AI-based Legal Document Summarization and Case Prediction System

Nagamani Thanuboddi
Assistant Professor
Department of CAI
KKR&KSR Institute of Technology and Sciences, Guntur,
Andhra Pradesh,India

Mudiga Prachay Kumar
Student
Department of CAI
KKR&KSR Institute of Technology and Sciences, Guntur,
Andhra Pradesh,India

Garikapati Nani
Student
Department of CAI
KKR&KSR Institute of Technology and Sciences, Guntur,
Andhra Pradesh,India

Nayeni Sai Yuvan Chaitanya Reddy
Student
Department of CAI
KKR&KSR Institute of Technology and Sciences, Guntur,
Andhra Pradesh, India

Pasupula Pradeep Kumar
Student
Department of CAI
KKR&KSR Institute of Technology and Sciences, Guntur,
Andhra Pradesh, India

ABSTRACT
The increasing volume of digital legal documents such as court judgments, FIR records, contracts, and legal orders has made manual analysis both time-consuming and difficult for legal professionals. This study proposes LawTech AI, an AI-powered legal document intelligence system designed to transform unstructured legal text into structured and searchable legal information. The system follows a multi-stage pipeline where legal documents are first processed to extract key entities using Legal Named Entity Recognition (NER). The extracted information is then analyzed through the FASSI workflow (Fetch, Analyze, Summarize, Store, and Interact) to understand the legal context of the document. To enable efficient legal search and precedent discovery, document embeddings are generated and stored in a FAISS vector database, allowing the system to retrieve the top similar cases. These retrieved cases are used within a Retrieval Augmented Generation (RAG) framework to assist a fine-tuned legal language model in generating structured summaries, identifying legal issues, and providing insights based on relevant precedents. The proposed approach aims to support lawyers, legal researchers, and students by simplifying legal document analysis and improving access to meaningful legal knowledge.

General Terms
Artificial Intelligence, Natural Language Processing, Deep Learning, Information Retrieval Systems, Legal Analytics, Text Processing, FAISS Vector Database, Retrieval Augmented Generation (RAG).

Keywords
Legal Document Intelligence, Legal Named Entity Recognition (NER), FAISS Vector Database, Retrieval Augmented Generation (RAG), Legal AI, Case Law Analysis, Legal Document Processing

1. INTRODUCTION
The rapid digitization of judicial systems and legal institutions has resulted in a massive increase in the availability of digital legal documents, including court judgments, case records, statutes, and contracts. While this transformation has improved accessibility, it has also introduced significant challenges in efficiently analyzing and extracting meaningful information from large volumes of unstructured legal text. Legal professionals, researchers, and students often spend considerable time manually reviewing lengthy documents to identify relevant facts, legal provisions, and precedents.

Traditional methods of legal document analysis primarily rely on keyword-based search and manual interpretation, which are often inefficient and fail to capture the semantic relationships within legal texts. Legal language is inherently complex, containing domain-specific terminology, intricate sentence structures, and contextual dependencies that make automated processing difficult. As a result, there is a growing need for intelligent systems that can assist in understanding, summarizing, and retrieving relevant legal information effectively.

Recent advancements in **Natural Language Processing (NLP)** and **Artificial Intelligence (AI)** have enabled the development of automated tools for text analysis, information extraction, and summarization. Techniques such as Named Entity Recognition (NER), semantic embeddings, and transformer-based language models have shown promising results in processing complex textual data. However, many existing solutions focus on isolated tasks such as classification, retrieval, or summarization, without providing an integrated framework for comprehensive legal document analysis.

To address these challenges, this paper proposes **LawTech AI**, an AI-based legal document summarization and case prediction system. The proposed system combines multiple advanced techniques, including Legal Named Entity Recognition for extracting key legal entities, semantic embedding generation for capturing contextual meaning, and FAISS-based vector search for efficient retrieval of similar cases. Furthermore, the system incorporates a **Retrieval-Augmented Generation (RAG)** framework, which leverages retrieved case information to generate structured summaries and insights using a fine-tuned language model.

The primary objective of this research is to transform unstructured legal documents into structured, searchable, and interpretable information, thereby reducing manual effort and improving the efficiency of legal research. The proposed system not only facilitates quick understanding of complex legal texts but also assists in identifying relevant precedents, making it valuable for legal practitioners, researchers, and students.

2. RELATED WORK

The application of Artificial Intelligence (AI) and Natural Language Processing (NLP) in the legal domain has gained significant attention in recent years, particularly for tasks such as legal document classification, case prediction, information retrieval, and summarization. Early research by Ion Androutsopoulos and Ilias Chalkidis [1] explored neural approaches for legal judgment prediction using machine learning techniques. Similarly, Nikolaos Aletras et al. [2] demonstrated that machine learning models can predict decisions of the European Court of Human Rights by analyzing textual features of legal cases. Although these approaches showed promising results, they often struggled with interpretability and generalization across diverse legal systems.

