ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

Briefly: AI-Powered Legal Document Summarization System

Harsh Vardhan Singh Raj[1], Prateek Singh Bedi[2], Prathmesh Rakesh Mali[3], Rahmat Pathan[4], Sheetal Kapse[5]

[1,2,3,4]*Student, STES's Smt. Kashibai Navale College of Engineering, Pune, India*

[5]*Assistant Professor, STES's Smt. Kashibai Navale College of Engineering, Pune, India*

Abstract: *Legal professionals routinely analyze lengthy and complex documents such as court judgments, contracts, petitions, and legal notices. Manual review of these documents is both time-consuming and susceptible to human error, making it difficult to efficiently identify critical legal information. Although recent advances in Natural Language Processing (NLP) have enabled automated text summarization, many existing solutions focus only on summary generation and provide limited support for comprehensive legal document analysis. This paper presents Briefly, an AI-powered legal document summarization and analysis system that combines transformer-based abstractive summarization with multiple NLP techniques to improve the accessibility of legal information. The proposed system is developed using a React-based frontend and a FastAPI backend, providing a scalable web application for processing legal documents in PDF, DOCX, TXT, and Markdown formats. The summarization module employs the pre-trained DistilBART transformer model from Hugging Face to generate concise summaries while maintaining computational efficiency. To overcome transformer input-length limitations, the system adopts a hierarchical two-pass summarization strategy consisting of sentence chunking, chunk importance scoring, batch summarization, and final summary compression. Beyond document summarization, the system integrates Named Entity Recognition (NER), rule-based Rhetorical Role Labeling (RRL), case brief generation, legal insight extraction, timeline generation, analytics, document-based question answering, and multi-format export functionality into a unified platform. The implementation further incorporates caching mechanisms and automatic CPU/GPU fallback to improve system responsiveness and reliability during document processing. The developed prototype successfully performs end-to-end legal document analysis through an intuitive web interface, demonstrating the practical integration of modern NLP techniques for legal information retrieval and summarization. The proposed system provides an efficient and scalable solution that can assist legal professionals, researchers, and students in understanding lengthy legal documents while reducing manual effort.*

Keywords: Legal Document Summarization, Natural Language Processing, DistilBART, Transformer Models, FastAPI, React.js, Named Entity Recognition, Rhetorical Role Labeling, Legal AI.

I. NTRODUCTION

The legal domain generates a large volume of documents such as court judgments, contracts, petitions, and legal notices. These documents are often lengthy, contain complex legal terminology, and require significant time for manual review. Legal professionals and researchers must carefully examine these documents to identify important facts, legal arguments, statutory references, and judicial decisions. Manual analysis is time-consuming and increases the possibility of overlooking important information.

 DOI: 10.5281/zenodo.21317970

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

Recent developments in Artificial Intelligence (AI) and Natural Language Processing (NLP) have enabled automated analysis of textual documents. Transformer-based models such as BART, DistilBART, and T5 have demonstrated promising performance in abstractive text summarization. In addition, techniques such as Named Entity Recognition (NER) and Rhetorical Role Labeling (RRL) assist in extracting structured legal information from unstructured legal documents. However, most existing systems primarily focus on summarization and provide limited support for comprehensive legal document analysis.

To address these challenges, this paper presents Briefly, an AI-powered legal document summarization and analysis system. The proposed system accepts legal documents in PDF, DOCX, TXT, and Markdown formats and performs text extraction, preprocessing, hierarchical summarization using DistilBART, Named Entity Recognition, rule-based Rhetorical Role Labeling, case brief generation, timeline extraction, legal insight generation, analytics, and document-based question answering. The system is implemented using a React frontend and a FastAPI backend to provide an interactive web-based platform for legal document analysis.

The primary objective of the proposed system is to reduce the manual effort involved in reviewing lengthy legal documents while improving access to important legal information through modern NLP techniques.

1.1 Problem Statement

Existing legal document summarization systems mainly generate summaries but often lack additional analytical features such as legal insight extraction, timeline generation, case brief preparation, and interactive document querying. Furthermore, processing lengthy legal documents remains challenging because of transformer input limitations and computational overhead. Therefore, there is a need for an integrated system that performs comprehensive legal document analysis in an efficient and user-friendly manner.

