International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 5, Sep-Oct 2025
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Automated Legal Document Analysis Platform

Lubna Kausar Inamdar*, Prof. Afreen Mujawar**
*(Department of Computer Science and Engineering, SIET, Vijayapura
Email: lubsinamdar@gmail.com)
** (Department of Computer Science and Engineering, SIET, Vijayapura
Email: mujawarafreen19@gmail.com)

----------**********----------

Abstract:

The Legal-AI Project presents an intelligent platform for automated legal document analysis, leveraging Natural Language Processing (NLP) and Machine Learning (ML) models to assist in the review, classification, and comprehension of legal texts. The system integrates clause extraction, question answering, and paraphrasing modules using transformer-based models such as RoBERTa and T5. It simplifies complex contractual language and provides sentiment-based risk assessment. The architecture includes a Flask backend, a Next.js frontend, and containerized deployment through Docker. The platform enhances efficiency, reduces human error, and democratizes access to legal understanding. This paper outlines the methodology, implementation, and performance outcomes of the Legal-AI platform.

Keywords — **Legal AI, NLP, RoBERTa, T5, Flask, Contract Analysis, Machine Learning.**

----------**********----------

I. INTRODUCTION

Legal documents such as contracts, agreements, and case laws are complex, making manual review both time-consuming and error-prone. With the rapid advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP), automation in legal text analysis has become feasible. The proposed system, the Automated Legal Document Analysis Platform, addresses the inefficiencies of manual contract review through intelligent models capable of clause classification, paraphrasing, and sentiment analysis.

One of the key motivations behind this project is to bridge the gap between legal expertise and accessible technology. While legal professionals are highly skilled at interpreting law, they often spend a significant portion of their time on repetitive and routine tasks such as identifying standard clauses or checking for missing provisions. Automating such processes allows lawyers to focus more on strategic analysis and client advisory roles, thereby enhancing efficiency.

II. PROBLEM STATEMENT AND OBJECTIVES

A. Problem Statement

Legal documents such as contracts, agreements, and policies are often lengthy, complex, and filled with jargon, making them difficult to review and understand for both professionals and non-experts. Manual review of these documents is time-consuming, error-prone, and costly, especially when it involves identifying critical clauses, assessing risks, or simplifying legal language for stakeholders. Traditional contract review processes do not scale well for organizations that must analyze large volumes of legal documents, leading to inefficiencies, delays in decision-making, and potential oversight of important clauses.

B. Objectives

The primary objectives are:

- To allow users to ask questions and get instant answers from contract documents using a legal QA model.
- To enhance comprehension of legal clauses through automated paraphrasing.

- To offer a User-Friendly Interface for smooth user interaction.
- To offer insights into the implications of contract clauses via sentiment and risk analysis, aiding informed decision-making.

III. LITERATURE REVIEW

Shelar (2024) explores the application of machine learning (ML) and deep learning (DL) models in the legal domain. The study highlights that ML/DL techniques significantly outperform traditional rule-based systems in tasks such as **legal document classification**, **judgment prediction**, and summarization.

Usman and Abutalha (2024) propose an innovative framework that integrates **Natural Language Processing (NLP) with blockchain technology for smart contract generation**. Their research focuses on automating the conversion of natural language contracts into executable smart contracts.

D. Hendrycks et al. (2021) introduced the **Contract Understanding Atticus Dataset (CUAD)**, which consists of more than 500 contracts annotated for 41 types of legally significant clauses [1]. Transformer-based models such as BERT, RoBERTa, and ALBERT were tested, and the overall system demonstrated strong potential for automated clause extraction.

I. Chalkidis et al. (2020) developed **LEGAL-BERT**, a family of BERT-based models pretrained on a large collection of legal documents including contracts, case law, and legislation [2]. LEGAL-BERT significantly outperformed generic BERT models in tasks such as contract classification and clause detection.

P. Rajpurkar et al. (2016) introduced the **Stanford Question Answering Dataset (SQuAD)**, a benchmark dataset for extractive machine reading comprehension [4]. This work laid the foundation for question answering approaches adapted in legal document analysis.

C. Raffel et al. (2020) presented the **Text-to-Text Transfer Transformer (T5)**, a unified framework that reformulates all NLP tasks into text-to-text problems [5]. T5 showed superior performance in tasks like summarization and paraphrasing, inspiring the paraphrasing module in this project.

IV. SYSTEM ANALYSIS

A. Existing System

Traditional approaches to legal document review and contract analysis rely heavily on manual processes performed by lawyers and contract managers. This involves carefully reading each section of a contract, identifying key clauses, and assessing risks or obligations. While this ensures accuracy, it is extremely time-consuming, labor-intensive, and costly, especially when dealing with large volumes of contracts.