The introduction of transformer-based models, particularly BERT [6], significantly improved the performance of NLP tasks. Building on this, Legal-BERT [3] was developed as a domain-specific model trained on legal corpora, enabling better understanding of legal terminology and context. These models achieved improved accuracy in tasks such as classification and entity recognition; however, they require substantial computational resources and lack efficient retrieval mechanisms.

For semantic search and document similarity, Sentence-BERT [7] introduced an effective approach to generate dense vector representations of text, enabling similarity-based retrieval. Additionally, FAISS [8] provides an efficient indexing mechanism for large-scale vector search, making it suitable for retrieving similar legal documents from extensive datasets. Nevertheless, retrieval-based systems alone do not generate meaningful summaries or insights, limiting their practical usability.

To overcome these limitations, Retrieval-Augmented Generation (RAG) [9] combines document retrieval with generative models, enabling context-aware response generation. RAG-based approaches have shown significant improvements in knowledge-intensive NLP tasks by leveraging external document sources during generation. Furthermore, large-scale language models such as GPT [10] have demonstrated strong capabilities in few-shot learning and text generation, contributing to advancements in legal text understanding.

In addition, benchmark datasets such as RCV1 [5] have supported research in text categorization, while models like the Multilingual Universal Sentence Encoder [11] have enabled efficient semantic retrieval across multiple languages. Recent studies [4] also highlight the broader impact of NLP in legal systems, including document summarization and legal analytics. Despite these advancements, existing solutions often focus on individual components such as prediction, retrieval, or summarization. There remains a lack of integrated systems that combine these functionalities into a unified framework. The proposed LawTech AI system addresses this gap by incorporating entity extraction, semantic retrieval, and RAG-based summarization into a single pipeline, thereby improving both efficiency and usability in legal document analysis.

3. MATERIALS AND METHODS

3.1 DATASETS USED

The dataset used in this study was obtained from the Kaggle repository **"Legal Dataset: SC Judgments India (1950–2024)"**. It contains Supreme Court of India judgment records collected from publicly available legal sources. The dataset includes case details and judgment texts, which were used as the primary data for extracting legal information and analyzing documents in the proposed system.

3.2 SYSTEM ARCHITECTURE

The proposed **LawTech AI** system is designed as a modular pipeline that processes unstructured legal documents and converts them into structured, searchable, and meaningful information. The architecture integrates multiple Artificial Intelligence (AI) and Natural Language Processing (NLP) components to enable efficient legal document analysis, retrieval, and summarization.

The overall architecture of the system is illustrated in **Fig. 1**, which depicts the flow of data from input processing to final output generation

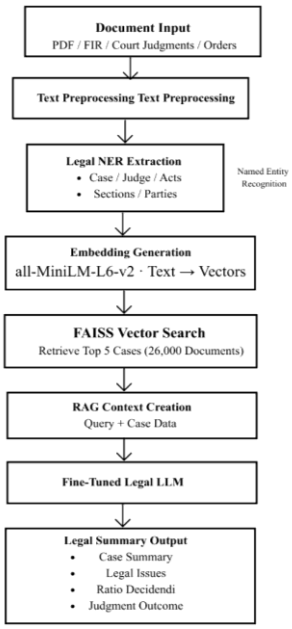

Fig. 1: System Architecture of LawTech AI

3.2.1 Input and Preprocessing Module

The system accepts legal documents such as court judgments, case records, and legal reports in textual format. The preprocessing stage includes:

- Text cleaning (removal of noise, special characters)
- Tokenization and sentence segmentation
- Stop-word removal and normalization

This step ensures that the input data is in a structured format suitable for further analysis.

3.2.2 Legal Named Entity Recognition (NER) Module

In this stage, domain-specific Named Entity Recognition is applied to extract key legal entities, including:

- Case names
- Courts and Judges

- Legal sections and acts
- Parties involved

This module transforms raw legal text into structured data, enabling better understanding and downstream processing.

3.2.3 Embedding Generation Module

The processed text is converted into dense vector representations using sentence embedding models. These embeddings capture the semantic meaning of legal documents, allowing the system to understand contextual similarities beyond keyword matching.

3.2.4. FAISS-Based Vector Database

The generated embeddings are stored in a **FAISS (Facebook AI Similarity Search)** index. This component enables efficient similarity-based retrieval of legal documents by performing approximate nearest neighbor (ANN) search. It significantly reduces retrieval time while maintaining high accuracy.