1.2 Objectives

The objectives of the proposed work are:

1. To develop an AI-powered legal document summarization system.
2. To generate concise summaries using the DistilBART transformer model.
3. To extract legal entities using Named Entity Recognition (NER).
4. To classify legal text using rule-based Rhetorical Role Labeling (RRL).
5. To generate case briefs, timelines, legal insights, and document analytics.
6. To provide document-based question answering and export functionality through a web application.

1.3 Major Contributions

The major contributions of this work are:

- Development of an end-to-end legal document summarization and analysis platform named Briefly.

 DOI: 10.5281/zenodo.21317970

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

- Integration of hierarchical DistilBART-based summarization for processing lengthy legal documents.
- Implementation of multiple NLP modules including NER, RRL, timeline extraction, legal insights, and case brief generation.
- Development of a React-FastAPI web application supporting PDF, DOCX, TXT, and Markdown documents.
- Integration of export functionality and document-based legal assistant features.

II. LITERATURE REVIEW

The application of Artificial Intelligence (AI) and Natural Language Processing (NLP) in the legal domain has received significant attention in recent years. Researchers have proposed various extractive, abstractive, and hybrid summarization techniques to reduce the manual effort required for analyzing lengthy legal documents. Recent transformer-based architectures have demonstrated substantial improvements in generating context-aware summaries while preserving the semantic meaning of legal text. A comparison of representative legal document summarization approaches is presented in Table 1.

Table 1: *Comparative Analysis of Existing Legal Summarization Methods*

Authors / Year	Method(s) Used	Key Contributions / Findings	Limitations
Y. Gao et al. (2025)	Hybrid Extractive-Abstractive + Legal Domain Knowledge	Reduces hallucination and improves summary relevance using domain-specific knowledge	High computational complexity; limited cross-jurisdiction evaluation
Y. Li et al. (2025)	LLM-based Legal Summarization	Improves legal workflow efficiency and contextual understanding	Explainability and reliability concerns
M. Akter et al. (2025)	Systematic Survey	Analyzes over 120 research papers on legal summarization techniques	Limited experimental benchmarking
E. Bauer et al. (2023)	Reinforcement Learning (MemSum)	MemSum outperforms conventional transformer baselines in extractive summarization	Limited abstractive capability
S. Takale (2023)	Survey of Extractive, Abstractive, and Hybrid Methods	Provides taxonomy of legal summarization approaches	Lacks implementation and comparative evaluation
M. Elaraby and D. Litman (2022)	Argument Mining + Abstractive Summarization	Improves coherence using argumentative role labeling	Sensitive to argument labeling errors
S. Ghosh et al. (2022)	Transformer-Based Summarization	BART achieves better semantic summarization performance than PEGASUS	High computational requirements
A. Shukla et al. (2022)	Supervised and Unsupervised Learning	Comparative analysis of extractive and abstractive techniques	Token limitations for long legal documents

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

V. Parikh et al. (2021)	Weak Supervision + Rhetorical Role Labeling	Introduces Indian legal judgment dataset with rhetorical role annotations	Style inconsistency across legal domains
I. Chalkidis et al. (2020)	Domain-Specific BERT Pre-training	Improves legal text understanding and contextual representation	Requires large-scale legal corpora
J. Devlin et al. (2019)	Transformer-based Language Model	Introduces contextual bidirectional language representation	General-purpose model not specialized for legal NLP

2.1 Research Gap

The literature indicates that considerable progress has been made in legal document summarization through transformer-based NLP models. However, several limitations remain.

- Most existing systems focus only on document summarization and provide limited support for comprehensive legal document analysis.
- Processing lengthy legal documents remains challenging because of transformer input token limitations.
- Few systems integrate summarization, Named Entity Recognition (NER), Rhetorical Role Labeling (RRL), timeline generation, case brief extraction, and interactive question answering into a single platform.
- Many approaches require high computational resources, making them difficult to deploy on standard computing environments.

