© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1713158

An AI-Based Smart Legal Assistant for Automated Legal Document Analysis

RUMMANA FIRDAUS[1], SUNIDHI S BABU[2], SNEHA NAYAK M S[3], MANVITHA P[4]
[1]*Assistant Professor, Department of Computer Science & Engineering,*
GSSS Institute of Engineering & Technology for Women, Mysuru, Affiliated to VTU, Belagavi
[2, 3, 4]*Student, Department of Computer Science & Engineering,*
GSSS Institute of Engineering & Technology for Women, Mysuru, Affiliated to VTU, Belagavi

Abstract- ***Understanding legal paperworks are a challenge as it's filled with hard-to-understand legal jargon and long sentences. Going through these papers by hand takes a lot of time and can lead to mixed results. Within this document is introduced the Smart Legal Assistant a system rooted in AI geared for the automation of legal document scrutiny through the avenues of extracting clauses, sorting them semantically, and evaluating risks. In practice the assistant is capable of taking both PDF and DOCX files breaking them down into segmented clauses by abiding to a set of predetermined rules and then goes on to classify each clause utilizing derivations from the models from transformer-based natural language processing techniques. Clauses are given a risk severity rating to point out terms that could be harmful or critical. FastAPI used in the backend for development and use of React in the frontend to make a review interface that interact with users. Exploring the output of the suggested setup it was observed that it notably slashes the time spent on going through documents and steps up the quality of how clear and consistent they are structured; all of these together these characteristics have made it a fitting choice for both academic work and the initial stages of legal reviews.***

Index Terms- ***Legal Document Analysis, Clause Classification, Transformer Models, Risk Assessment, Contract Review***

I. INTRODUCTION

The rapid digitalization of legal workflows across organizations, enterprises, and regulatory environments has significantly increased the volume and complexity of electronic legal documents. Contracts, policies, compliance agreements, and consent forms are now created, exchanged, and stored primarily in digital formats. These documents are typically long, structurally dense, and written in formal legal language that includes layered conditions, exceptions, cross-references, and obligations distributed across multiple sections. As a result, understanding the full legal impact of a document often requires careful interpretation of several interdependent parts rather than reading isolated paragraphs.

Despite this shift toward digital documentation, legal review practices remain largely manual. Legal professionals and reviewers usually rely on traditional document readers to scroll through contracts line by line, repeatedly revisiting sections to understand intent and implications. This process is time-consuming, cognitively demanding, and highly dependent on individual experience and attention. When document volumes increase or deadlines become tight, the likelihood of inconsistent interpretation, missed obligations, or overlooked risks also increases. Existing document-handling tools mainly provide basic reading, annotation, and keyword search capabilities, that are enough not for capturing deeper legal meaning or structural dependencies within documents.

Recent research in legal natural language processing (NLP) has shown that accurate legal interpretation cannot be achieved through keyword- based or document-level analysis alone [1][2]. Legal meaning is often expressed at finer structural levels, where obligations, rights, and conditions are embedded within some segments of text and connected through long-range dependencies. Benchmark studies such as CUAD, ContractNLI, and LexGLUE highlight that legal understanding tasks require semantic reasoning over these structured units rather than shallow text matching [3][4][5]. These findings indicate that legal analysis systems must move beyond surface-level processing to capture contextual and semantic intent.

In response to these insights, several AI-based legal analysis tools have emerged in recent years. While these systems aim to support contract review, many still face limitations related to reliable text segmentation, consistency in interpretation, and standardization of risk assessment outputs [6][7]. Inconsistent segmentation can lead to misclassification, and varying scoring mechanisms often make it difficult to compare risk levels across documents. Additionally, many tools store results externally or present them in unstructured formats, requiring reviewers to manually consolidate findings during reporting or auditing stages.

Challenges to address, this paper presents Smart Legal Assistant, an AI-driven system designed to support automated legal document summarization and risk assessment through structured semantic analysis. The system follows a systematic processing approach in which legal documents are transformed from unstructured text into organized, machine-interpretable representations. A domain-trained transformer model is used to analyze legal semantics and interpret intent, while a structured severity scoring mechanism prioritizes high-risk portions of the document. This combination enables faster, more consistent review while reducing reliance on repetitive manual reading.

By converting complicated legal text into structured analytical outputs, the proposed system reduces manual effort, improves consistency in interpretation, and supports scalable analysis across large collections of legal documents. The design emphasizes practical usability, interpretability, and extensibility, making it suitable for real-world legal workflows. Overall, this work explains how domain- aware AI models combined with structured processing pipelines can enhance legal document understanding and provide a reliable foundation for modern legal analysis systems [8][9].