3.2.5 Retrieval Module

When a query or new legal document is provided, the system searches the FAISS database to retrieve the most relevant cases based on semantic similarity. This ensures that retrieved documents share contextual and legal relevance rather than simple keyword overlap.

3.2.6. Retrieval-Augmented Generation (RAG) Module

The retrieved documents are passed to a Retrieval-Augmented Generation framework, where a fine-tuned language model generates:

- Structured summaries
- Key legal insights
- Relevant case interpretations

This module enhances the quality of output by incorporating external knowledge into the generation process.

3.2.7 Output and Interaction Module

The final output is presented to the user in a structured and readable format, including:

- Summarized case content
- Extracted legal entities
- Suggested similar cases

This enables users to quickly understand complex legal documents and make informed decisions.

3.3 METHODOLOGY

The proposed **LawTech AI** system follows a structured and multi-stage methodology for analyzing legal documents, retrieving relevant cases, and generating meaningful summaries. The methodology integrates Natural Language Processing (NLP), semantic search, and generative AI techniques into a unified workflow.

Step 1: Data Acquisition and Preprocessing

Legal documents are collected from publicly available datasets, including court judgments and case records. The preprocessing stage ensures data quality and consistency through:

- Removal of noise, special characters, and irrelevant symbols
- Tokenization and sentence segmentation
- Lowercasing and normalization

This step converts raw legal text into a clean and structured format suitable for further analysis.

Step 2: Legal Named Entity Recognition (NER)

A domain-specific Named Entity Recognition model is applied to extract key legal entities such as:

- Case name
- Court and judge
- Legal acts and sections
- Parties involved

This process transforms unstructured legal text into structured information, enabling better interpretability and downstream processing.

Step 3: Semantic Embedding Generation

The preprocessed text is converted into dense vector representations using sentence embedding models. These embeddings capture semantic relationships within legal documents, allowing the system to identify similarity based on meaning rather than exact keyword matches.

Step 4: Vector Storage using FAISS

The generated embeddings are stored in a **FAISS vector database**, which enables efficient similarity search using Approximate Nearest Neighbor (ANN) techniques. This approach significantly reduces retrieval time while maintaining high accuracy.

Step 5: Similar Case Retrieval

Given a query or new legal document, the system computes its embedding and retrieves the top-K most similar documents from the FAISS database. The similarity between documents is measured using cosine similarity:

$$\cos(\theta) = \frac{A \cdot B}{||A|| \; ||B||}$$

This step ensures that retrieved cases are contextually relevant.

Step 6: Retrieval-Augmented Generation (RAG)

The retrieved documents are used as contextual input to a Retrieval-Augmented Generation framework. A fine-tuned language model generates structured summaries and insights based on both the input document and retrieved cases.

The probability of generating an output is defined as:

$$P(y \mid x) = \sum_{z \in D} P(y \mid x, z) \cdot P(z \mid x)$$

where:

- $x$represents the input document
- $z$represents retrieved documents
- $y$represents the generated output

Step 7: Summary Generation and Output

The final output includes:

- Concise summary of the legal document
- Extracted key entities
- List of similar cases

The generated summaries are evaluated using ROUGE metrics to measure their quality and relevance.

3.4 Evaluation Metrics

For To evaluate the performance of the proposed **LawTech AI** system, multiple quantitative metrics are used to assess the effectiveness of entity extraction, document retrieval, and summary generation. These metrics ensure a comprehensive evaluation of the system across different functional components.

1. ROUGE Score (Summarization Evaluation)

The quality of generated summaries is evaluated using the **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** metric. Specifically, **ROUGE-1**, **ROUGE-2**, and **ROUGE-L** are used to measure the overlap between the generated summary and the reference summary.

$$ROUGE\text{-}1 = \frac{\sum_{w \in Ref} Coun\,t_{match}(w)}{\sum_{w \in Ref} Count(w)}$$

ROUGE-1 measures unigram overlap, ROUGE-2 evaluates bigram overlap, and ROUGE-L captures the longest common subsequence between the generated and reference summaries. Higher ROUGE scores indicate better summarization quality and content relevance.

2. Cosine Similarity (Retrieval Evaluation)

To measure the similarity between legal documents, cosine similarity is used on vector embeddings. It determines how closely two documents are related in semantic space.

$$\cos(\theta) = \frac{A \cdot B}{||\,A\,||\;\;||\,B\,||}$$

where $A$and Brepresent embedding vectors of two documents. A value closer to 1 indicates higher similarity.