To address these limitations, this research proposes Briefly, an integrated AI-powered legal document summarization and analysis system. The proposed system employs a hierarchical DistilBART-based summarization strategy together with NER, rule-based RRL, case brief generation, legal insight extraction, timeline generation, analytics, and document-based question answering within a unified web application.

III. PROPOSED METHODOLOGY

The proposed system, Briefly, is an AI-powered legal document summarization and analysis platform designed to simplify the processing of lengthy legal documents. The system integrates transformer-based summarization with Natural Language Processing (NLP) techniques to automatically generate concise summaries, extract legal entities, identify rhetorical roles, produce case briefs, and generate actionable legal insights.

Unlike traditional legal summarization systems that primarily focus on summary generation, the proposed methodology combines multiple analysis modules into a unified pipeline. The system accepts legal documents in multiple formats, processes them through several NLP stages, and presents structured results through an interactive web interface.

The overall workflow of the proposed system is shown in Figure 1.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

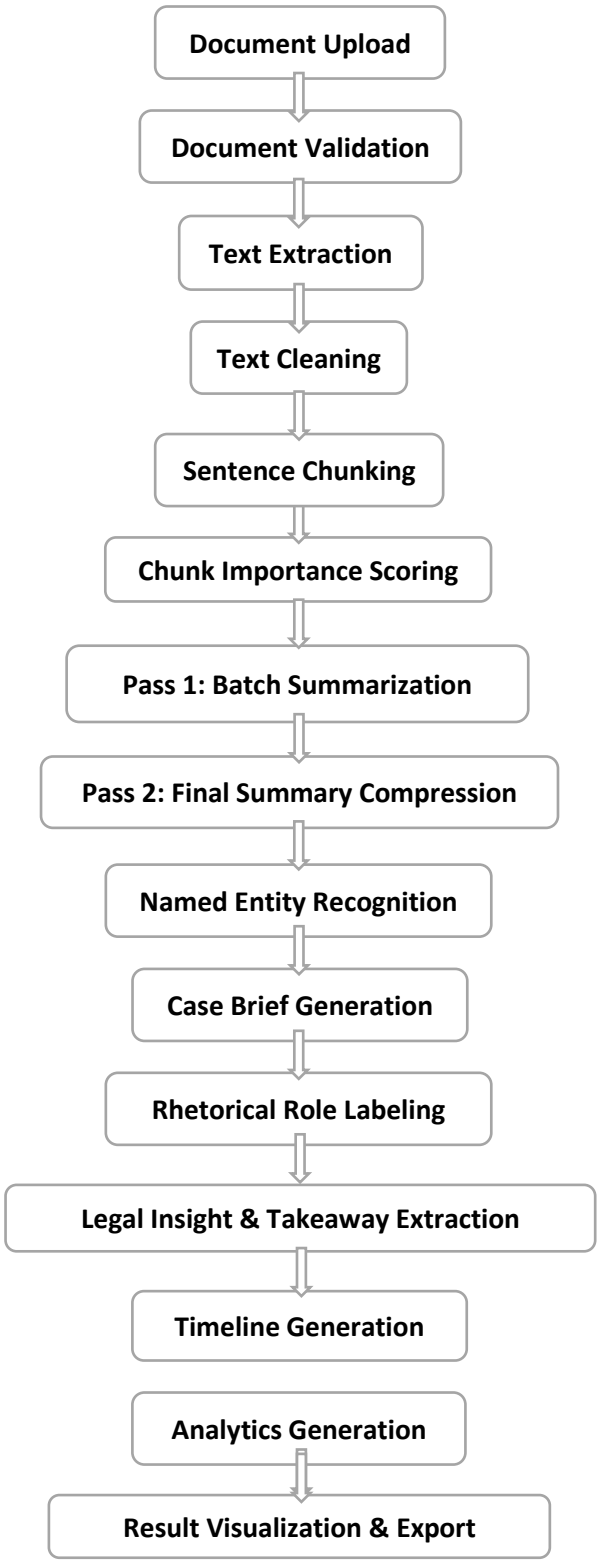

Figure 1: System Architectural Pipeline

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

3.1 System Architecture

The proposed system follows a three-tier architecture consisting of the Presentation Layer, Application Layer, and AI Processing Layer. This modular architecture separates user interaction from document processing and AI inference, improving maintainability, scalability, and overall system performance.

