Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20260501041
Volume 5, Issue1 (January-April 2026), PP: 293-300.
www.indjcst.com
ISSN No: 2583-5300

An Intelligent Framework for Financial Contract Risk Analysis using Transformer-Based Clause Modelling and Graph Neural Networks

Alok Kushwaha[1], Dr. Mohammed Abuzar A[2]
[1]M.Sc. Data Science, SIES College of Arts, Science and Commerce, (Empowered Autonomous), Mumbai, India.
[2]Head of Department, Data Science, SIES College of Arts, Science and Commerce (Empowered Autonomous), Mumbai, India.

To Cite this Article: *Alok Kushwaha[1], Dr. Mohammed Abuzar A.[2], "An Intelligent Framework for Financial Contract Risk Analysis using Transformer-Based Clause Modelling and Graph Neural Networks", Indian Journal of Computer Science and Technology, Volume 05, Issue 01 (January-April 2026), PP: 293-300.*

Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution License; Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: *With the increasing number of legal and financial contracts, the analysis of contracts through a manual system of analysis has become inefficient. Hence, the need for an automated system that can not only extract the clauses from the contracts but also interpret them to assess the risks involved with the contracts arises. This paper proposes a machine learning-based system for the analysis of legal contracts through the application of Natural Language Processing techniques. This system can be applied to both scanned documents and digital documents by the application of a hybrid approach to the extraction of the content from the documents. The system will be capable of analysing the content of the documents by the application of a Legal-BERT model that has been fine-tuned on a dataset comprising the General Financial Rules (GFR) and additional legal contract clause datasets. The system will be capable of risk analysis by the application of a risk-aware clause classification system that not only classifies the clauses into various types of laws but also determines the risk level of the contracts by the application of three risk levels: low, medium, and high. To enhance interpretability and robustness, the framework includes a hybrid risk assessment mechanism that leverages the power of the transformer model's predictions and rule-based compliance verification. Moreover, the framework includes a GNN approach to analyse the structural relationships between clauses in the contract document. The experimental results demonstrate that the proposed approach can effectively improve the accuracy and reliability of the automated risk detection process using the domain-specific transformer models and rule-based reasoning approach. The proposed approach can be considered an efficient and intelligent way to analyse contracts and can be effectively utilized by legal professionals and financial analysts.*

Key Words: *Legal Natural Language Processing, Contract Risk Analysis, Legal-BERT, Clause Classification, Graph Neural Networks, Automated Contract Analysis*

I. INTRODUCTION

The rise of the digital age has seen financial, legal, and regulatory documents becoming more digitized. This has, therefore, led to a rise in the quantity of contracts and policy documents. Every organization, irrespective of the industry, has the obligation to analyse legal documents to identify obligations, potential risks, and compliance with regulations. However, the process of manually analysing contracts is time-consuming, expensive, and prone to human error, especially where there is the use of intricate legal jargon and long interdependent clauses. As such, the use of Artificial Intelligence (AI) and Natural Language Processing (NLP) has been seen as an alternative for the efficient analysis of legal documents.

However, the recent advancements in NLP techniques, specifically in transformer-based language models, have shown significant potential in the ability of machines to comprehend domain-specific textual data. For example, the use of pre-trained language models such as Legal BERT has shown promising results in the processing of legal texts using large-scale legal documents for domain adaptation [1]. Moreover, the availability of datasets such as the Contract Understanding Atticus Dataset (CUAD) enables the use of expert-labelled legal contracts for the development of various research studies in clause classification, contract review automation, and risk identification [2].

Several research studies have also examined the effectiveness of machine learning and deep learning methods for the identification of risks and extraction of key clauses from contracts. Transformer models were used to identify risky clauses from construction contracts, showcasing the ability of the model to employ contextual embeddings for the identification of legal obligations and liabilities [3]. Previous research studies have also employed machine learning with the help of heuristic methods

for the extraction of legal clauses from contracts, showcasing the effectiveness of automated contract analysis systems [7]. Further research studies on the survey of legal NLP highlight the need for the development of appropriate datasets for the improvement of performance for legal text understanding tasks [5], [12].