B. Proposed System

1. Document Acquisition and Preprocessing: Legal documents in PDF/DOCX format are converted to plain text using tools like PyPDF2. The text is then normalized, segmented into clauses, and processed with Named Entity Recognition (NER) to extract key entities.
2. Clause Identification and Classification: Transformer-based models like LEGAL-BERT or RoBERTa, fine-tuned on datasets like CUAD and LEDGAR, are used to classify clauses into predefined categories (e.g., liability, dispute resolution).
3. Question Answering (QA) Module: An extractive QA model, inspired by SQuAD and implemented using a RoBERTa model fine-tuned on CUAD, allows users to ask natural-language questions and receive precise answers extracted from the document.
4. Paraphrasing and Simplification: A T5 transformer model is used to paraphrase complex legal sentences into plain, understandable English without altering the original meaning.
5. Sentiment and Risk Analysis: The TextBlob library is used to perform sentiment analysis on clauses, classifying them as positive, neutral, or negative to provide a preliminary risk assessment.
6. System Architecture: The system follows a client-server architecture with a Next.js frontend for the user interface and a Flask backend that orchestrates all the ML models.

The application is containerized using Docker for portability.

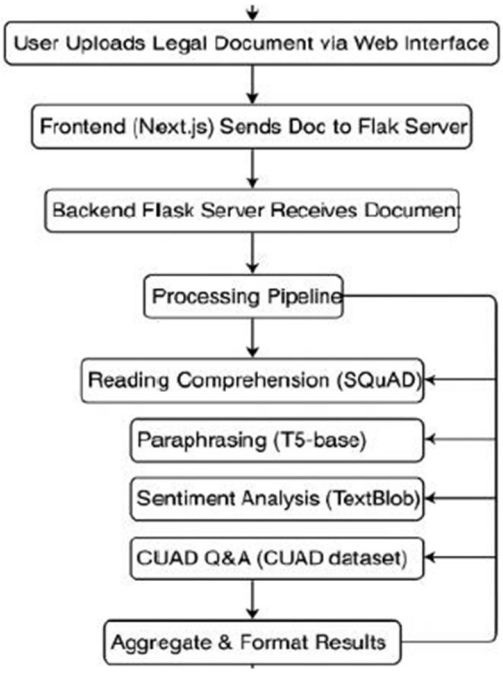

Fig. 1 Flowchart of proposed model

As illustrated in the system workflow of Fig. 1, the process begins when a user uploads a legal document through the web interface. The frontend, built with Next.js, then transmits the document to the backend Flask server. Upon receipt, the server initiates a parallel processing pipeline where the document undergoes multiple AI-driven analyses simultaneously. This includes reading comprehension using a SQuAD-based model for general understanding, paraphrasing via a T5-base model to simplify legal jargon, sentiment analysis with TextBlob to assess clause tone and risk, and specialized legal question-answering powered by a CUAD-trained model for precise clause extraction. Finally, the results from all these modules are aggregated and formatted by the backend before being sent back to the frontend for presentation to the user.

IV. RESULTS AND DISCUSSIONS

The Legal AI platform presents users with a modern and professional interface that immediately establishes trust and credibility. Its design communicates the platform's objectives clearly, with the introductory message emphasizing innovation and collaboration. A visual element depicting a human and a robot shaking hands reinforces the idea that the system is meant to act as an assistant rather than a replacement, symbolizing cooperation between legal professionals and artificial intelligence. This thoughtful introduction helps build confidence and creates a positive first impression, which is particularly critical in the legal domain where reliability, professionalism, and trustworthiness are paramount.

The platform also provides a dedicated section that outlines its key features in a clear and structured way. By highlighting the core modules—clause classification, question answering, paraphrasing, and risk detection users are guided to understand the breadth of functionalities available. Presenting these capabilities in one place not only improves usability for newcomers but also positions the system as a comprehensive solution for legal document analysis, rather than being limited to a single task like question answering. This design choice demonstrates the platform's versatility and communicates its value as an end-to-end legal assistant.

When it comes to system outputs, the platform demonstrates a strong focus on transparency and reliability. In scenarios where the uploaded contract does not contain the requested information, or where the model is unable to locate it, the system clearly states that no answer is available. This approach, rather than producing potentially misleading results, builds user confidence. In legal applications, it is far better to acknowledge uncertainty than to provide inaccurate outputs, and this design choice reflects a commitment to maintaining professional standards.