II. MOTIVATION FOR THE PAPER

The motivation for this research is the increasing reliance on digital legal documents and the practical challenges involved in reviewing them manually. Legal contracts often contain complex clause structures, conditional statements, and domain-specific terminology that make accurate interpretation difficult, particularly for non-legal professionals. Identifying critical clauses related to obligations, penalties, and termination conditions requires significant time and expertise, which is not always readily available.

To explore potential solutions, a structured review of existing research in legal natural language processing was conducted. Prior studies on contract analysis, clause classification, and transformer-based language models were examined to understand how semantic understanding of legal text has evolved. In addition, existing legal document analysis tools were analyzed to identify their capabilities and limitations. This research that showed many current approaches focus on document-level processing or keyword- based methods, offering limited support for fine- grained clause interpretation.

Based on these findings, the research direction was refined to emphasize clause-level semantic analysis combined with risk identification. The collected insights provided a foundation for designing a system that leverages advanced language models to improve the efficiency, consistency, and interpretability of legal document review.

III. LITERATURE SURVEY

[1] LEGAL-BERT: The Muppets Straight Out of Law School – Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos (2020).
This work introduced LEGAL-BERT, a transformer-based language model pretrained exclusively on legal-domain corpora such as contracts, statutes, case law, and legal opinions. The authors demonstrated that domain-adaptive pretraining significantly improves performance in legal clause classification,

legal entailment, and semantic understanding tasks when compared to generic BERT models. However, the model retains the standard transformer token-length limitation, restricting its impacts for very long legal documents without clause-level segmentation.
[2] CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review – Dann Hendrycks, Collin Burns, Anya Chen, and Spencer Ball (2021).

This paper presented the Contract Understanding Atticus Dataset (CUAD), consisting of real-world commercial contracts annotated by legal experts across multiple critical clause categories. The study exposed major challenges in contract analysis, including ambiguous clause boundaries, extensive cross-referencing, and diverse drafting styles. While transformer models achieved reasonable accuracy on frequently occurring clause types, the authors observed reduced reliability for rare and complex clauses, highlighting limitations in current automated contract review systems.

[3] LexGLUE: Legal General Languages Understanding Evaluation – Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikollaos Aletras (2021).

LexGLUE introduced a comprehensive points for legal language evluating understanding across multiple tasks such as legal text classification, legal inference, and contract analysis. The benchmark revealed that even advanced transformer models struggle with deep legal reasoning, contradiction detection, and interpretation of context-sensitive clauses. The findings emphasize that legal document analysis requires structured reasoning beyond surface-level pattern recognition.

[4] ContractNLI: Document-Level Natural Language Inference for Contracts – Yuta Koreeda and Christopher D. Manning (2021).

This study reformulated contract interpretation as a document-level natural language inference task. The authors showed that determining contractual obligations often requires reasoning across multiple clauses and sections rather than isolated sentence analysis. Their results demonstrated that keyword-

based and sentence-level models fail to capture contractual intent effectively, underscoring the need for clause-aware and context-sensitive legal analysis frameworks.

[5] LayoutLM: Pretraining of Text and Layout for Documents Images Understanding – Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhouu (2019).

LayoutLM proposed a document understanding model that combines textual content with two- dimensional layout information derived from document structure. The model achieved strong results in structured document analysis tasks, particularly for contracts containing tables, headings, and multi-column layouts. However, its dependence on OCR accuracy limits performance when dealing with low-quality or scanned legal documents.

[6] LayoutLMv2: Multi-Modal Pretraining for Documents Understanding – Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furru Wei, Guoxin Wang, and Yijuan Lu (2020).

LayoutLMv2 extended the original model by integrating textual, layout, and visual features within a unified transformer architecture. The authors demonstrated improved robustness and accuracy in document segmentation and information extraction tasks. Despite its effectiveness, the model requires higher computational resources, making deployment more challenging in constrained environments.

[7] Longformer: The Long-Document Transformer – Iz Beltagy, Matthew E. Peters, and Arman Cohan (2020).

This paper introduced Longformer, a transformer architecture designed to long documents process efficiently using sparse attention mechanisms. The model enables contextual understanding across extended legal texts, making it suitable for analyzing long contracts with cross-referenced clauses. However, careful configuration is required to balance performance and computational cost.

