International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 05 | May 2026 www.irjet.net p-ISSN: 2395-0072

Court Sense for Case Analysis: Legal Document Classification and Analysis using NLP and Deep Learning

Mrs. Megharani Deore, Dr. Kiran Rathore

Student CES Department Oriental University Indore, MP, India
Assistant professor CES Department Oriental University Indore, MP, India

---

Abstract - *Legal documents are often lengthy, complex, and difficult to analyze manually due to their unstructured nature and technical language. This paper presents **CourtSense for Case Analysis**; an intelligent legal document analysis system developed using Natural Language Processing (NLP) and deep learning techniques. The proposed system is designed to automate important legal tasks such as document summarization, case outcome prediction, and similar case retrieval. The system accepts legal content in the form of PDF files, direct text input, or web URLs. Extracted text is pre-processed using cleaning and normalization techniques to remove unnecessary patterns and improve text quality. Since transformer-based models have token limitations, a chunking approach similar to a sliding window mechanism is used to process long legal documents without losing contextual information. The T5 transformer model is utilized to generate concise summaries, while a fine-tuned LegalBERT model is used for predicting legal outcomes with confidence scores. The system also identifies key legal issues and jurisdiction-related information from the input text. In addition, CourtSense provides similar case suggestions using a search-based retrieval mechanism, with future scope for semantic retrieval using dense vector embeddings. The application is implemented using Flask and PyTorch, providing a scalable and user-friendly interface for legal analysis. Experimental evaluation can be performed using ROUGE and METEOR metrics for summarization, and Precision, Recall, and F1-score for prediction tasks. The proposed system aims to reduce manual effort, improve efficiency, and support intelligent decision-making in legal workflows through automated document understanding and analysis.*

Key Words: Natural Language Processing (NLP), Legal Document Analysis, Deep Learning, LegalBERT, T5 Transformer, Text Summarization, Case Outcome Prediction, Semantic Retrieval, Flask, Artificial Intelligence in Law

1. INTRODUCTION

The legal industry generates a massive amount of digital information every day in the form of case documents, judgments, petitions, and legal reports. Most of these documents are unstructured, lengthy, and written using complex legal terminology, making manual analysis a difficult and time-consuming process. Legal professionals often spend significant time reading and understanding case documents to identify important information, summarize content, and search for related cases. With the increasing growth of digital legal data, there is a strong need for intelligent systems that can automate legal document analysis efficiently.

Recent advancements in Natural Language Processing (NLP) and deep learning have created new opportunities for developing automated systems capable of understanding and processing human language. Transformer-based models such as T5 and LegalBERT have shown strong performance in tasks like text summarization, classification, and semantic understanding. These models can capture contextual meaning more effectively compared to traditional machine learning techniques. However, applying these models to legal documents remains challenging because legal texts are usually very long, while transformer models have limited token-processing capacity.

To address these challenges, this project proposes **CourtSense for Case Analysis**, an intelligent legal document analysis system that combines NLP and deep learning techniques to automate important legal tasks. The system is capable of processing legal documents from multiple input sources such as PDF files, direct text input, and web URLs. It performs preprocessing and text cleaning to remove unwanted formatting and improve text quality before analysis.

For handling long legal documents, the system uses a chunking mechanism similar to a sliding window approach, where large documents are divided into smaller segments and processed individually. The T5 transformer model is used to generate concise summaries of legal content, while a fine-tuned LegalBERT model is used for legal outcome prediction. The system also identifies key legal issues, jurisdiction information, and retrieves similar legal cases to assist users in legal research and analysis.

CourtSense is implemented using Flask and PyTorch, providing a scalable and user-friendly web-based platform. The proposed system aims to reduce manual effort, improve efficiency in legal workflows, and support faster decision-making by automating the understanding and analysis of complex legal documents.

1.1 BACKGROUND

The legal field has experienced significant digital transformation in recent years, resulting in the generation of large volumes of electronic legal documents such as court

judgments, case reports, contracts, and petitions. Most of these documents are unstructured and contain complex legal terminology, making manual analysis difficult and time-consuming. Legal professionals often spend considerable time reviewing documents to identify important information, summarize case details, and search for related precedents.

Traditional document management systems mainly rely on keyword-based searching techniques, which are often unable to understand the actual context and meaning of legal text. As the amount of legal data continues to grow, there is an increasing demand for intelligent systems that can process and analyze legal documents more efficiently.

Advancements in Natural Language Processing (NLP) and deep learning have made it possible to automate several language-related tasks such as text classification, summarization, and semantic understanding. These technologies can help improve the efficiency of legal workflows by reducing manual effort and providing faster access to relevant legal information.