In addition to the above-mentioned transformer models, the application of graph-based methods to the analysis of the relationship between legal entities and clauses within the documents has also received considerable attention. Graph Neural Networks (GNNs) enable the analysis of legal documents as structured graphs, thereby effectively modelling the relationship between the clauses within the documents for the improvement of legal case retrieval tasks and article retrieval tasks [4],[8]. Graph-based learning methods have also been applied to the prediction of legal judgments, thereby effectively modelling the structural relationship between the legal documents [9].Despite such advances, existing systems are largely focused on the analysis of particular, isolated functions such as clause classification or risk detection, rather than the provision of a comprehensive framework for the complete analysis of contracts. Furthermore, the variety of media in which legal documents are represented,such as scanned images of documents and computer-generated documents, is a complicating factor for the analysis of such documents. To address the shortcomings of existing systems, the present research proposes an artificial intelligence-based framework for the automated analysis of legal and financial documents, which incorporates the use of the transformer family of language models, rule-based risk evaluation, and graph-based dependency modelling. A hybrid text extraction method is employed for the analysis of legal documents, along with a fine-tuned Legal-BERT model for the classification of clauses and the detection of potential risks. Furthermore, a graph neural network is employed for the analysis of the relationships between clauses, thereby enabling the more detailed structural analysis of legal documents.

II.LITERATURE REVIEW

A.Legal Natural Language Processing

Natural Language Processing (NLP) has emerged as a promising area for the analysis of legal documents to obtain relevant information from the documents. Legal documents often involve intricately written sentences with unique terminology and long-range dependencies. These complexities make it difficult for rule-based systems to effectively process legal documents. However, recent advancements in the application of deep learning techniques have enabled the effective analysis of legal documents. Surveys on the application of NLP for legal documents emphasize the fact that the application of machine learning techniques for the analysis of legal documents has been significantly effective for tasks like legal document classification, clause extraction, etc. [5], [12].

B.Transformer-Based Models for Legal Text

Transformer-based language models have greatly improved the performance of NLP systems for a particular domain. One of the most prominent language models for a particular domain is Legal-BERT. Legal-BERT is pre-trained on vast amounts of legal texts to effectively comprehend legal terminology and context relationships compared to general language models [1]. With the introduction of the Contract Understanding Atticus Dataset (CUAD) dataset, it is also possible to train a model for a particular task of contract review and clause classification [2]. Recent research demonstrates the effectiveness of the transformer-based model to automatically identify the obligations, liabilities, and risks of a particular legal agreement by learning the context of legal terminology [3].

C.Clause Extraction and Risk Detection

Automatic clause extraction and risk detection have also become prominent research areas for legal document analysis. Earlier research on legal document analysis primarily focused on the application of rule-based systems and heuristic approaches for the identification of specific clauses within contracts. For instance, the application of machine learning with the incorporation of heuristic enhancements has been reported for the extraction of legal clauses as well as the identification of key clauses within contracts [7]. Recently, the application of deep learning approaches for the detection of risky clauses within legal documents, specifically for construction contracts and financial contracts, has also been reported[14]. These research studies indicate the effectiveness of the application of contextual models for the identification of specific patterns related to legal contracts. However, most of the existing research studies primarily focus on the application of the system for specific tasks rather than a comprehensive framework.

D.Graph-Based Legal Document Analysis

In addition to textual analysis, the application of graph-based learning techniques has also been proposed to learn the structural information contained in legal documents. Graph Neural Networks (GNNs) enable the modelling of legal information as a series of nodes and edges. Existing research has shown the effectiveness of the application of graph-based learning techniques to the tasks of legal case retrieval and article retrieval [4],[8]. Graph-based learning techniques have also been applied to the prediction of legal judgments, where the relationship between legal arguments can be used to enhance the accuracy of the prediction [9].

E.Retrieval-Enhanced Legal Analysis

Retrieval-based methods have also been examined to enhance the precision of legal NLP systems. Retrieval-Augmented Generation (RAG) combines external information with language models to enable the system to retrieve relevant information during the prediction process. Studies have demonstrated that retrieval mechanisms enhance the performance of knowledge-intensive tasks by minimizing hallucinations and increasing the accuracy of information [24]. In the context of legal NLP

systems, retrieval mechanisms can be used to retrieve information from a database of legal cases or statutes to increase the accuracy of AI-generated insights.