[8] BigBird: Transformers for Longer Sequences Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chriss Alberti, Santiago

Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed (2020).

BigBird proposed a sparse attention-based transformer capable of handling very long sequences while maintaining theoretical expressiveness. The authors demonstrated its effectiveness in long- document classification and reasoning tasks. Although well-suited for legal documents, the model's complexity and resource requirements pose challenges for lightweight deployment.

IV. SURVEY FINDINGS

The literature survey conducted across multiple legal natural language processing studies and legal document analysis systems reveals a clear evolution from traditional keyword-based document handling

toward clause-centric, semantically driven analysis frameworks. Early legal document processing tools focused mainly on text retrieval and keyword highlighting, which were found to be insufficient for accurately interpreting legal intent, obligations, and conditional dependencies embedded within contracts. The surveyed works consistently emphasize that legal meaning is distributed across individual clauses rather than entire documents, making clause-level processing a fundamental requirement for reliable legal analysis.
A key finding from benchmark datasets and evaluation frameworks is that accurate clause segmentation plays a cruciapl role in downstream legal interpretation tasks. Studies highlight that incorrect or inconsistent clause boundary detection significantly degrades classification accuracy and risk interpretation, even when advanced language models are applied. This observation establishes clause segmentation as a foundational step that must be handled with precision before semantic analysis can be effective.

The survey further indicates that general-purpose language models struggle with legal terminology, structured drafting styles, and long-range dependencies commonly found in contracts. Domain-adapted transformer models trained on legal corpora demonstrate superior performance in clause intent

classification, legal entailment reasoning, and obligation detection. These findings reinforce the importance of using legal-domain-trained models rather than relying on generic NLP approaches for contract analysis.

Another important observation from the surveyed literature is the lack of standardized risk assessment and severity tagging mechanisms in many existing legal analysis tools. While several systems provide clause classification or summarization, few offer consistent, interpretable, and reusable severity scoring that supports downstream auditing or report generation. The absence of structured storage and uniform severity logic often forces reviewers to manually consolidate findings, increasing effort and reducing reliability.

Overall, the survey findings conclude that an effective legal document analysis system must integrate accurate clause segmentation, domain- aware semantic classification, consistent risk severity assessment, and structured data storage. These insights directly inform the design of the proposed system, motivating the adoption of a clause-centric pipeline, transformer-based legal semantics, rule- driven severity scoring, and database-backed clause intelligence storage. By addressing the gaps identified in existing research and tools, the proposed approach aligns with current advancements while offering a more practical and scalable solution for real-world legal document review.

V. METHODOLOGY

The methodology adopted for developing the Smart Legal Assistant follows a structured, clause- centric approach for automated legal document analysis. The complete workflow is divided into sequential stages to ensure accurate extraction, interpretation, and assessment of legal clauses.

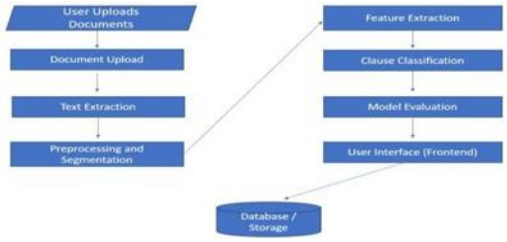

Figure 1: Architecture of the System

- User Authentication and Access Control: The system begins with secure user authentication to ensure controlled access to legal documents. User registration and login are validated using OTP-based verification, preventing unauthorized usage and maintaining confidentiality of uploaded legal files.
- Document Upload and Validation: Users upload legal documents in supported patterns such as PDF and DOCX. The system validates file type and size before processing, and rejects unsupported or corrupted files to prevent processing errors.
- Document Parsing and Text Extraction: Uploaded documents are converted into machine-readable text using format-specific parsing libraries. PDF files are processed with PyMuPDF, while DOCX documents are handled using python-docx to ensure accurate extraction of raw legal text and preserve structural consistency.

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1713158

– Clause Detection and Segmentation: The extracted text is segmented into individual clauses using a rule- based boundary detection mechanism. Numbering formats, section headers, bullet points, and paragraph structures are analyzed to identify clause start and end positions, enabling fine-grained legal analysis.

