Comparative Analysis Of Various Extractive Summarization Techniques For Legal Domain

Ralph Pereira
Computer Engineering
St. Francis Institute of Technology
pereiraralph2102@gmail.com

Samit Fernandes
Computer Engineering
St. Francis Institute of Technology
samitfernandes019@gmail.com

Jaden Franco
Computer Engineering
St. Francis Institute of Technology
francojaden04@gmail.com

K. Priya
Computer Engineering
St. Francis Institute of Technology
kpriya@sfit.ac.in

***Abstract*—This study conducts a comparative analysis of various summarization methods applied in the legal domain, focusing on their effectiveness in generating coherent and informative summaries. The results indicate that InLegalBert extractive summarization significantly outperforms other techniques, providing summaries that accurately retain crucial legal terms and sections [1]. In comparison, TF-IDF (Term Frequency-Inverse Document Frequency) effectively extracts key terms, while LDA (Latent Dirichlet Allocation) identifies thematic content but lacks coherence. TextRank, despite its speed, offers less accuracy in capturing the depth of legal information [2]. The findings suggest potential future research avenues, including the hybridization of these methods to leverage their strengths and refining the InLegalBert model for enhanced domain specificity. Furthermore, integrating abstractive summarization techniques may improve the sentence structure and overall coherence of legal summaries, advancing the effectiveness of legal documentation analysis.**

***Index Terms*—Extractive summarization, InLegalBERT, TF-IDF(Term Frequency-Inverse Document Frequency), LDA(Latent Dirichlet Allocation), TextRank, legal domain, IPC(Indian Penal Code) sections.**

I. Introduction

In recent years, the exponential growth of legal documents has created an urgent need for effective summarization techniques that can aid legal professionals in quickly extracting relevant information. Traditional methods of reviewing legal texts are time-consuming and often inefficient, prompting the exploration of automated summarization techniques. Extractive summarization, which identifies and compiles significant sentences from the original text, has gained prominence due to its ability to retain critical information while reducing reading time.

Among the various summarization approaches, BERT (Bidirectional Encoder Representations from Transformers) has emerged as a powerful tool for natural language processing tasks, particularly in the legal domain. BERT's capacity to understand context and semantics makes it highly suitable for generating coherent summaries that encompass essential legal terms and concepts. In addition to BERT, other methods such as TF-IDF (Term Frequency-Inverse Document Frequency), LDA (Latent Dirichlet Allocation), and TextRank provide alternative strategies for summarizing legal texts, each with its own strengths and limitations.

This study conducts a comparative analysis of these summarization methods, focusing on their effectiveness in generating summaries of legal documents. By evaluating metrics such as coherence, relevance, and readability, the analysis aims to determine the best-performing method for legal summarization. The findings will contribute to the development of more refined summarization tools, ultimately enhancing the efficiency of legal research and documentation processes.

II. Related Work

The field of text summarization has witnessed significant advancements, particularly with the emergence of pre-trained language models tailored for specific domains, including the legal sector. Paul et al. [1] highlight the effectiveness of pre-trained language models in enhancing legal document summarization, showcasing their ability to understand legal terminologies and context. This work underscores the potential of leveraging such models to improve the coherence and relevance of summaries in legal applications.

In comparative analyses of extractive summarization techniques, Rani and Bidhan [2] examined traditional methods such as TextRank, TF-IDF, and LDA. Their findings revealed the strengths and weaknesses of each method, emphasizing that while TF-IDF is efficient for term extraction, TextRank provides a more nuanced approach to capturing contextual relationships within text. This study serves as a foundation for understanding the performance of different summarization algorithms and their applicability to various domains, including legal texts.

Further comparative studies, such as those by Palliyali et al. [3] and Issam et al. [4], delve into the efficacy of various extractive techniques, reinforcing the significance of topic modeling and semantic analysis in summarization. Palliyali et al. [3] present a comprehensive evaluation of extractive summarization techniques, affirming the need for continuous improvement and adaptation of these methods to better serve specialized fields. Meanwhile, Issam et al. [4] explore topic modeling as a means to enhance extractive summarization, providing insights into how these approaches can be tailored for specific content domains.