F.Limitations of Existing Approaches

However, despite the progress achieved in legal NLP, there are still several challenges to overcome. Most of the legal NLP systems developed to date focus on specific tasks like clause classification, legal document retrieval, judgment prediction, etc., but not on providing a framework that can perform legal document analysis. In addition to that, legal documents can appear in several forms, including scanned PDF files and digital files, which can pose several challenges during text processing as well. Surveys conducted on legal NLP studies emphasize the need to develop hybrid systems that can make use of domain-specific language models along with rule-based reasoning to improve accuracy and understanding [5], [12].

III. RESEARCH GAP AND PROBLEM DEFINITION

A. A.Research Gaps in Existing Systems

Despite the various advancements in Artificial Intelligence and Natural Language Processing techniques in the domain of contract analysis, certain limitations still persist in the existing systems. For example, although various techniques like Legal BERT, and transformer-based language model, have been effective in understanding the nuances of the law and classifying the same, it has been observed that these models rely heavily on the availability of large datasets. Another limitation that has been identified in the existing contract analysis systems is that various techniques that have been proposed in the domain of contract analysis are mostly focused on individual aspects of contract analysis, rather than providing an overall system that can perform comprehensive contract analysis.

For example, although datasets like CUAD have been effective in contract review and clause extraction, it has been observed that these datasets do not perform an overall analysis of the contract. Similarly, although techniques like deep learning have been effective in identifying risky clauses in contracts, especially in specialized domains, it has been observed that these techniques perform only textual classification. Graph-based methods were proposed to deal with relationships in legal documents by using clauses and legal entities as nodes in a graph structure [4], [8]. Although these methods help in understanding the relationships between various legal concepts, they are generally used in case retrieval or judgment prediction rather than contract risk analysis. Thus, the integration of text-based risk detection with clause dependency is not well explored. In addition, most of the existing methods consider that the legal documents are available in structured textual form. However, in practice, it is common that the contract is available in various forms like scanned PDFs or images.

B.Problem Definition

Based on the identified gaps in the research, the primary issue that this study seeks to resolve is the absence of an integrated framework that enables the automated analysis of contracts from the time they are extracted from the contract until the time they are evaluated. Most of the automated systems have focused on specific aspects and not the entire process. This research seeks to develop an AI-based framework that includes aspects such as text extraction from documents, clause classification, risk evaluation, and dependency analysis. The proposed framework uses a domain-specific transformer model and incorporates graph-based models.

The challenges that the framework seeks to address include:

1.Text extraction from both digital and scanned legal documents.
2.Identification and classification of contractual clauses.
3.Identification of possible risks using machine learning and rule-based approaches.
4.Representation of the relationships between clauses for better contract understanding.

By resolving these challenges, the proposed framework seeks to enhance the efficiency and accuracy of automated contract analysis systems.

IV. PROPOSED SYSTEM ARCHITECTURE

A.Overview of the Framework

This section describes the architecture of the proposed AI-based framework for legal document analysis. The system architecture integrates various components for document preprocessing, clause classification using the transformer architecture, graph modelling for clause dependencies, and automated risk evaluation for the analysis of legal contracts. The system workflow for the proposed framework is depicted in Fig. 1. The system workflow begins with the upload of the legal contract or financial agreement by the user. The uploaded documents may be in digital or scanned formats. The uploaded documents are fed into the hybrid document extraction module for converting the documents into machine-readable formats using the PDF parsing and Optical Character Recognition (OCR) techniques. The extracted documents are further fed into the clause extraction module for the identification of relevant clauses from the legal documents. For the analysis of the identified clauses, the system utilizes the Legal BERT model, which has been proven effective for various legal document understanding tasks [1]. To identify the dependencies between the clauses, the system generates the clause dependency graph and utilizes the Graph Neural Network (GNN) for the processing of the graph. The graph-based approach has been successful for various legal information retrieval and reasoning tasks [4], [8].

