International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 15 Issue 3, March-2026

Vision-Based Contract Document Risk Analyzer

Varsha Raju Gujrathi[1], Yash Ajay Parab[2], Jinay Dipak Patel[3], Prof. Devanand Bathe[4]

[1,2,3,4]Department of Artificial Intelligence and Data Science, K J Somaiya Institute of Technology, Mumbai.
University of Mumbai, Maharashtra, India.

ABSTRACT

In India Legal contracts are drafted in complex legal clauses which are difficult to interpret by non-expert people, which increase the risk of signing legally harmful clauses. Existing tools primarily focus on summarization, clause detection, or broad risk categorization, but they lack severity level of risk clause. To address this, our paper presents a two-stage clause-level risk classification framework that integrates document understanding and clause-level risk identification to aid non-expert users to take better decision. The proposed system extracts text form contracts, analyzes it and finds clauses and then applies hierarchical transformer-based classification approach. In first stage clauses are separated on the basis of low risk and risky using threshold optimization to improve recall for risky content. In the second stage risky clauses are again classified into medium or high-risk content along with a short explanation of risk. The proposed system is evaluated on manually annotated dataset containing 800 clauses extracted from real world contracts. Experimental results shows that the proposed system efficiently identifies high risk clauses, demonstrating its potential for assisting non expert users also.

Keywords: Clause-level analysis, supervised text classification, transformer, TF-IDF, natural language processing, legal contract processing, risk severity detection.

INTRODUCTION

In India, it is quite challenging for people to understand legal contracts which are drafted in complex language. Many drafts use legal terminologies that are not common and frame sentences such that it seems normal but can interpreted in many forms. This lack of clarity put the individuals in risk of agreeing to terms without fully understanding the implications of specific clauses and exposing them to legal and financial risks [1], [2]. Traditional approach to get insights from such documents is to involve a third person who has legal knowledge. This process is also a little time consuming and can cost much only for interpretation alone.

Recent development in artificial intelligence (AI) and natural language processing (NLP) are significantly helping legal documents analysis. Several studies demonstrate that automated legal documentation systems, conversational bots and workflow automation tools improves efficiency and accessibility in legal process [1], [2], [6]. Also, transformer models have been successfully applied to legal document summarization, clustering and semantic understanding. It enables improved relevant information extraction from large legal texts [5], [7], [8].

While these approaches improves document accessibility, most of the existing system operate at document level and focus on clause identification and general categorisation. They do not assess the risk associated with individual clauses [3], [4], [7]. Risk analysis methods that are available are often limited to professional use only limiting its use case to enterprise-oriented contracts or paragraph level classification. They do not provide fine grained assessment of contractual clause level risk severity [9], [10]. Also, large language model based approach depends heavily on prompt driven inference, which raise concerns regarding consistency in output for each query.

From a normal citizen perspective, the absence of clause level risk evaluation is a gap in existing solutions. Identifying potential impact of clauses on contracting parties is far helpful to non-expert user rather than identifying just cluses. Non-expert users need a system that can convey their relative risk in an interpretable and structured manner to take informed decision-making [15].

To address such limitations in a two-stage clause level risk classification framework focusing on risk severity identification is proposed. The proposed approach extracts contractual clauses from the documents and classify them into distinct risk levels. It enables users to efficiently identify potential unfavourable conditions. By using supervised learning approach tailored to legal text, the framework bridges the gap between automated document processing and decision supporting. The contribution of this work includes training supervised model on manual labeled data of clause level risk assessment, user-oriented contract analysis along with some features like missing data.

1. Related Work

Recent studies shows the application of artificial intelligence and natural language processing (NLP) for legal document analysis. Use of AI for simplifying and summarizing document that include LegalBERT and T5 based legal document assistant, integrated Long former based summarization, M2M-100 translation to multiple language [1], [2]. AI driven legal assistant to automate legal research, summarization and query answering [6]. These framework has shown improvement the efficiency and accessibility but they are not designed to identify risk severity which remains critical factor.

Transformer models are getting more recognition as used for text summarization, clustering and semantic analysis. Pre-trained model such as BART and PEGASUS have shown effectiveness in text summarization [5]. Context aware document clustering using DistilBERT, summarization using Pegasus and NLP driven legal assistants integrated with summarization and case similarity analysis improve document retrieval and accessibility [7], [8]. However, these approaches primarily focused on information condensation and accessibility not focusing on risk evaluation.

AI powered contract management system has shown efficiency gains, including time saving and manual error reduction, through automated analysis [4], [10]. These system are effective for operational purpose to increase efficiency but they do not provide risk associated with clauses.

Risk assessment in documents has been addressed in only few works. Chakrabarti et el. proposed early framework to identify risk prone paragraphs in legal documents using paragraph embedding and supervised classification [9]. Just recent studies have utilised pretrained models like Legal-T5 and Law-GPT for contract risk categorization through prompt based interface [14]. These are very effective methods but still these method explicitly clause level risk classification and interpretability.