1.2 Role of NLP and Deep Learning in Legal Analysis

Natural Language Processing (NLP) and deep learning techniques have become important tools for analyzing large amounts of textual data. In the legal domain, these technologies help automate tasks such as document classification, text summarization, information extraction, and legal outcome prediction. By understanding the context and structure of language, NLP models can process legal documents more efficiently than traditional rule-based systems.

Recent transformer-based models such as T5 and LegalBERT have significantly improved the performance of language understanding tasks. These models are capable of capturing contextual relationships within legal text, allowing better interpretation of complex legal terminology and case information. LegalBERT is specifically designed for legal language processing, while T5 is widely used for generating meaningful summaries from lengthy documents.

Deep learning models also support semantic analysis, which helps in identifying related legal cases based on meaning rather than exact keyword matching. This improves the accuracy of legal research and reduces the time required for manual document review. As a result, NLP and deep learning technologies are becoming increasingly valuable in developing intelligent legal assistance systems.

2. LITERATURE SURVEY

The use of Natural Language Processing (NLP) and deep learning in the legal domain has increased significantly in recent years. Researchers have developed various systems for legal document classification, judgment prediction, document summarization, and case retrieval. Traditional legal analysis systems mainly depended on keyword-based search methods, which often failed to understand the contextual meaning of legal text. As legal datasets became larger and more complex, machine learning and transformer-based approaches started gaining importance in legal text processing.

Transformer architectures such as BERT, T5, and LegalBERT have improved the performance of legal language understanding tasks. LegalBERT, which is trained on legal corpora, performs better in understanding legal terminology and contextual relationships compared to general NLP models. Similarly, the T5 transformer model has shown effective results in abstractive text summarization by generating concise summaries from lengthy legal documents. Despite these advancements, challenges such as long-document handling, semantic retrieval, and interpretability of predictions still remain important research areas.

Table -1: Sample Table format

Author / Model	Technique Used	Application Area	Limitations
Traditional Legal Retrieval Systems	Keyword-based Search	Legal Document Retrieval	Unable to understand semantic meaning
SVM / Naive Bayes Models	Machine Learning Classification	Legal Text Classification	Requires manual feature engineering
BERT-based Models	Contextual Embeddings	Legal Text Understanding	Limited token handling
LegalBERT	Domain-Specific Transformer	Legal Outcome Prediction	High computational requirements
T5 Transformer	Abstractive Summarization	Legal Document Summarization	Difficulty handling very large documents
Existing Retrieval Systems	Search-based Matching	Similar Case Retrieval	Limited semantic relevance

The analysis of existing research indicates that transformer-based models provide better contextual understanding and improved prediction performance compared to traditional machine learning techniques. However, most existing systems focus only on individual tasks such as summarization or prediction. The proposed system, **CourtSense for Case Analysis**, integrates document summarization, legal outcome prediction, and similar case retrieval within a single framework. Additionally, the system uses chunk-based processing techniques to handle long legal documents more effectively.

3. PROPOSED METHODOLOGY

The proposed system, **CourtSense for Case Analysis**, is designed to automate the analysis of legal documents using Natural Language Processing (NLP) and deep learning techniques. The system performs multiple tasks such as legal document summarization, outcome prediction, extraction of key legal information, and retrieval of related cases. The complete workflow of the system is divided into several stages, including document acquisition, preprocessing, long-document handling, summarization, prediction, and result Generation.

3.1 System Overview

CourtSense is implemented as a web-based application that accepts legal content in the form of PDF files, direct text input, or web URLs. Once the input is provided, the system extracts and preprocesses the text before passing it through transformer-based models for analysis. The final output includes a summarized version of the document, predicted legal outcome, confidence score, key legal issues, jurisdiction details, and related case suggestions.

3.2 Document Acquisition and Processing

The system supports multiple input formats to improve usability and flexibility. Legal documents uploaded in PDF format are processed using text extraction libraries, while web-based legal content is collected using HTML parsing techniques. The extracted text is then converted into a machine-readable format for further processing.
To improve the quality of extracted text, unnecessary patterns such as page numbers, repeated headings, formatting symbols, and unwanted spaces are removed. This preprocessing stage helps in reducing noise and improving model performance.

3.3 Text Processing and Analysis

After document extraction, the legal text undergoes several preprocessing operations to improve data quality and prepare it for deep learning models. The preprocessing stage includes tokenization, normalization, removal of irrelevant content, and elimination of duplicate text patterns. Unnecessary formatting elements such as repeated headers, page numbers, and unwanted symbols are removed to maintain consistency across different legal documents.

Since legal documents are generally lengthy and transformer-based models such as T5 and LegalBERT support limited input tokens, the proposed system uses a chunk-based processing mechanism similar to the sliding window approach. In this method, the document is divided into smaller text segments, and each segment is processed independently. The outputs generated from these chunks are later combined to preserve contextual continuity and improve analysis quality.