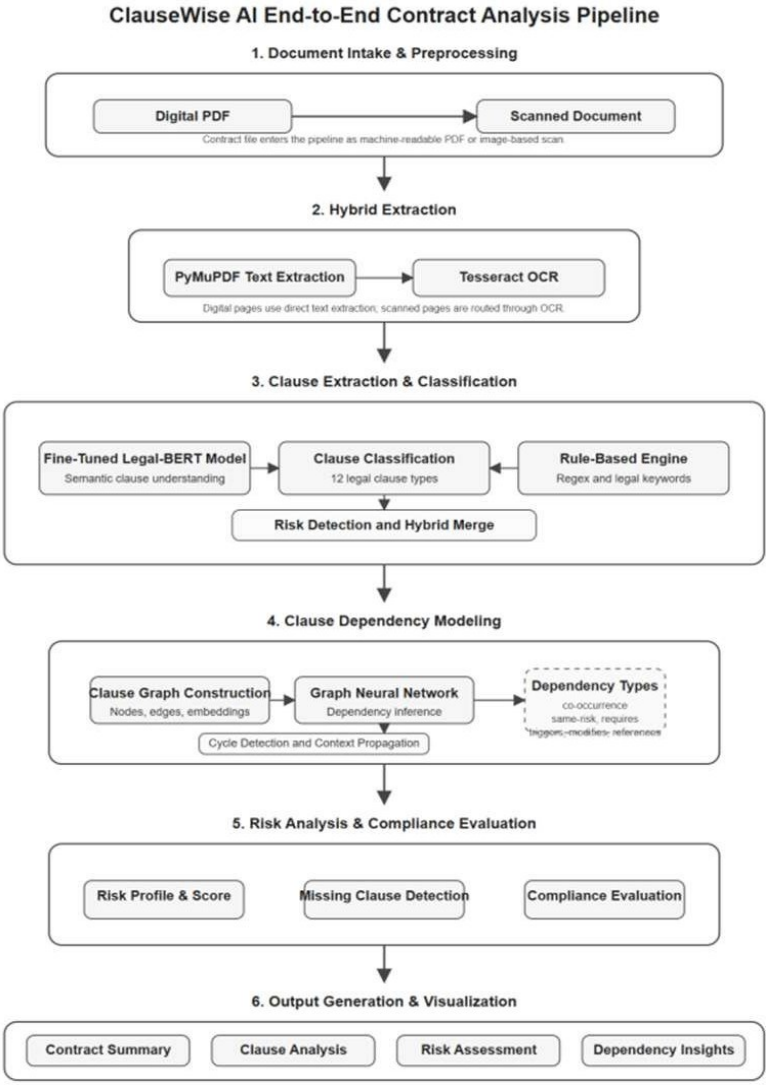

Fig. 1. Overview of the proposed AI-driven legal document analysis framework for automated contract risk assessment.

B.Document Intake and Hybrid Text Extraction

The first stage of the system is document intake and preprocessing. The user is expected to upload legal documents in different formats, including digital PDF formats or scanned images. Legal documents contain unstructured as well as structured data. Therefore, a hybrid extraction mechanism is employed. For digital documents, PyMuPDF is employed to obtain text content directly from the digital documents in PDF format. For scanned documents, Tesseract OCR is employed to obtain text content. The text obtained is then processed using Natural Language Processing techniques. Let's assume that if the extracted document is in the form of a sequence of tokens, then:

$$\boldsymbol{T} = \{\boldsymbol{t1}, \boldsymbol{t2}, \boldsymbol{t3}, \dots, \boldsymbol{tn}\}$$

where $t$irepresents individual tokens and $n$is the total number of tokens in the document.

C.Clause Extraction and Classification

After the pre-processing, the system identifies the important contractual clause types, such as termination, confidentiality, liability, and payment conditions. Then, each clause is fed into a fine-tuned Legal-BERT model for classification. Legal-BERT is a pre-trained model based on the transformer architecture, pre-trained on legal texts, and is specifically designed for legal text understanding [1].

Let the representation of a clause be denoted as:

$$\boldsymbol{h} = \boldsymbol{BERT}(\boldsymbol{C})$$

where C represents the clause text and $h$is the contextual embedding generated by the Legal-BERT encoder. The probability of a clause belonging to class $k$is calculated using the softmax function:

$$\boldsymbol{P}\ (\boldsymbol{y} = \boldsymbol{k}|\boldsymbol{h}) = \frac{e^{W_k h + b_k}}{\sum_{j=1}^{K} e^{W_j h + b_j}}$$

The predicted clause label is:

$$\hat{\boldsymbol{y}} = \arg\max_{k}\ \boldsymbol{P}\ (\boldsymbol{y} = \boldsymbol{k} \mid \boldsymbol{h})$$

Similar transformer-based methods have been widely used for contract clause detection and legal text classification [3], [13].