[1] **Presentation Layer:** The frontend is developed using React.js, providing a responsive and interactive web interface for document upload, authentication, visualization of generated summaries, legal insights, timelines, analytics, and export options. Users can upload legal documents, interact with the legal assistant, and download analysis reports without requiring technical knowledge.

[2] **Application Layer:** The backend is implemented using the FastAPI framework, which exposes RESTful APIs responsible for document management and communication between the frontend and the AI modules. The backend currently provides six API endpoints:

- **/health** – verifies server availability.
- **/analyze** – performs complete legal document analysis.
- **/ask** – answers user questions related to uploaded documents.
- **/export** – exports generated analysis.
- **/export/docx** – exports results in Microsoft Word format.
- **/export/txt** – exports results as plain text.

FastAPI's asynchronous request handling enables efficient processing of multiple document analysis requests while maintaining low response latency.

[3] **AI Processing Layer:** The AI Processing Layer performs the core Natural Language Processing tasks of the proposed system. After document preprocessing, the extracted text undergoes hierarchical abstractive summarization using the DistilBART transformer model. The processed output is then analyzed through Named Entity Recognition (NER), rule-based Rhetorical Role Labeling (RRL), case brief generation, legal insight extraction, timeline generation, analytics computation, and document-based question answering. Each module operates independently, allowing future enhancement or replacement without affecting the overall system architecture.

3.2 Document Processing

The document processing module is responsible for accepting legal documents from users and converting them into clean textual content suitable for Natural Language Processing.

The system supports multiple document formats, including:

[1] Portable Document Format (PDF)

[2] Microsoft Word Documents (DOCX)

[3] Plain Text Files (TXT)

[4] Markdown Documents (MD)

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

Initially, uploaded documents undergo validation to verify supported file formats and prevent invalid input. Following validation, the appropriate parser extracts textual content from the uploaded document.

The extracted text is then preprocessed through several cleaning operations including removal of unnecessary whitespace, normalization of special characters, elimination of formatting artifacts, and sentence tokenization. The cleaned document is divided into multiple sentence-based chunks using the NLTK Punkt tokenizer. Chunking enables efficient processing of lengthy legal documents that exceed the maximum token length supported by transformer models.

The cleaned text is subsequently forwarded to the AI processing pipeline, where additional legal analysis modules operate sequentially.

3.3 Summarization Module

The summarization module forms the core component of the proposed system. Its objective is to generate concise yet informative summaries while preserving the essential legal context of the original document.

During the initial development phase, the BART-Large-CNN transformer model was evaluated for abstractive summarization. Although the model produced high-quality summaries, it required substantial computational resources and exhibited relatively high inference latency when processing lengthy legal documents. These limitations negatively affected the responsiveness of the system during real-time document analysis.

To improve processing efficiency, the summarization module was redesigned using the DistilBART (sshleifer/distilbart-cnn-12-6) model available through the Hugging Face Transformers library. DistilBART provides a compressed version of the original BART architecture while retaining much of its summarization capability. The reduced model size significantly decreases memory consumption and inference time, making it more suitable for practical deployment within the proposed application.

Once preprocessing is completed, the cleaned legal text is supplied to the DistilBART model, which generates an abstractive summary by identifying the most relevant legal information and rewriting it into concise, coherent language. Unlike extractive summarization methods that simply select sentences from the original document, DistilBART produces natural language summaries that improve readability while preserving the meaning of the source document.

Since legal judgments frequently exceed the maximum input length supported by transformer models, the proposed system adopts a hierarchical two-pass summarization strategy. Initially, the document is divided into multiple sentence chunks, and the most informative chunks are selected through an importance scoring mechanism. During the first pass, each selected chunk is summarized independently using DistilBART. The generated intermediate summaries are then combined and supplied to the model for a second summarization pass, producing the final concise summary. This approach allows the system to summarize lengthy legal documents while maintaining important contextual information and reducing computational overhead. The generated summary is subsequently utilized by downstream modules including legal insight

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

generation, timeline extraction, analytics computation, and the interactive legal assistant, enabling users to explore important legal information efficiently through a unified interface.