3. Precision@K

Precision@K measures the proportion of relevant documents among the top-K retrieved results:

$$Precision@K = \frac{\text{Number of relevant documents retrieved in top K}}{K}$$

This metric evaluates the accuracy of the retrieval system.

4. Recall@K

Recall@K measures the ability of the system to retrieve all relevant documents:

$$Recall@K = \frac{\text{Number of relevant documents retrieved in top K}}{\text{Total number of relevant documents}}$$

Higher Recall@K indicates better coverage of relevant cases.

5. Latency (Retrieval Efficiency)

Latency measures the time required to retrieve relevant documents from the FAISS database. It is typically measured in milliseconds (ms). Lower latency indicates a more efficient retrieval system.

6. RAG-Based Generation Probability

The Retrieval-Augmented Generation (RAG) framework models the probability of generating output $y$given input $x$as:

$$P(y \mid x) = \sum_{z \in D} P(y \mid x, z) \cdot P(z \mid x)$$

where:

- $x$is the input query or document
- $z$represents retrieved documents
- $y$is the generated output

This formulation ensures that the generated summary is conditioned on both the input and relevant retrieved context.

4. RESULTS AND DISCUSSION

The performance of the proposed **LawTech AI** system was evaluated using a dataset of approximately **5,000 Indian Supreme Court judgments (1950–2024)**. The system was tested across three major components: entity extraction, document retrieval, and summary generation. The results were compared with two baseline approaches, namely **keyword-based retrieval (TF-IDF/BM25)** and **Legal-BERT-based dense retrieval**.

4.1 Entity Extraction Performance

The Legal Named Entity Recognition (NER) module was able to accurately extract key legal entities such as case names, judges, courts, legal sections, and involved parties. The extracted entities were consistent across most documents, demonstrating the effectiveness of domain-specific NER in handling legal text.

The structured representation of entities significantly improved downstream tasks such as indexing and retrieval. However, minor challenges were observed in cases involving ambiguous legal terminology and complex sentence structures.

4.2 Retrieval Performance

The retrieval performance of the system was evaluated using **Precision** and **Recall** metrics. The proposed system outperformed both baseline methods due to the use of semantic embeddings and FAISS-based similarity search.

Table 1. Precision and Recall comparasion between proposed and Baseline Models

Metric	Keyword-Based	Legal-BERT	Proposed (LawTech AI)
Precision	0.41	0.63	0.79
Recall	0.37	0.58	0.74

The results show that:

- The proposed system improves precision by approximately **25.4%** over Legal-BERT
- Recall is improved by approximately **27.6%**, indicating better coverage of relevant cases

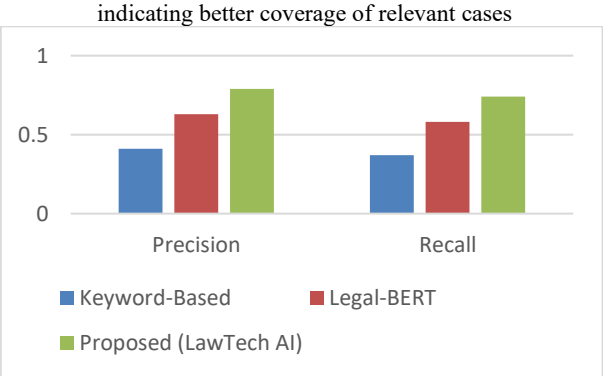

Fig. 2: Improvement in Precision and Recall over Legal-BERT

This improvement is attributed to the use of **semantic similarity search**, which captures contextual relationships rather than relying on keyword matching.

4.3 Summarization Performance

The quality of generated summaries was evaluated using ROUGE metrics. The RAG-based approach produced more informative and context-aware summaries compared to baseline methods.

Table 2. ROUGE Score Comparison of Summarization Methods

Metric	Keyword-Based	Legal-BERT	Proposed (LawTech AI)
ROUGE-1	0.31	0.52	0.71
ROUGE-2	0.18	0.39	0.58
ROUGE-L	0.27	0.48	0.66

The improvements indicate that:

- The summaries generated by the proposed system retain more relevant information
- Context from retrieved documents enhances the quality of generated output
- The RAG framework effectively reduces information loss during summarization

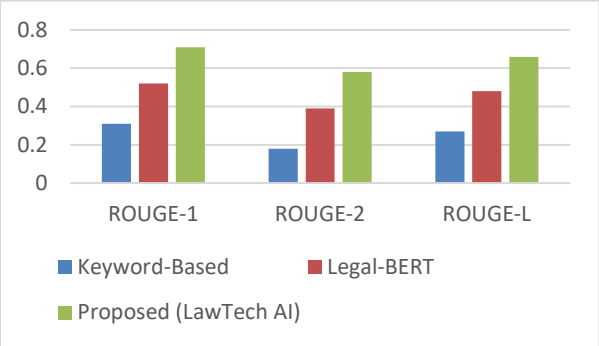

Fig. 3: Improvement in ROUGE Scores over Legal-BERT

4.3 Retrieval Efficiency (Latency Analysis)

The use of the FAISS vector database significantly reduced retrieval time. The proposed system achieved an average latency of **94 milliseconds**, which is substantially lower than traditional retrieval methods.