D.Clause Dependency Modelling using Graph Neural Networks

Legal contracts often contain dependencies between clauses, where the interpretation of one clause may depend on another. To model these relationships, the system constructs a clause dependency graph:

$$G = (V, E)$$

where Vrepresents clauses as nodes and Erepresents relationships between clauses.

Graph Neural Networks propagate information across nodes to capture structural dependencies. The node update function is defined as:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in N(i)} W^{(l)} h_j^{(l)}\right)$$

where $N(i)$represents neighboring clauses.

Graph-based approaches have been shown to improve legal document retrieval and reasoning by modelling relationships between legal entities [4], [8], [25].

E.Risk Analysis and Compliance Evaluation

After clause classification and dependency modelling, the system evaluates contract risks. Each clause is assigned a risk score based on the classification probability and dependency influence.

The risk score for clause $i$is calculated as:

$$Ri = \alpha Pi + \beta Di$$

where

Pi = predicted risk probability from Legal-BERT

Di= dependency influence score from the GNN

$\alpha$and $\beta$are weighting parameters.

The overall contract risk score is computed as:

$$R_{contract} = \frac{1}{n} \sum_{i=1}^{n} R_i$$

Risk detection using machine learning and NLP has been widely explored for automated contract review and legal compliance analysis [14].

F.Output Generation and Visualization

The final stage generates a structured output summarizing the analysis results. The output includes:

1. Contract Summary
2. Clause Classification Results
3. Risk Assessment Scores
4. Clause Dependency Insights

The system also highlights risky clauses directly in the original document and generates downloadable reports. This automated pipeline assists legal professionals and compliance analysts in efficiently reviewing complex contracts.

V. CASE STUDY: LEGAL CONTRACT ANALYSIS APPLICATION

To validate the efficacy of the proposed framework, a case study was conducted in the domain of automated legal contract analysis. Legal contracts tend to be lengthy and involve complex clause structures. Such complex structures need to be analysed in detail to identify potential risks and ensure compliance with legal regulations. The aim of this case study is to validate the feasibility of the proposed system in the real-world legal contract analysis domain.

A.Case Description

A financial and employment-type agreement was presented as the case for the evaluation of the system. The contract was presented in PDF format to mimic a real-world situation where organizations would need to deal with both electronically created and scanned versions of legal agreements.

The suggested system utilizes the hybrid document extraction pipeline to extract machine-readable text from the contract. The text is processed to extract individual clauses from the agreement using contextual clues and rule-based boundary detection. This allows the framework to process the unstructured legal text into structured units for further analysis.

B.Clause Classification and Risk Detection

The clauses were processed using the fine-tuned transformer model for classification to determine their function within the contract. The system classified the clauses into different categories.

The framework also detects clauses that may indicate potential risks within the contract. High-risk clauses are related to clauses that indicate one-sided penalties, ambiguity in contract obligations, and restrictions related to intellectual property. Medium-risk clauses are related to clauses that indicate limited flexibility in contract terms and clauses that indicate insufficient legal safeguards. The overall contract risk score was computed by aggregating clause-level risk estimates, providing a quantitative measure of contractual safety.

C.Compliance Assessment

In the compliance assessment module, the extracted clauses were compared to predetermined contract requirements related to employment and financial contracts. The system checked for the inclusion of essential clauses such as termination and confidentiality clauses.The compliance level provided a structured approach to evaluating the completeness and preparedness of contracts.

D.Clause Dependency Analysis

In addition to the relationships between clauses in a contract, the framework also created a clause dependency graph to represent structural relationships between clauses. This was done to understand the semantic and functional relationships between clauses. In the Graph Neural Network module, the structural relationships between clauses were analysed to identify dependency patterns. This included indirect risk propagation and circular relationships. This provided a deeper understanding of how specific clauses in a contract relate to one another.