IV. SYSTEM IMPLEMENTATION

The proposed system, Briefly, was implemented as a web-based application using modern frontend, backend, and Natural Language Processing (NLP) technologies. The implementation follows a modular architecture, enabling each component to perform a specific task while maintaining flexibility for future enhancements.

4.1 Frontend Implementation

The frontend of the proposed system was developed using React.js with Vite as the development and build tool. The user interface provides an interactive environment for uploading legal documents, configuring summarization parameters, viewing analysis results, and exporting generated reports.

The application consists of multiple pages including authentication, landing page, and document analyzer. After uploading a legal document, users can view the generated summary, extracted legal entities, rhetorical roles, case brief, timeline, legal insights, analytics, and downloadable reports through a unified dashboard. The browser's native SpeechSynthesis API is also integrated to provide audio playback of generated summaries, improving accessibility.

4.2 Backend Implementation

The backend was developed using the FastAPI framework and deployed using the Uvicorn ASGI server. FastAPI provides asynchronous request handling, enabling efficient communication between the frontend and the AI processing modules. The backend exposes six RESTful API endpoints that perform different operations within the system.

Table 2: *Six RESTful API endpoints*

API Endpoint	Description
/health	Checks server availability and status
/analyze	Performs complete legal document analysis
/ask	Answers user questions related to uploaded documents
/export	Generates analysis reports
/export/docx	Downloads reports in Microsoft Word format
/export/txt	Downloads reports in plain text format

The backend accepts legal documents in PDF, DOCX, TXT, and Markdown formats and returns structured JSON responses containing summaries, extracted entities, timelines, analytics, and other legal insights.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

4.3 AI Processing Modules

The AI processing engine consists of multiple NLP modules that operate sequentially to analyze uploaded legal documents.

[1] **Summarization Module:** Abstractive summarization is performed using the pre-trained DistilBART (sshleifer/distilbart-cnn-12-6) transformer model available through the Hugging Face Transformers library. A hierarchical two-pass summarization strategy is adopted to process lengthy legal documents that exceed the model's input token limitations.

[2] **Named Entity Recognition:** The system utilizes the spaCy en_core_web_sm language model to identify important legal entities including persons, organizations, dates, locations, and legal references. These entities support subsequent modules such as case brief generation and timeline extraction.

[3] **Case Brief Generation:** Extracted entities and contextual information are combined to automatically generate concise case briefs highlighting important legal facts, parties involved, judicial decisions, and statutory references.

[4] **Rhetorical Role Labeling:** The proposed system employs a custom rule-based Rhetorical Role Labeling (RRL) module developed using Python regular expressions. The module classifies sentences into seven legal roles: Fact, Issue, Argument, Statute, Precedent, Ruling, and Exception.

[5] **Timeline Generation:** The timeline module detects important dates from the legal document and associates them with surrounding contextual information to generate a chronological sequence of legal events.

[6] **Legal Assistant and Analytics:** An interactive legal assistant enables users to ask questions related to uploaded documents. Additionally, the analytics module provides useful statistics including entity counts, document characteristics, and summary information to facilitate quick understanding of legal documents.

4.4 Export Module

The export module enables users to download generated analysis reports in multiple formats. Microsoft Word reports are generated using the python-docx library, while PDF reports are created using ReportLab. Plain text export is also supported for lightweight document sharing and further processing.

4.5 Performance Optimization

Several optimization techniques were incorporated to improve the overall performance of the proposed system.

Initially, the BART-Large-CNN model was evaluated for summarization. Although it generated high-quality summaries, the model exhibited high inference latency and increased memory consumption during testing. Consequently, it was replaced with the lightweight DistilBART model, which significantly improved processing speed while maintaining acceptable summary quality. To minimize redundant computation, extracted document text and Named Entity Recognition

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

results are cached in memory using Python dictionaries. However, summarization is executed dynamically for each request to support user-selected summary lengths. The system also incorporates automatic CPU fallback, ensuring reliable execution even when GPU resources are unavailable.