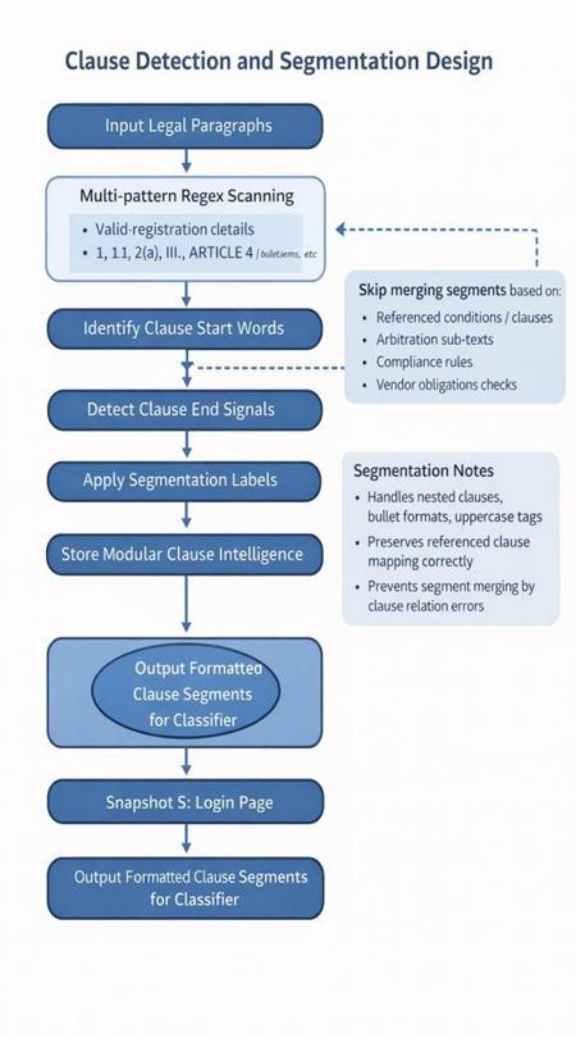

Figure 2: Clause Detection and Segmentation Design Flow

– Clause Intent Classification: Each extracted clause is analyzed using a transformer-based classification model trained on legal-domain data. The model assigns intent labels such as liability, payment, termination, confidentiality, or jurisdiction based on semantic interpretation rather than keyword matching.

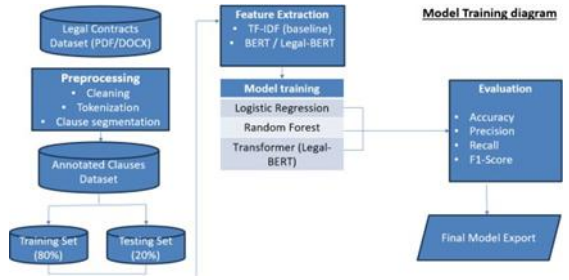

Figure 3: Model Training Workflow for Clause Intent Classification

– Risk Severity Scoring: A rule-based severity scoring engine evaluates classified clauses to determine potential legal risk. Based on predefined logic and clause characteristics, each clause is assigned a severity level, allowing prioritisation of critical and high-risk clauses.

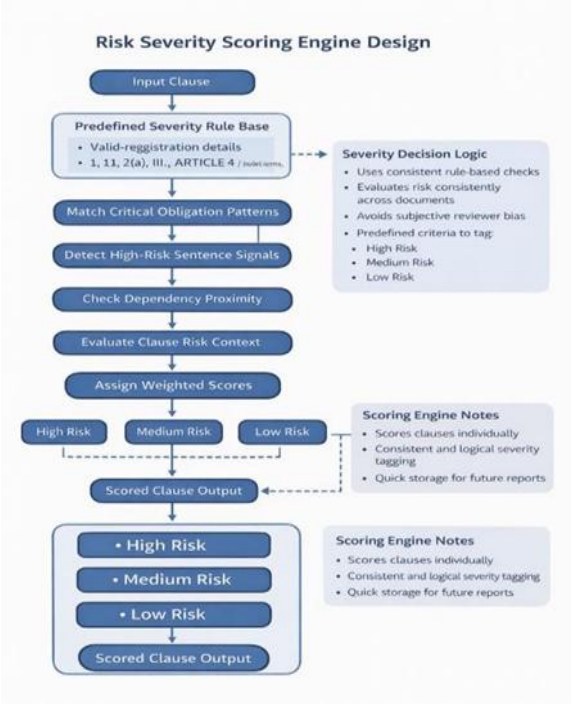

Figure 4: Risk Severity Scoring Engine Design Flow

– Structured Data Storage: All analysed clause data, including clause text, intent label, severity score, and document metadata, is stored in an SQLite database. This structured storage ensures efficient retrieval, reuse of analysis results, and consistency across multiple document reviews.
– Backend Processing and API Management: The backend system is implemented using FastAPI and served through Uvicorn. It manages document

processing requests, model inference operations, database interactions, and secure data communication between system components.