E.Analytical Output and Interpretation

The final system output consisted of a summarized contract overview, classification of clauses, risk severity indicators, evaluation of compliance, and dependency visualization. The clauses that were deemed to be risky were highlighted in the original contract for better interpretability. The above case study proves that the proposed framework can automate various stages of contract analysis with significant legal insights. The proposed framework reduces the overall human intervention in contract analysis, ensuring increased consistency in identifying risks.

VI.RESULTS AND DISCUSSION

A.Experimental Setup and Evaluation

The proposed AI framework was implemented as a complete end-to-end contract intelligence solution that included document ingestion, hybrid OCR-based text extraction, transformer-based clause classification, rule-based contractual risk identification, and graph-based dependency modelling. This document intelligence pipeline has also been explored in legal NLP research studies [3], [7].

The experimental evaluation of the proposed AI framework was conducted on a dataset of curated multi-page financial and employment contracts that mimic real-world contractual complexity. The dataset included a total of 530 contractual clauses, which were extracted from publicly available regulatory documents, including the Government of India General Financial Rules (GFR) dataset [Dataset]. A train-test split ratio of 80:20 was used for the experimental evaluation. The model was fine-tuned for three epochs, a batch size of 4, and a maximum sequence length of 512 tokens, using the Adam optimizer with linear learning rate scheduling. A two-layer Graph Attention Network (GAT) was used for inter-clause reasoning.

Evaluation metrics for the proposed AI framework included clause classification effectiveness, contractual risk detection capability, structural reasoning, and efficiency. The results for clause classification effectiveness are shown in Table no 1.

Table no 1. Clause Classification Performance

Model	Precision	Recall	F1-Score
Rule-based baseline	0.72	0.68	0.7
Transformer classifier	0.88	0.86	0.87
Proposed Framework	0.91	0.89	0.9

The results demonstrate that the proposed framework achieves improved contextual understanding of contractual clauses compared to traditional rule-driven approaches.

Table no 2. Risk Detection Accuracy

Approach	Risk Detection Accuracy
Keyword matching system	74%
Semantic classifier only	84%
Proposed hybrid framework	89%

The framework successfully identified high-risk contractual provisions such as one-sided liability clauses, restrictive termination conditions, and missing compliance clauses.

Table no 3. System Efficiency Metrics

Metric	Observed Value
Average document length	25 pages
Average clauses extracted	80+
Processing time per document	15–20 sec
Circular dependency loops detected	Up to 16

The dependency modelling module was also useful in identifying complex inter-clause relationships and structural risk propagation patterns, which is an identified challenge in automated legal reasoning systems [8], [9].

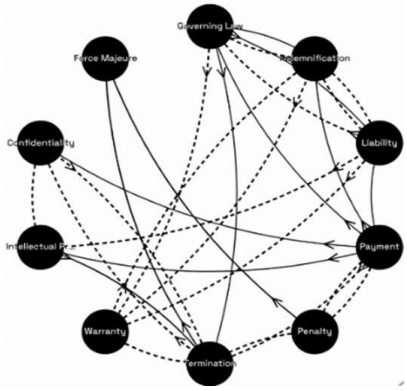

Fig. 2. Inter-clause dependency graph for structural relationships and circular contractual obligations, as identified by the proposed framework.

From the graph, it is evident that the different clauses, such as liability, termination, indemnification, and payment, are all interrelated and affect each other in some way. This is a clear example of how the proposed framework is useful for visualizing contractual risks. The classifier was also able to converge with a minimum evaluation loss of 0.097, which is a clear indicator that the model is stable and functional.

B.Comparative Insights

In comparison to traditional approaches to legal analysis based on rules or keywords [3], the presented framework allows for deeper semantic interpretation based on contextual modelling. Previous approaches to automated argumentation and legal reasoning were limited by the lack of dynamic grounding in external knowledge and structural information [6], [14]. By incorporating elements of semantic retrieval and adaptive knowledge enrichment based on retrieval-augmented approaches [15], [19], the presented system allows for more holistic approaches to contract intelligence.

C.Limitations and Challenges

Although the framework has produced promising results, its performance still largely depends on the quality of the language models used, as well as their consistency of reasoning, which has been a limitation of evaluating LLMs as discussed in recent studies [10], [11]. In addition, the quality of the documents as well as the presence of OCR noise might also pose a challenge. The complexity of processing the extracted data through a series of steps, especially during the construction of the dependency graph as well as the analysis of the risks, might pose a computational complexity challenge, as observed in studies related to multi-agent as well as graph-based reasoning systems [8], [9]. In addition, as the knowledge base increases, scalability might become a challenge, as discussed in adaptive RAG systems [15], [23].