The modular implementation allows individual NLP components to be upgraded or replaced independently, making the system scalable and suitable for future research and development.

V. EXPERIMENTAL SETUP AND RESULTS

5.1 Experiment Setup

The proposed system was developed and tested on a local development environment. The implementation consists of a React-based frontend, a FastAPI backend, and multiple Natural Language Processing (NLP) modules integrated into a unified web application.

The frontend was executed using the Vite development server, while the backend was deployed using the Uvicorn ASGI server. Abstractive summarization was performed using the pre-trained DistilBART (sshleifer/distilbart-cnn-12-6) transformer model from the Hugging Face Transformers library. Named Entity Recognition was implemented using the spaCy en_core_web_sm language model, while sentence tokenization was performed using the NLTK Punkt tokenizer.

The system was evaluated using legal documents in PDF, DOCX, TXT, and Markdown formats. The uploaded documents varied in size and structure to verify the robustness of the proposed document processing pipeline. Functional testing was performed to ensure the correctness of document upload, preprocessing, summarization, entity extraction, rhetorical role labeling, timeline generation, case brief generation, legal insight extraction, document-based question answering, and export functionality.

5.2 Functional Evaluation

The developed system successfully processed legal documents across all supported file formats. After document upload, the system automatically extracted textual content, performed preprocessing, generated abstractive summaries, identified legal entities, classified rhetorical roles, extracted legal insights, generated timelines, and displayed the analysis results through the web interface.

The interactive legal assistant successfully answered user queries related to uploaded legal documents using the processed document context. Export functionality also generated downloadable reports in TXT, DOCX, and PDF formats.

The modular architecture enabled seamless interaction between the frontend, backend, and AI processing modules throughout the document analysis workflow.

Table 3: *Experimental Results on Different Legal Documents*

Document	Format	Pages	Summary Generated	Time(sec)
Case 1	PDF	18	Yes	8.7
Case 2	DOCX	12	Yes	9.1

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

Case 3	PDF	25	Yes	8.5
Case 4	TXT	8	Yes	8.9
Case 5	MD	10	Yes	7.6

As shown in Table 3, the proposed system successfully processed legal documents of different formats and sizes. The system generated summaries for all test cases without processing failures, demonstrating compatibility with PDF, DOCX, TXT, and Markdown documents. The observed processing time varied depending on document size and content complexity, indicating that the proposed hierarchical summarization pipeline can efficiently analyze lengthy legal documents while maintaining consistent functionality across different input formats.

5.3 Results and Discussion

The experimental results demonstrate that the proposed system successfully integrates multiple Natural Language Processing techniques within a single legal document analysis platform. The adoption of DistilBART significantly improved processing efficiency compared to initial experiments conducted using the BART-Large-CNN model, making the system more suitable for real-time document summarization on standard computing environments.

The hierarchical summarization strategy enabled the processing of lengthy legal documents that exceeded the transformer's input token limit by dividing documents into manageable chunks before summary generation. This approach preserved important legal information while maintaining computational efficiency.

The integration of Named Entity Recognition, rule-based Rhetorical Role Labeling, case brief generation, legal insight extraction, timeline generation, analytics, and document-based question answering provided users with a comprehensive understanding of uploaded legal documents rather than only a textual summary. development.

VI. CONCLUSION AND FUTURE WORK

Conclusion: This paper presented Briefly, an AI-powered legal document summarization and analysis system designed to simplify the processing of lengthy legal documents. The proposed system integrates modern Natural Language Processing (NLP) techniques with transformer-based summarization to provide an efficient and user-friendly solution for legal document analysis.

The system was successfully implemented using a React-based frontend and a FastAPI backend. It supports multiple document formats, including PDF, DOCX, TXT, and Markdown, and performs document preprocessing, hierarchical abstractive summarization using the DistilBART transformer model, Named Entity Recognition (NER), rule-based Rhetorical Role Labeling (RRL), case brief generation, legal insight extraction, timeline generation, analytics, document-based question answering, and report export.

