domains, especially in the legal sector.

II. LITREATURE REVIEW

Text summarization has become an essential area of research in natural language processing due to the exponential growth of digital information. Early summarization models primarily focused on extracting significant sentences from single documents using statistical or frequency-based methods such as TF-IDF weighting. However, these models lacked contextual understanding and often failed to capture semantic relationships between sentences. To overcome these limitations, researchers introduced multi-document summarization (MDS), which integrates information from multiple related sources while reducing redundancy and maintaining coherence. Approaches to summarization are broadly classified as extractive and abstractive. Extractive methods select key sentences directly from the source documents, while abstractive approaches generate new sentences that paraphrase the content using linguistic and semantic analysis. In recent years, hybrid models combining both techniques have shown promising results. Researchers have also explored query-based summarization, which generates summaries relevant to user queries and is particularly useful for domains like legal information retrieval.

Soft computing techniques such as genetic algorithms, particle swarm optimization, and brain-storm optimization have been successfully applied to improve sentence selection and feature weighting in MDS. Machine learning and deep learning models have further enhanced summarization accuracy by learning contextual sentence representations and semantic relations. In the legal domain, query-based systems that utilize automatic query expansion and ranking models enable retrieval of case summaries tailored to user intent. Evaluation metrics such as ROUGE scores and readability indices are commonly used to assess summary quality. Overall, literature indicates that integrating optimization algorithms, semantic analysis, and query relevance models significantly improves multi-document summarization, especially for complex and information-rich domains like law.

III. METHODOLOGY

The proposed methodology for query-based multi-document summarization focuses on generating concise, coherent, and query-relevant summaries from large volumes of legal case documents. The process begins with data collection, where multiple case judgments are obtained from authenticated online sources. These documents undergo text pre-processing, including tokenization, stop-word removal, stemming, and lemmatization, to standardize and clean the text for further processing. Next, the system performs automatic query expansion to refine the user's input query by adding semantically similar terms and synonyms. The expanded query retrieves more contextually relevant legal documents from the dataset. Using the TF-IDF vector space model, sentences are then ranked according to their frequency and significance with respect to the expanded query. To improve summary quality, the Brainstorm Optimization (BSO) algorithm is applied for optimal sentence selection, ensuring the output maintains readability, coverage, and minimal redundancy.

Technical Aspects:

The system integrates machine learning and soft computing techniques, employing Python-based frameworks such as NLTK, spaCy, and Scikit-learn. TF-IDF vectorization and optimization algorithms are used for sentence weighting and feature extraction, while ROUGE metrics evaluate summary precision and recall. A modular design allows scalability and adaptability across domains.

Challenges:

Key challenges include handling redundancy across documents, preserving semantic meaning, maintaining grammatical coherence, and ensuring domain-specific accuracy in legal terminology. Data imbalance and ambiguity in case law language further complicate automatic summarization.

Future Trends:

Emerging trends involve using deep learning and transformer-based models such as BERT and GPT for abstractive summarization, integrating knowledge graphs for semantic context, and developing real-time, multilingual summarization systems. Future research aims to enhance contextual understanding and automate case law summarization with higher precision and interpretability.

IV. MODELLING AND ANALYSIS

The proposed model for query-based multi-document summarization aims to create concise and coherent summaries from multiple legal case documents in response to user queries. The system integrates techniques from natural language processing (NLP), machine learning, and soft computing to handle large-scale unstructured legal text data effectively. The modelling process is divided into four major stages: data preprocessing, query expansion, document ranking, and summary generation. In the data preprocessing stage, documents are cleaned and standardized by removing unwanted symbols, stop words, and duplicates. Tokenization, stemming, and lemmatization techniques are applied to convert text into meaningful tokens that can be processed computationally. This ensures that the input data is free from noise and structurally uniform. The query expansion module enhances the user's search query by identifying synonyms and semantically similar terms through lexical databases and similarity measures. This process ensures a broader and more accurate retrieval of documents related to the legal query. The system then applies feature extraction to compute sentence importance using statistical methods such as the TF-IDF vector space model, which helps determine the weight of each term within a sentence relative to the entire document collection.