D.Practical Implications

The experimental results show that the suggested framework can significantly help automate various contract intelligence processes such as compliance checking, enterprise risk analysis, and legal document summarization. This helps to detect any contractual risks at an early stage, thus facilitating risk-based decision support during the review of contracts.

The suggested framework can be effectively used in a large organizational environment where it becomes critical to interpret contractual agreements as well as maintain compliance with various regulatory requirements. This helps to automate the process of identifying critical contractual clauses that have risks or are not properly addressed from a legal perspective. This helps to enhance efficiency during the process of contract negotiation. Furthermore, the suggested knowledge enrichment process helps to enhance the analytical capabilities of the system over time.

The effectiveness of similar AI-based decision support systems has already been proven in various other domains such as healthcare analytics and policy evaluation. [1], [4].

VII. CONCLUSION AND FUTURE WORK

A.Summary of Contributions

In this paper, an AI-based framework was proposed for automated legal contract analysis that incorporates a hybrid document extraction mechanism, a transformer model for clause classification, a rule-based system for risk evaluation, and a dependency modelling mechanism based on graph theory. Contrary to traditional legal document processing systems that are primarily based on individual clause identification, this proposed system can provide an entire contract intelligence system based on semantic contract understanding and structural dependency relationships. The proposed system was implemented as a working prototype and was tested with actual styled financial and employment contract examples. The experimental results show that the system can provide valuable contract clauses that are risky in nature, identify missing compliance elements in the contract, and provide meaningful analytical results. In addition to this, adaptive knowledge enhancement can also be included in the proposed system to update its performance.

B.Key Findings

The research demonstrated the efficacy of integrating contextual language modelling and dependency-aware reasoning in the analysis of contracts. The system was successful in extracting clauses, maintaining risk scores, and enhancing interpretability by visual analytics. Dependency graph modelling was also successful in creating a deeper level of understanding about interactions in contracts. This demonstrates the potential efficacy of developing holistic AI-based contract intelligence systems.

C.Future Work

Future developments in this area would involve developing the scalability and computational efficiency of the system. This would involve developing features such as improved indexing techniques, light-weight language modelling, and legal knowledge graphs. This would help in enhancing the efficacy and factual accuracy of the system. Additionally, expanding the scope to different domains would also help in enhancing the system's applicability. However, it is also important to ensure the ethical and transparent usage of such AI-based contract analysis systems.