Recent advancements in latent semantic analysis also contribute to the summarization landscape, as explored by Onah et al. [5], who apply LDA topic modeling to automatic text summarization. This approach enhances the understanding of contextual relationships within the text, further refining the summarization process. Additionally, Ramadhan et al. [6] investigate the implementation of the TextRank algorithm in summarizing product reviews, which demonstrates the versatility of this method across different types of content.

Jewani et al. [7] provide a brief overview of various extractive summarization methods, offering insights into emerging trends and methodologies that enhance summarization capabilities. Gupta [8] and Jain [9] discuss the application of TextRank and TF-IDF, respectively, in their practical implementations, contributing to a broader understanding of how these algorithms can be employed for effective text summarization. Collectively, these studies illustrate the evolving landscape of text summarization techniques, highlighting the necessity of combining traditional methods with advanced models to address the specific challenges of summarizing legal documents effectively.

III. METHODOLOGY

This section outlines the methodology for evaluating various extractive summarization techniques within the legal domain. The process begins with the collection and preprocessing of legal documents, followed by the application of four extractive summarization methods—TF-IDF, TextRank, LDA, and a BERT-based model. Each method generates summaries, which are analyzed for their effectiveness in retaining key legal terms and case-specific details.

A. Data Preprocessing

The methodology for this study involves implementing various extractive summarization techniques to evaluate their effectiveness in the legal domain. It begins with the collection of a diverse set of legal documents, followed by preprocessing steps such as text normalization, tokenization, and removal of irrelevant content. Four extractive summarization methods—TF-IDF, TextRank, LDA, and a BERT-based model—are then applied to generate summaries.

B. BERT-based Extractive Summarization

BERT embeddings are used to compute sentence similarities, forming the basis of extractive summarization. BERT-based models, particularly BERT embeddings, are used due to their ability to capture deep contextual information. In this method, each sentence in the document is represented as a dense vector using BERT embeddings. We compute sentence similarities and rank them based on their importance (see Fig. 1).

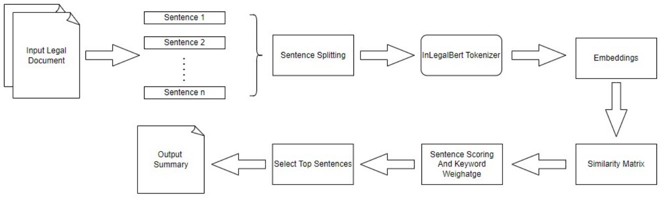

Fig. 1. BERT-Based Extractive Summarization Architecture

- Input Text: Legal document is input.
- Sentence Splitting: Document is split into sentences.
- Tokenization: Sentences are tokenized using InLegalBERT tokenizer.
- Generate Embeddings: BERT model generates embeddings for each sentence.
- Similarity Matrix Calculation: Inner product of embeddings to form similarity matrix.
- Sentence Scoring: Sentences are scored based on similarity matrix and length.
- Keyword Weighting: Legal terms in each sentence are weighted.
- Sentence Selection: Top sentences are selected based on combined score.
- Output Summary: A fixed-length summary is generated.

C. TF-IDF Summarization

The TF-IDF (Term Frequency-Inverse Document Frequency) method emphasizes frequently occurring terms while downplaying commonly used words. For summarization, we compute the TF-IDF scores of all terms in the document and rank sentences based on the sum of the TF-IDF scores of the words they contain (see Fig. 2). This method is effective for extracting the most information-rich sentences in the document. However, since TF-IDF doesn't consider sentence structure, there may be occasional coherence issues.