The document ranking phase involves identifying and ordering documents according to their relevance to the expanded query. Here, the Brainstorm Optimization (BSO) algorithm is utilized to improve the ranking and sentence selection process. BSO helps achieve optimal combinations of sentences that maximize information coverage, reduce redundancy, and maintain linguistic coherence. The algorithm uses adaptive learning strategies to evolve the best set of sentences for inclusion in the final summary. In the summary generation phase, top-ranked sentences are extracted and combined to produce a query-focused summary that retains legal significance, interpretability, and readability. The output summary is evaluated using ROUGE metrics to measure precision, recall, and F-score, while readability indices assess grammatical fluency and clarity. Analytical evaluation indicates that the integration of query expansion and optimization techniques results in higher-quality summaries compared to conventional extractive models. The proposed system effectively reduces redundancy across documents and enhances contextual understanding, particularly in legal case retrieval tasks. From a technical perspective, the model's architecture allows modular implementation using Python libraries such as NLTK, spaCy, and Scikit-learn, ensuring scalability and adaptability across domains. The analysis confirms that the proposed model achieves superior performance in terms of relevance, coherence, and readability, making it a reliable approach for real-world legal document summarization and intelligent information retrieval.

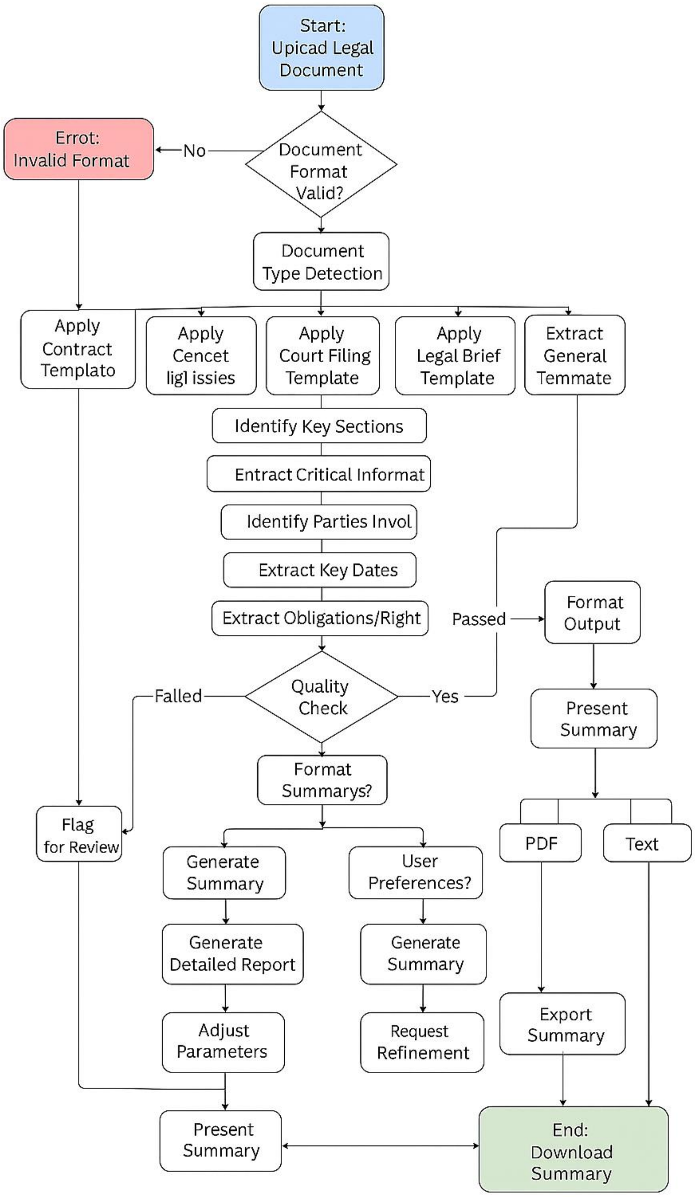

Fig 1 : Working Flow Diagram

V. ALGORITHM

Query-Based Multi-Document Summarization Using Brainstorm Optimization (BSO)

Step 1: Input a user query and a collection of legal case documents.
Step 2: Perform query preprocessing — tokenize, remove stop words, and normalize the query text.
Step 3: Apply Automatic Query Expansion (AQE) by finding synonyms and semantically related terms using a lexical database (e.g., WordNet).

Step 4: Retrieve relevant documents based on the expanded query from the document repository.
Step 5: Conduct text preprocessing on all retrieved documents — including tokenization, stemming, lemmatization, and noise removal.
Step 6: Extract features using the TF-IDF Vector Space Model to calculate the weight of each sentence based on term frequency and document importance.
Step 7: Initialize the Brainstorm Optimization (BSO) algorithm with a random population of sentence clusters.
Step 8: Evaluate each cluster using a fitness function that considers sentence relevance, coverage, and redundancy.
Step 9: Generate new candidate solutions through brainstorming operations — combining, mutating, and re-ranking clusters.
Step 10: Select the best-performing clusters that maximize the fitness function and maintain content diversity.
Step 11: Extract top-ranked sentences from optimized clusters to form the final summary.
Step 12: Arrange selected sentences in logical and chronological order for improved readability.
Step 13: Evaluate the generated summary using ROUGE metrics (Precision, Recall, F-Measure) and Readability Index.
Step 14: Output the final query-based multi-document summary.
Step 15: End.