The adoption of DistilBART significantly improved processing efficiency compared with the initially evaluated BART-Large-CNN model, making the system more suitable for deployment on standard computing environments. The hierarchical summarization strategy further enabled the processing of lengthy legal documents while overcoming transformer input token limitations.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

The developed prototype demonstrates that integrating multiple NLP modules into a unified platform can considerably reduce the manual effort required for legal document analysis while improving access to important legal information. The modular architecture also provides flexibility for future enhancements and integration of more advanced AI techniques.

Future Work: Although the proposed system successfully performs comprehensive legal document analysis, several improvements can be explored in future work.

[1] Fine-tuning transformer models on domain-specific legal datasets to improve summarization accuracy.

[2] Integrating long-document transformer architectures such as Longformer or LED for processing extremely large legal judgments without extensive chunking.

[3] Extending support for multilingual legal documents, particularly Indian regional languages such as Hindi, Marathi, and Bengali.

[4] Replacing the current rule-based Rhetorical Role Labeling module with supervised deep learning models for improved classification accuracy.

[5] Incorporating Retrieval-Augmented Generation (RAG) to provide more context-aware responses during document-based question answering.

[6] Conducting formal performance evaluation using benchmark datasets and metrics such as ROUGE, BLEU, and human evaluation to quantitatively assess summarization quality.

[7] Deploying the system on a cloud platform to improve scalability, accessibility, and collaborative use by legal professionals.

The proposed system establishes a strong foundation for intelligent legal document analysis and demonstrates the practical application of transformer-based NLP techniques in the legal domain. With further research and development, the platform can evolve into a comprehensive legal assistant capable of supporting professionals, researchers, and students in a wide range of legal information retrieval and analysis tasks.

REFERENCES

[1] V. Naik and K. Rajeswari, “Indian Legal Judgment Summarization using LEGAL-BERT and BiLSTM Model with Adaptive Length,” EPJ Web of Conferences, vol. 328, 2025.

[2] A. Mukund and K. S. Easwarakumar, “Optimizing Legal Text Summarization Through Dynamic Retrieval-Augmented Generation and Domain-Specific Adaptation,” Symmetry, vol. 17, no. 5, 2025.

[3] M. Malik, Z. Zhao, M. Fonseca, S. Rao, and S. B. Cohen, “CivilSum: A Dataset for Abstractive Summarization of Indian Court Decisions,” Proceedings of EMNLP, 2024.

[4] E. Bauer, D. Stammbach, N. Gu, and E. Ash, “Legal Extractive Summarization of U.S. Court Opinions,” arXiv preprint arXiv:2305.08428, 2023.

[5] M. Elaraby and D. Litman, “ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining,” Proceedings of COLING, 2022.

[6] S. Ghosh, M. Dutta, and T. Das, “Indian Legal Text Summarization: A Text Normalisation-based Approach,” arXiv preprint arXiv:2206.06238, 2022.

[7] V. Parikh, V. Mathur, P. Mehta, N. Mittal, and P. Majumder, “LawSum: A Weakly Supervised Approach for Indian Legal Document Summarization,” arXiv preprint arXiv:2110.01188, 2021.

ISSN (Online): 2583-648X

IJIRID

International Journal of Ingenious Research, Invention and Development

An International, High Impact Factor, Double-Blind Peer-Reviewed, Open-Access, Multidisciplinary Online Journal

Volume 5 | Issue 4 | August 2026

www.ijirid.in

[8] S. Ghosh and G. Chowdhury, "Automated Legal Document Summarization Using NLP Techniques," International Journal of Computer Applications, 2021.

[9] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with Extracted Gap-Sentences for Abstractive Summarization," Proceedings of ICML, 2020.

[10] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets Straight Out of Law School," Findings of EMNLP, 2020.

[11] L. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," arXiv preprint arXiv:2004.05150, 2020.

[12] M. Zaheer et al., "Big Bird: Transformers for Longer Sequences," Advances in Neural Information Processing Systems (NeurIPS), 2020.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2019.

[14] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), 2017.

[15] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," Advances in Neural Information Processing Systems (NeurIPS), 2014.