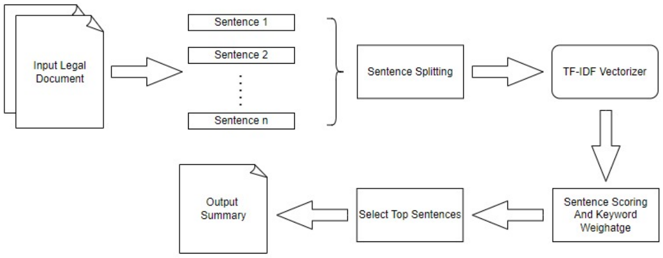

Fig. 2. TF-IDF Summarization Architecture

- Input Text: Legal document is input.
- Sentence Splitting: Document is split into sentences.
- TF-IDF Vectorization: Sentences are vectorized using TF-IDF.
- Sentence Scoring: Sentences are scored based on TF-IDF vector weights.
- Sentence Selection: Top sentences are selected based on scores.

- Output Summary: A fixed-length summary is generated.

D. Latent Dirichlet Allocation (LDA) Summarization

LDA is a topic modeling technique that helps identify latent topics within a document. After applying LDA to the document, sentences are categorized under specific topics. For summarization, a representative set of sentences from each topic is selected (see Fig. 3). This ensures that the summary provides a balanced representation of all major themes discussed in the legal document, such as case background, legal issues, and arguments from both sides.

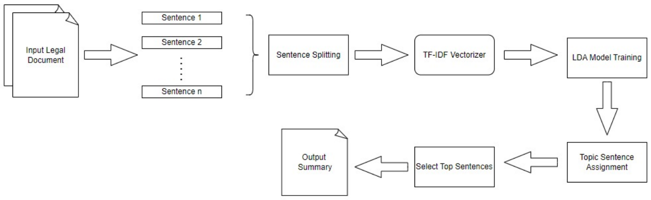

Fig. 3. LDA Summarization Architecture

- Input Text: Legal document is input.
- Sentence Splitting: Document is split into sentences.
- TF-IDF Vectorization: Sentences are vectorized using TF-IDF.
- LDA Model Training: Latent Dirichlet Allocation is applied to group sentences into topics.
- Topic Sentence Assignment: Sentences are assigned to topics based on LDA.
- Sentence Selection: Top sentences from each topic are selected.
- Output Summary: A fixed-length summary is generated.

E. TextRank Summarization

TextRank is a graph-based ranking algorithm where sentences are nodes, and edges represent sentence similarity based on lexical overlap. TextRank constructs a graph by calculating the cosine similarity between sentence embeddings, and then ranks sentences based on their centrality in the graph. The most central sentences (i.e., those that are most similar to others) are selected for the final summary (see Fig. 4). TextRank is computationally efficient and works well for general summarization tasks, though it may overlook legal-specific nuances.

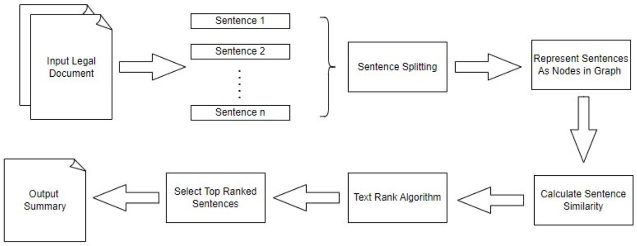

Fig. 4. TextRank Summarization Architecture

- Input Text: Legal document is input.
- Sentence Tokenization: Document is tokenized into sentences.
- Graph Construction: Sentences are represented as nodes in a graph.
- Sentence Similarity Calculation: Sentence similarity is computed based on overlap.
- PageRank Algorithm: TextRank algorithm is applied to rank sentences.
- Sentence Selection: Top-ranked sentences are selected.
- Output Summary: A fixed-length summary is generated.

F. Extraction of IPC Sections and Case Details

A specialized step for legal summarization involves extracting IPC sections and key case details. This is done using regex patterns to identify legal terms like section numbers, case titles, and party names. This information is crucial to the final summary as it provides explicit legal references that are often required by legal professionals.

IV. Evaluation

A. Summary Output Comparison

After generating summaries using BERT, TF-IDF, LDA, and TextRank, we compare them based on three criteria:

- Summary length: Ensuring that each summary is concise and within a fixed length.