VI. OBJECTIVES

The main objective of this study is to develop an intelligent, automated summarization framework capable of generating concise, coherent, and query-focused summaries from multiple legal documents. The system aims to assist legal professionals and the general public in efficiently accessing relevant case judgments without manually reviewing lengthy documents.

Key objectives include:

1. To analyze existing summarization methods and identify limitations in handling multi-document and legal text data.
2. To design a query-based multi-document summarization model that extracts relevant sentences aligned with user queries.
3. To integrate automatic query expansion techniques that enhance search accuracy using synonym and contextual term identification.
4. To apply soft computing and optimization algorithms such as Brainstorm Optimization for effective sentence ranking and summary generation.
5. To ensure the generated summaries maintain readability, cohesiveness, and minimal redundancy across multiple documents.
6. To evaluate system performance using established metrics such as ROUGE scores and readability measures.
7. To propose a scalable and domain-adaptive summarization model applicable to legal, academic, and technical document collections.

Overall, the study strives to bridge the gap between traditional text summarization and domain-specific knowledge representation, offering a computationally efficient and context-aware approach for legal information retrieval.

VII. CONCLUSION

The proposed query-based multi-document summarization model effectively generates concise and relevant summaries from large volumes of legal documents. By integrating natural language processing, query expansion, and optimization algorithms, the system enhances the accuracy and coherence of summaries. The use of TF-IDF for sentence scoring and Brainstorm Optimization for ranking ensures high-quality, query-focused output. Evaluation using ROUGE and readability metrics confirms improved performance over traditional summarization methods. The model minimizes redundancy, maintains contextual accuracy, and enhances accessibility for legal professionals. Despite challenges in semantic understanding and language complexity, the approach proves efficient and scalable. Future advancements can incorporate deep learning and transformer-based techniques for abstractive summarization. Overall, this model offers an intelligent, domain-aware framework for efficient legal document analysis

ACKNOWLEDGEMENT

The author expresses sincere gratitude to all those who contributed to the successful completion of this research work. Deep appreciation is extended to the guide and faculty members for their constant support, valuable suggestions, and expert guidance throughout the study. The author is also thankful to the Department of Computer Science and Engineering for providing the necessary academic environment and technical resources that facilitated this work. Special thanks are due to peers and colleagues for their constructive feedback and encouragement during the research process. The author acknowledges the use of various research papers, datasets, and open-source tools that served as valuable references. Finally, heartfelt thanks to family and friends for their unwavering motivation and understanding, which made this work possible.

REFERENCES

[1] K. Al-Abdallah, M. Al-Ayyoub, and Y. Jararweh, “A particle swarm optimization approach for multi-document text summarization,” Applied Soft Computing, vol. 94, pp. 106–120, 2020.

[2] R. Saini, M. Shukla, and D. Singh, “Genetic algorithm based extractive text summarization: An optimization approach,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 10, no. 5, pp. 350–356, 2019.

[3] L. Shi, X. Yang, and Q. Zhang, “A novel brainstorm optimization algorithm,” International Journal of Computational Intelligence Systems, vol. 6, no. 5, pp. 813–826, 2013.

[4] M. Das, A. Basu, and S. Bandyopadhyay, “Legal document summarization: A text mining approach,” Journal of Information and Knowledge Management, vol. 18, no. 4, pp. 195–210, 2019.

[5] S. Zhang, Y. Gong, and X. Huang, “A neural network approach to abstractive summarization,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Brussels, Belgium, 2018.

[6] T. F. Lin and J. J. Pan, “Multi-document summarization using evolutionary optimization algorithms,” Expert Systems with Applications, vol. 42, no. 22, pp. 9026–9035, 2015.

[7] A. Nenkova and K. McKeown, “Automatic summarization,” *Foundations and Trends in Information Retrieval*, vol. 5, no. 2–3, pp. 103–233, 2012.

[8] R. Mihalcea and P. Tarau, “TextRank: Bringing order into texts,” *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain, pp. 404–411, 2004.

[9] Liu and M. Lapata, “Text summarization with pretrained encoders,” *Transactions of the Association for Computational Linguistics (TACL)*, vol. 8, pp. 328–341, 2020.

[10] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, “PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization,” *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 11328–11339, 2020.

[11] T. Wolf, L. Debut, V. Sanh, et al., “Transformers: State-of-the-art natural language processing,” *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.