The processed text is then passed to the T5 transformer model for legal document summarization. The model generates concise summaries for individual chunks, which are later combined into a final coherent summary representing the complete legal document. This helps reduce manual reading effort while preserving important legal information.

For legal outcome prediction, the summarized text is analyzed using a fine-tuned LegalBERT model trained on legal datasets. The model predicts the possible legal outcome, such as “violation” or “no violation,” along with a confidence score. Using domain-specific transformer models improves contextual understanding and enhances prediction accuracy in legal analysis tasks.

The system also includes a similar case retrieval mechanism to support legal research. The summarized legal content is used to identify related legal cases through a search-based retrieval approach. Future improvements may include the use of dense vector embeddings and semantic similarity techniques for more context-aware retrieval.

In addition, the proposed framework extracts important legal insights such as key legal issues, jurisdiction details, and relevant legal terms. These extracted insights help users quickly understand the nature of the legal document without manually reviewing the complete text.

The performance of the proposed system is evaluated using standard NLP evaluation metrics. Summarization quality is measured using ROUGE-1, ROUGE-2, ROUGE-L, and METEOR scores, while prediction performance is evaluated using Precision, Recall, and F1-Score. These metrics help assess the effectiveness of both summarization and legal outcome prediction modules.

4. System Architecture

The system architecture of CourtSense for Case Analysis follows a multi-stage workflow for analyzing legal documents using Natural Language Processing (NLP) and deep learning techniques. The process begins with multiple input sources such as PDF documents, web URLs, or direct text input. The extracted text then undergoes preprocessing, where unnecessary content, formatting noise, and duplicate patterns are removed. Since legal documents are usually lengthy, the system applies a chunk-based processing approach similar to the sliding window technique to divide large documents into smaller segments. These processed chunks are passed to the T5 transformer model for generating concise summaries. The summarized content is further analyzed using a fine-tuned LegalBERT model to predict legal outcomes along with confidence scores. The system also performs similar case retrieval to identify relevant legal precedents and extracts important legal insights such as jurisdiction and key issues. Finally, the generated results, including summaries, predictions, and related cases, are presented through a user-friendly interface for efficient legal analysis and decision-making.

5. IMPLEMENTATION AND RESULTS

The proposed system, **CourtSense for Case Analysis**, was implemented as a web-based application using Python and Flask. The system integrates transformer-based deep learning models for performing legal document summarization and legal outcome prediction. PyTorch and Hugging Face transformer libraries were used for loading and processing the T5 and LegalBERT models.

The application accepts legal documents in the form of PDF files, direct text input, and web URLs. Text extraction from PDF documents is performed using document parsing libraries, while web content is processed using HTML parsing techniques. The extracted legal text is then cleaned and preprocessed before being passed to the analysis modules.

For summarization, the T5 transformer model generates concise summaries from lengthy legal documents. Since legal documents often exceed token limitations, the system processes the text using a chunk-based approach, where large documents are divided into smaller segments before summarization. The generated summaries are then combined into a final coherent summary.

The summarized content is further analyzed using a fine-tuned LegalBERT model to predict legal outcomes such as "violation" or "no violation." The model also produces a confidence score representing the probability of the predicted result. In addition, the system retrieves similar legal cases and extracts key legal insights such as jurisdiction and legal issues.

The developed system provides a user-friendly interface for uploading legal documents and viewing analysis results. Experimental evaluation of the system can be performed using metrics such as ROUGE and METEOR for summarization quality, and Precision, Recall, and F1-Score for prediction performance. The obtained results demonstrate that transformer-based models can effectively assist in automating legal document analysis and reducing manual effort in legal workflows.

6. CONCLUSION

This paper presented **CourtSense for Case Analysis**; a legal document analysis system developed using Natural Language Processing (NLP) and deep learning techniques. The system utilizes transformer-based models such as T5 and LegalBERT for legal document summarization, outcome prediction, and similar case retrieval. A chunk-based processing approach is used to handle lengthy legal documents efficiently. The proposed system helps reduce manual effort, improves legal document analysis, and provides a scalable solution for intelligent legal assistance and decision-making.

REFERENCES

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

[2] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, "LEGAL-BERT: The Muppets Straight Out of Law School," Findings of EMNLP, 2020, pp. 2898-2904.

[3] C. Raffel, N. Shazeer, A. Roberts, et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, vol. 21, no. 140, 2020, pp. 1-67.

[4] T. Wolf, L. Debut, V. Sanh, et al., "Transformers: State-of-the-Art Natural Language Processing," Proceedings of EMNLP: System Demonstrations, 2020, pp. 38-45.

[5] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 5998-6008.