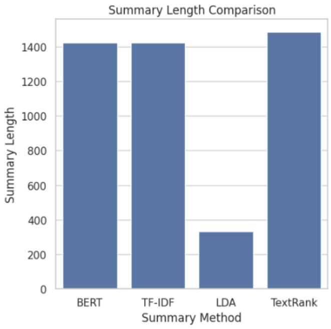

Fig. 5. Bar graph comparing Summary length of each approach

- Relevance: Measured by how well the summaries capture the key aspects of the original document, especially legal details like IPC sections and case outcomes.

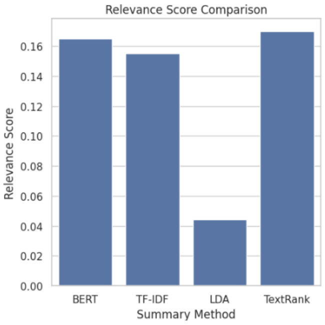

Fig. 6. Bar graph comparing Relevance of each approach

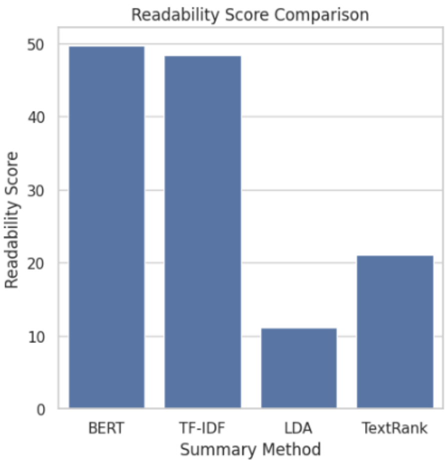

Fig. 7. Bar graph comparing Readability of each approach

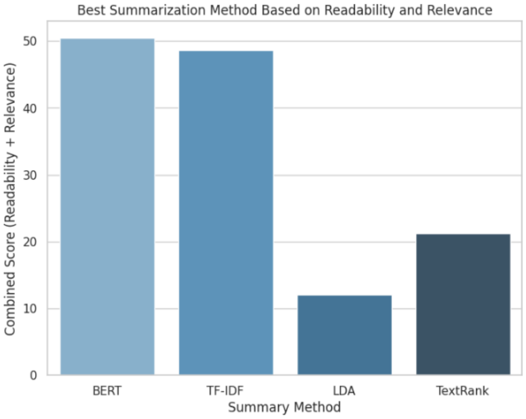

Fig. 8. Bar graph based on Readability and Relavance of each approach

B. Readability and Relevance Metrics

To quantify readability, we use the textstat library, which provides metrics like Flesch reading ease and Gunning Fog index. A lower reading difficulty score indicates better ease of understanding, which is essential for legal documents that are often complex (see Fig. 7). Relevance is assessed by checking the overlap of legal keywords between the original document and the summary. The more accurate the keyword preservation, the better the summarization method in retaining critical legal information (see Fig. 6).

C. Visualization of Results

The performance of each technique is visualized in terms of summary length, readability score, relevance score, and the number of legal references. These visualizations highlight the strengths and weaknesses of each approach, making it easier to identify the best-suited technique for legal document summarization (see Fig. 8).

V. Results and Discussion

The BERT-based extractive summarization technique outperformed the other methods in terms of sentence coherence and legal term retention due to its deep contextual embeddings. TF-IDF was effective in identifying key sentences but often struggled with sentence-level coherence. LDA provided a more thematic summary but sometimes missed critical legal details. TextRank, though fast and efficient, occasionally selected non-essential sentences, leading to less relevant summaries.

A. InLegalBert-based Summarization

The InLegalBert-based summarization method produced the best results in terms of sentence coherence and legal term retention. InLegalBert embeddings, which capture the contextual relationships between words, resulted in summaries that were not only relevant but also structured logically. This method was particularly effective in retaining IPC sections and other legal references due to the contextual weighting of legal terms during summarization (see Table [I]). However, InLegalBert's computational cost is higher compared to traditional methods.