This improvement is mainly due to:

- Approximate Nearest Neighbor (ANN) search
- Efficient vector indexing mechanisms

Lower latency ensures that the system can be used in real-time legal applications.

4.5 Discussion

The experimental results demonstrate that the integration of **NER, semantic embeddings, FAISS, and RAG** provides a comprehensive solution for legal document analysis. Unlike traditional keyword-based systems, the proposed approach captures semantic meaning, leading to more accurate retrieval and better summarization.

The system also shows strong scalability, as FAISS enables efficient handling of large datasets. Additionally, the use of Retrieval-Augmented Generation enhances contextual understanding, allowing the model to generate more meaningful summaries.

Table 3. Over all Performance Comparison of Proposed System with Baseline Methods

Evaluation Metric	Keyword -Based Search	Legal-BERT Retrieval	Law Tech AI	LawTech AI vs Legal-BERT
Precision	0.41	0.63	0.79	+25.4%
Recall	0.37	0.58	0.74	+27.6%
ROUGE-1	0.31	0.52	0.71	+36.5%
ROUGE-2	0.18	0.39	0.58	+48.7%
ROUGE-L	0.27	0.48	0.66	+37.5%
Summary Relevance (0–1)	0.43	0.61	0.75	+22.9%

However, some limitations remain:

- Ambiguity in legal language can affect entity extraction accuracy
- The system depends on the quality of embeddings for retrieval performance
- Fine-tuned models require computational resources

Despite these challenges, the proposed system demonstrates significant improvements over existing approaches and provides a practical solution for real-world legal applications.

5. CONCLUSION

In the present work, the possibility of using an AI-Powered Legal Document Intelligence System (LawTech AI) in the analysis of large groups of legal documents has been examined. The work has been specifically targeted at the transformation of unstructured legal texts into a more organized and queryable form through the application of modern Natural Language Processing techniques. The Legal Named Entity Recognition (NER) technique has been employed to recognize significant information such as the names of cases, courts, judges, legal acts, sections, and parties involved in the legal case.

The proposed workflow included incorporating the FASSI approach, which facilitated the system in fetching, analyzing, summarizing, storing, and interacting with legal document data in a systematic manner. To facilitate the retrieval of cases, document embeddings were created and stored in a FAISS vector database, allowing for similarity-based document search in legal cases. Once a query or document is given, the relevant cases are fetched and used as context in a model known as Retrieval Augmented Generation (RAG), allowing the model to generate a concise summary and provide insights that enable a user to easily understand the key points in a case. The results of the present work indicate that the integration of entity extraction, semantic search, and AI-based summarization can greatly contribute to the analysis of legal documents. The proposed system can be helpful to lawyers, legal researchers, as well as law students in effectively dealing with large amounts of legal information.

6. ACKNOWLEDGMENTS

I would like to thank my project guide for the time and guidance provided while working on this project. The discussions and suggestions provided during the development of the work were helpful in improving the overall study. I also thank my department for giving the opportunity to carry out this project. I am also grateful to my friends for the support provided to me throughout the project work.

7. REFERENCES

[1] D. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural Legal Judgment Prediction in English," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4317–4323, 2019.

[2] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective," *PeerJ Computer Science*, vol. 2, pp. 1–19, 2016.

[3] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets Straight Out of Law School," *Findings of the Association for Computational Linguistics (EMNLP)*, pp. 2898–2904, 2020.

[4] 4. J. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5218–5230, 2020.

[5] 5. D. Lewis, J. Yang, T. Rose, and F. Li, "RCV1: A New Benchmark Collection for Text Categorization Research," *Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.

[6] 6. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[7] 7. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, 2019.

[8] 8. J. Johnson, M. Douze, and H. Jégou, "Billion-scale Similarity Search with FAISS," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

[9] 9. P. Lewis, E. Perez, A. Piktus et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[10] 10. T. Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[11] 11. Y. Yang, Y. Cer, A. Ahmad, M. Guo, J. Law, and N. Constant, "Multilingual Universal Sentence Encoder for Semantic Retrieval," *Proceedings of ACL*, pp. 87–94, 2019