– Frontend Visualisation and User Interface: Analysis results are presented through an interactive web interface developed using React. It displays clause- wise classifications, severity indicators, and analysis summaries, allowing users to review findings without repeatedly scanning the original document.
– Result Consolidation and Review Support: The system provides organised clause-level intelligence that supports faster report generation, audit preparation, and compliance review. By storing reusable analysis outputs, the methodology reduces manual work and improves review efficiency.

VI. CONCLUSION AND FUTURE ENHANCEMENT

The Smart Legal Assistant system demonstrates an effective and reliable clause-based approach for automated legal document review and risk assessment. By integrating rule-driven clause segmentation, transformer-based intent classification, and structured severity scoring, the system successfully addresses major challenges seen in manual reviewing such as repetitive reading, clause boundary confusion, and inconsistent interpretation. The project also establishes a stable backend pipeline using Python, PyMuPDF, python-docx, HuggingFace Transformers, and PyTorch for accurate semantic analysis, while the React-based frontend delivers clear visual outputs that simplify document navigation and contract comprehension. Secure OTP authentication ensures controlled usage, and the SQLite database enables reusable clause intelligence storage, making the solution scalable and dependable for real-world legal environments.

Looking ahead, the system can be enhanced to support more advanced analytical features and wider deployment capabilities. Future improvements may include further training of the transformer model using larger and diverse legal corpora to achieve stronger clause intent accuracy across varying drafting styles. Clause comparison modules, batch document processing, and severity-based heat-map visualizations can be incorporated to support large-scale legal review workflows. The platform can also be extended to mobile interfaces for portable contract analysis, along with improved dashboard responsiveness and document history management. Security and performance can be strengthened by enabling encrypted database storage, advanced authentication layers, multi-user collaboration support, and faster text extraction pipelines.

Overall, the system lays a strong foundation for AI-driven legal document analysis and holds significant potential to evolve into a comprehensive automated contract management solution capable of supporting enterprise-level review, compliance assistance, and legal decision support.

ACKNOWLEDGMENT

The authors express sincere gratitude to the Department of Computer Science & Engineering, GSSS Institute of Engineering & Technology for Women, Mysuru, and the project guide for guidance and support.

REFERENCES

[1] Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. "LEGAL-BERT: The muppets straight out of law school." arXiv preprint arXiv:2010.02559 (2020).

[2] Hendrycks, Dan, Collin Burns, Anya Chen, and Spencer Ball. "Cuad: An expert-annotated nlp dataset for legal contract review." arXiv preprint arXiv:2103.06268 (2021)..

[3] Chalkidis, Ilias, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. "LexGLUE: A benchmark dataset for legal language understanding in English." In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4310-4330. 2022.

[4] Koreeda, Yuta, and Christopher D. Manning. "ContractNLI: A dataset for document-level natural language inference for contracts." arXiv preprint arXiv:2110.01799 (2021).

[5] Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter Liu. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." In International conference on machine learning, pp. 11328-11339. PMLR, 2020.

[6] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

[7] Xu, Y., M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou. "LayoutLM: pre-training of text and layout for document image understanding. ArXiv." arXiv preprint arXiv:1912.13318 (2019).

[8] Xu, Yang, Yiheng Xuu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu et al. "Layoutlmv2: Multi-modal pre-training for visually-rich document understanding." In Proceeding of the 59th Annual Meeting of the Association for Computational Linguistic and the International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 2579-2591. 2021.

[9] Stanisławek, Tomasz, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. "Kleister: key information extraction datasets involving long documents with complex layouts." In International Conference on Document Analysis and Recognition, pp. 564-579. Cham: Springer International Publishing, 2021.

[10] Garncarek, Łukasz, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. "Lambert: Layout-aware language modeling for information extraction." In International conference on document analysis and recognition, pp. 532-547. Cham: Springer International Publishing, 2021.

[11] Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150 (2020).

[12] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

[13] Zaheer, Manzil, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham et al. "Big bird: Transformers for longer sequences." Advances in neural information processing systems 33 (2020): 17283-17297.

[14] Mathew, Minesh, Ruben Tito, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. "Document visual question answering challenge 2020." arXiv preprint arXiv:2008.08899 (2020).

[15] Krpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. "Dense Passage Retrieval for Open-Domain Question Answering." In EMNLP (1), pp. 6769-6781. 2020.