B. TF-IDF

TF-IDF performed well in extracting key sentences based on term importance. The selected sentences often contained critical information but lacked coherence as the sentence selection process ignored the relationship between sentences. This made the summary somewhat disjointed, although it was effective for extracting important keywords and legal references (see Table [I]).

C. LDA

LDA provided a more thematic summary by grouping sentences around specific topics. This method excelled at capturing the broader themes within legal documents, offering insight into overarching topics. However, LDA failed to retain crucial legal details like IPC sections, as it prioritized thematic elements over specific sentences (see Table [I]). This occasionally resulted in summaries that missed essential case-specific information, which is critical in the legal domain.

TABLE I
Summary Analysis Results for Different Methods

Method	***Num Sentences***	***Summary Length***	***Relevance Score***	***Legal References***	***Readability Score***
BERT	3	1461	0.1649	4	50.3
TF-IDF	3	1422	0.1552	4	48.5
LDA	3	478	0.0650	1	12.0
TextRank	3	1483	0.1697	4	21.1

D. TextRank

TextRank was efficient in terms of speed and simplicity, making it a practical choice for quick summarization. It produced coherent summaries, which was a notable strength. However, the method occasionally prioritized non-critical sentences, leading to summaries that lacked some relevance. While legal references were adequately included, TextRank exhibited lower readability compared to TF-IDF and BERT-based models(see Table [I]). This reduced readability, coupled with its reliance on sentence structure rather than deeper content, limited its effectiveness for capturing the nuances of legal documents.

Table [I] Presents the results of a comparative analysis of various summarization methods used for legal texts. Each method is evaluated based on key metrics: the number of sentences in the summary, summary length, relevance score, number of legal references, and readability score.

BERT achieved the highest relevance score of 0.1649, with a summary length of 1461 words and a readability score of 50.3. It also referenced 4 legal sources, indicating its effectiveness in generating coherent and contextually rich summaries.

TF-IDF followed closely with a relevance score of 0.1552, a summary length of 1422 words, and a readability score of 48.5. Like BERT, it also included 4 legal references, demonstrating its capability in identifying significant terms within the text.

TextRank, with a relevance score of 0.1697, produced a longer summary of 1483 words but had a lower readability score of 21.1, suggesting that while it captures relevant content, it may lack coherence in presentation.

LDA performed the least effectively, achieving a relevance score of 0.0650 and a shorter summary length of 478 words, with a readability score of 12.0. Its lower legal reference count of 1 indicates its limited applicability in generating meaningful summaries in the legal context.

TABLE II
COMBINED SCORES FOR SUMMARIZATION METHODS

Method	Combined Score
InLegalBert	**50.46**
TF-IDF	48.66
LDA	12.06
TextRank	21.27

Table [II] summarizes the combined scores of various summarization methods evaluated for their effectiveness in generating legal text summaries. The combined score is a crucial metric that reflects the overall performance of each method based on both readability and relevance.

BERT achieved the highest combined score of 50.46, demonstrating its superiority in generating coherent and contextually relevant summaries for legal documents. This score highlights BERT's ability to integrate complex language patterns and maintain the necessary context.

TF-IDF follows closely with a combined score of 48.66, indicating its effectiveness in identifying key terms and phrases. While slightly lower than BERT, it still demonstrates a strong performance in generating meaningful summaries.

TextRank obtained a combined score of 21.27. Although it shows some capability in summarization, its lower score suggests that it may not effectively maintain readability and relevance compared to BERT and TF-IDF.

LDA received the lowest combined score of 12.06. This result indicates its limited efficacy in summarizing legal texts, likely due to its inability to capture the contextual nuances required for effective summarization in this domain.

Overall, the combined scores clearly indicate that InLegalBERT is the most suitable method for summarizing legal texts, significantly outperforming the other methods in this evaluation.