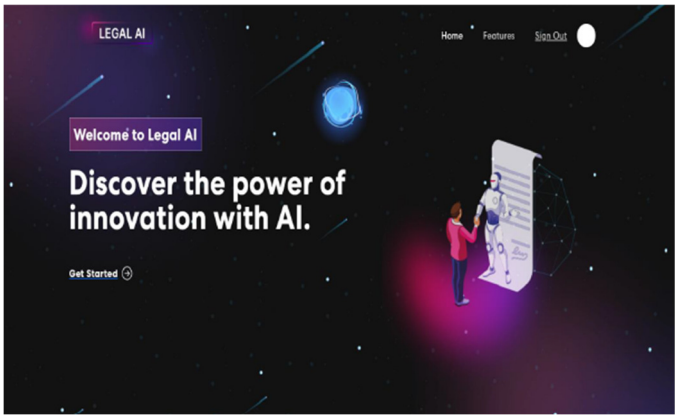

Fig. 2 Welcome screen of platform

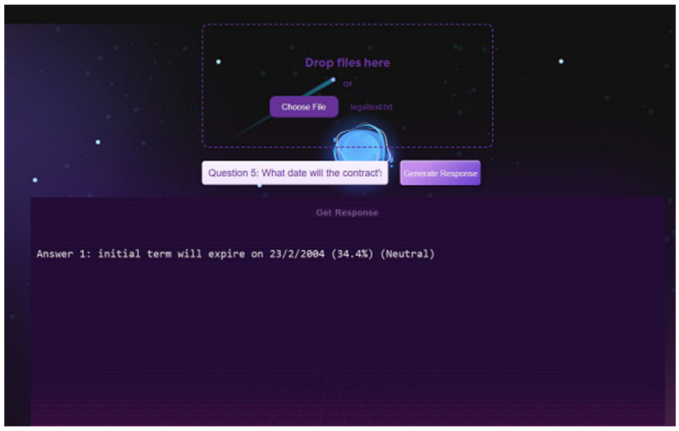

Fig. 3 Output generated by platform

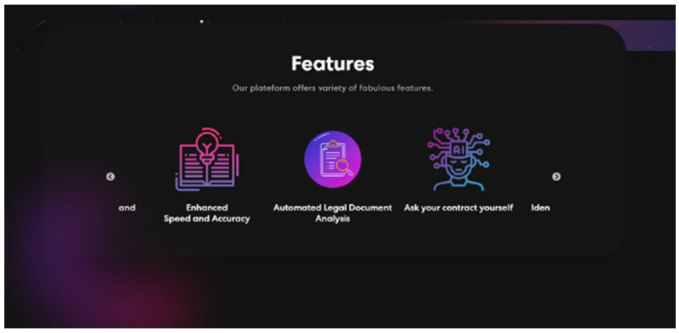

Fig. 4 Features of platform

VI. CONCLUSIONS

The Legal-AI Project demonstrates the significant potential of artificial intelligence to transform legal document analysis. By integrating advanced NLP techniques into an accessible platform, it provides a practical solution that enhances efficiency, promotes accessibility, and supports better decision-making for both legal professionals and non-experts. The project successfully bridges a gap between complex legal knowledge and everyday users, marking a step toward democratizing access to legal insights.

ACKNOWLEDGMENT

I would like to express my profound gratitude and sincere appreciation to all the individuals who have supported and guided me throughout the journey of this project, "Automated Legal Document Analysis Platform."

First and foremost, I wish to extend my deepest thanks to my guide, **Prof. Afreen Mujawar**, Department of Computer Science and Engineering, SECAB Institute of Engineering and Technology, Vijayapura. Her unwavering guidance, invaluable feedback, and constant encouragement.

I am also profoundly grateful to **Dr. S.A. Quadri**, Head of the Department of Computer Network Engineering, His leadership and support for academic pursuits are highly appreciated.

REFERENCES

[1] D. Hendrycks, C. Burns, A. Chen, and S. Ball, "CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review," in *Proc. NeurIPS Datasets and Benchmarks*, 2021.
[2] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and N. Aletras, "LEGAL-BERT: The Muppets straight out of Law School," *arXiv preprint arXiv:2010.02559*, 2020.
[3] D. Tuggener, P. von Däniken, T. Peetz, and M. Cieliebak, "LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts," in *Proc. LREC*, 2020, pp. 1235-1241.
[4] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proc. EMNLP*, 2016, pp. 2383-2392.
[5] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1-67, 2020..
[6] Rabelo, J., Goebel, R., & Da Silva, D. (2022). Legal Information Retrieval and Entailment Based on BM25, Transformer and Semantic Thesaurus Methods. Artificial Intelligence and Law.