References

1. I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, Nov. 2020, pp. 2898–2904, doi: 10.18653/v1/2020.findings-emnlp.261. arXiv:2010.02559.
2. D. Hendrycks, C. Burns, A. Chen, and S. Ball, "CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review," arXiv preprint arXiv:2103.06268, Mar. 2021. (Presented at NeurIPS 2021 Datasets and Benchmarks Track.)
3. S. Moon, J. Lee, S. Lee, J. Oh, and S. Chi, "Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (BERT)," *Automation in Construction*, vol. 142, art. 104486, Oct. 2022, doi: 10.1016/j.autcon.2022.104486.
4. A. Louis, G. van Dijck, and G. Spanakis, "Finding the Law: Enhancing Statutory Article Retrieval via Graph Neural Networks," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)*, Dubrovnik, Croatia, May 2023, pp. 2761–2776, doi: 10.18653/v1/2023.eacl-main.203. arXiv:2301.12847.
5. F. Ariai, J. Mackenzie, and G. Demartini, "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges," *ACM Computing Surveys*, vol. 58, no. 6, art. 163, Apr. 2026, pp. 1–37, doi: 10.1145/3777009. arXiv:2410.21306 [cs.CL].
6. K. Vuthoo, S. Khetarpaul, and L. Venkata Subramaniam, "Datasets, Models and NLP Techniques for Legal Contracts – A Survey," Authorea preprint, Oct. 2025, doi: 10.22541/au.176044880.08032215/v1.
7. P. Shah, S. Joshi, and A. K. Pandey, "Legal Clause Extraction From Contract Using Machine Learning with Heuristics Improvement," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, Dec. 2018, pp. 1–5, doi: 10.1109/CCAA.2018.8777602.
8. Y. Tang, X. Wang, Y. Zhang, and J. Xiao, "CaseGNN: Graph Neural Networks for Legal Case Retrieval with Text-Attributed Graphs," arXiv preprint arXiv:2312.11229, Dec. 2023.
9. S. Tong, Y. Liu, and Z. Li, "Legal Judgment Prediction via graph boosting with constraints," *Information Processing & Management*, vol. 61, no. 3, art. 103678, May 2024, doi: 10.1016/j.ipm.2024.103678.
10. J. Niklaus, V. Sovukluk, M. Stürmer, I. Chalkidis, and M. Grabmair, "MultiLegalPile: A 689GB Multilingual Legal Corpus," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, Bangkok, Thailand, Aug. 2024, pp. 14947–14965, doi: 10.18653/v1/2024.acl-long.805.
11. C. Niklaus, M. Stürmer, I. Chalkidis, and M. Grabmair, "Legal-XLM-R: A Multilingual Legal Pre-trained Language Model," arXiv preprint (related to MultiLegalPile extensions), 2024.
12. A. Singh, "A Survey of Classification Tasks and Approaches for Legal Contracts," *Artificial Intelligence Review*, vol. 58, art. 380, Oct. 2025, doi: 10.1007/s10462-025-11359-8. arXiv:2507.21108.
13. S. Jha, "Natural Language Processing in Legal Practice: An Empirical Study on Contract Review and Risk Assessment," in *2025 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2025, pp. 1–8, doi: 10.1109/BigComp.2025.11135214.
14. M. H. Kazemi, S. Moon, and S. Chi, "Application of NLP-based models in automated detection of risky clauses in construction contracts," *Expert Systems with Applications*, vol. 255, art. 124678, Dec. 2025, doi: 10.1016/j.eswa.2024.124678.
15. Q. Wang, Y. Zhang, and Z. Li, "D2GCLF: Document-to-Graph Classifier for Legal Document Classification," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, WA, Jul. 2022, pp. 170–183.
16. M. Dechtiar, "Software engineering meets legal texts: LLMs for auto detection of contract smells," *Journal of High Technology Management Research*, vol. 36, no. 1, art. 100022, 2025, doi: 10.1016/j.hitech.2025.100022.
17. R. Ricciardi, "A multilingual BERT-based classification of reviews for enhanced visitors' experience analysis," *Scientific Reports*, vol. 15, art. 9418, Aug. 2025, doi: 10.1038/s41598-025-09418-9.
18. I. Chalkidis et al., "LLMs for Law: Evaluating Legal-Specific LLMs on Contract Understanding," arXiv preprint arXiv:2508.07849, Aug. 2025.
19. Y. Koreeda and C. D. Manning, "ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic, Nov. 2021, pp. 1907–1919.
20. I. Chalkidis, A. Xenos, and I. Androutsopoulos, "Legal Judgment Prediction: If You Are Going to (Predict) the Future, You Need to Look at the Past," *Artificial Intelligence and Law*, vol. 30, no. 1, pp. 1–35, Mar. 2022, doi: 10.1007/s10506-021-09295-0.
21. F. Ariai and G. Demartini, "Challenges and Opportunities in Multilingual Legal NLP: From Monolingual to Cross-Lingual Contract Analysis," in *Proceedings of the 2025 ACL Workshop on Legal NLP (LegalNLP 2025)*, Vienna, Austria, Jul. 2025.
22. M. Lippi and G. Torroni, "Context-independent claim detection for argument mining," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, Buenos Aires, Argentina, Jul. 2015, pp. 185–191.
23. M. Ruggeri, F. Zollo, and W. Quattrociocchi, "Risky clause detection in legal contracts via deep learning," arXiv preprint, 2022.
24. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Proc. 34th Conf. Neural Inf. Process. Syst. (NeurIPS 2020)*, Vancouver, Canada, 2020.
25. Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A Comprehensive Survey on Graph Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, Jan. 2021, doi: 10.1109/TNNLS.2020.2978386.
26. [Dataset] Government of India, General Financial Rules (Updated2024). https://doe.gov.in/files/circulars_document/FInal_GFR_upto_31_07_2024.pdf