CONCLUSION

The comparative analysis of various summarization methods reveals that InLegalBert-based extractive summarization outperforms other techniques in the legal domain. InLegalBert excels in generating coherent summaries that accurately capture critical legal terms and sections, ensuring that the nuances of legal language are preserved. Its contextual understanding makes it particularly suitable for summarizing complex legal documents, where precision is essential for maintaining the integrity of the information.

In contrast, TF-IDF is effective for extracting key terms but may not fully capture the relationships between them, potentially impacting summary quality. LDA focuses on thematic summarization by uncovering hidden topics but can lack coherence and contextual accuracy in summarizing individual cases. TextRank, while fast, falls short in depth and accuracy compared to InLegalBert and may produce summaries that lack detail.

Future work could involve hybridizing these techniques to combine their strengths, thereby creating a more comprehensive summarization framework. Additionally, refining the InLegalBert model for domain-specific applications could enhance its performance, ensuring a better grasp of legal intricacies. Exploring the integration of abstractive summarization methods with extractive techniques may further improve the structure and readability of legal summaries, paving the way for more effective legal summarization tools.

REFERENCES

[1] S. Paul, A. Mandal, P. Goyal, and S. Ghosh, “Pre-trained Language Models for the Legal Domain,” Jun. 2023, doi: https://doi.org/10.1145/3594536.3595165.

[2] U. Rani and K. Bidhan, “Comparative Assessment of Extractive Summarization: TextRank, TF-IDF and LDA,” Journal of scientific research, vol. 65, no. 01, pp. 304–311, 2021, doi: https://doi.org/10.37398/jsr.2021.650140.

[3] A. W. Palliyali, M. A. Al-Khalifa, S. Farooq, J. Abinahed, A. Al-Ansari and A. Jaoua, ”Comparative Study of Extractive Text Summarization Techniques,” 2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA), Tangier, Morocco, 2021, pp. 1-6, doi: 10.1109/AICCSA53542.2021.9686867.

[4] K. A. R. Issam, S. Patel, Subalalitha C. N., “Topic Modeling Based Extractive Text Summarization,” International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 4, pp. 1710–1719, Apr. 2020, doi: https://doi.org/10.35940/ijitee.f4611.049620.

[5] D. F. O. Onah, E. L. L. Pang and M. El-Haj, ”A Data-driven Latent Semantic Analysis for Automatic Text Summarization using LDA Topic Modelling,” 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 2771-2780, doi: 10.1109/BigData55660.2022.10020259.

[6] M. R. Ramadhan, S. N. Endah and A. B. J. Mantau, ”Implementation of Textrank Algorithm in Product Review Summarization,” 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia, 2020, pp. 1-5, doi: 10.1109/ICICoS51170.2020.9299005.

[7] K. Jewani, O. Damankar, N. Janyani, D. Mhatre and S. Gangwani, ”A Brief Study on Approaches for Extractive Summarization,” 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Coimbatore, India, 2021, pp. 601-608, doi: 10.1109/ICAIS50930.2021.9396031.

[8] M. Gupta, “Text summarization using TextRank in NLP,” Data Science in your pocket, Jul. 18, 2022. https://medium.com/data-science-in-your-pocket/text-summarization-using-textrank-in-nlp-4bce52c5b390.

[9] A. Jain, “Automatic Extractive Text Summarization using TFIDF,” Medium, Apr. 07, 2019. https://medium.com/voice-tech-podcast/automatic-extractive-text-summarization-using-tfidf-3fc9a7b26f5.

[10] R. C. Kore, P. Ray, P. Lade, and A. Nerurkar, “Legal Document Summarization Using Nlp and Ml Techniques,” International Journal of Engineering and Computer Science, vol. 9, no. 05, pp. 25039–25046, May 2020, doi: https://doi.org/10.18535/ijecs/v9i05.4488.

[11] K. Jewani, O. Damankar, N. Janyani, D. Mhatre, and S. Gangwani, “A Brief Study on Approaches for Extractive Summarization,” IEEE Xplore, Mar. 01, 2021. https://ieeexplore.ieee.org/abstract/document/9396031.