Additional research has been done on optimising NLP models for multilingual document analysis and large scale legal document processing, improving task efficiency such as named entity recognition (NER) and document classification [11], [12]. End to end NLP framework Legal-BERT for generation of embedding followed by BART and T5 for summarization also for predictive task like document type classification and risk detection [13]. A survey was conducted highlighting work on judicial document analysis but shows limited focus on contracts and their risk assessment for non-expert user [15].

2. Proposed Work

3.1 System Overview

We propose a clause-level risk classification framework which will identify text from documents or images of contracts identifying the different clauses and analysing them to identify risk associated with them in hierarchical order. Also giving one liner explanation for the associated risk level. Primary objective is to help non-expert users to understand unfavourable contract provisions.

Document level risk classification dilute the risk across entire document, thus limiting interpretability. Our system provides fine grained clause level evaluation. Every contract is decomposed into independent clauses and enable localized risk detection. This allows the user to adapt minor changes at particular clause before signing.

Overall system follows a structured pipeline:

1) document acquisition
2) normalisation, preprocessing and clause segmentation
3) Hierarchical risk classification
4) concise explanation for predication.

The system architecture of the proposed framework is illustrated in Fig. 1.

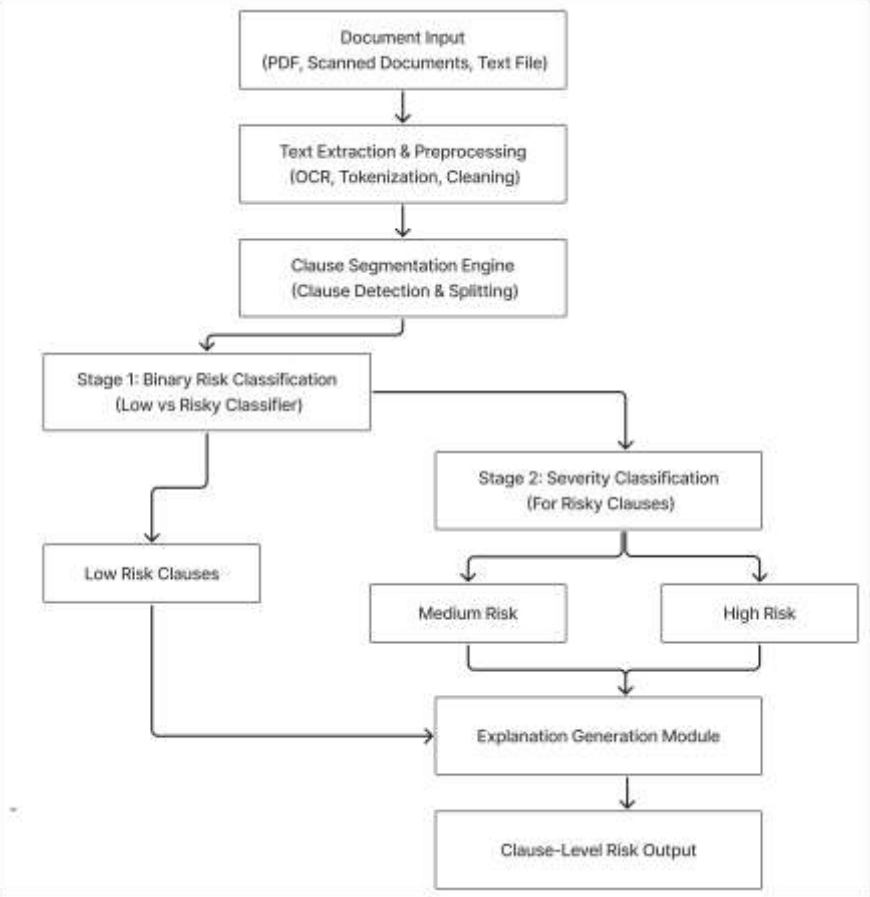

Fig. 1 The system architecture of the proposed framework

3.2 Clause Extraction

The proposed system starts with processing legal contract documents in text format or images. In case of images Optical Character Recognition (OCR) techniques are used to extract test. Extracted text is normalised and sentence segmentation is done.

Contracts are decomposed into individual clauses, which are then treated as fundamental units for clause level risk identification. Clause level processing provides fined grained evaluation of contract content and one liner explanation for evaluated risk level. This method overcomes the limitation of document level analysis.

3.3 Risk Severity Classification

Clause risk detection is formulated as a supervised learning problem at the clause level. Each clause assigned as their risk level as low, medium or high. Two modelling approach were explored: (i) NLP based pipeline using TF-IDF feature extraction with Logistic regression for multi class classification, (ii) transformer based approach by fine tuning legal domain language model. Preliminary experiments shows class imbalance and interclass confusion between medium and high risk.

To overcome class imbalance in dataset, methods like oversampling were applied. Model were evaluated using standard metrics including accuracy, precision, recall and F-1 score.

3.4 Two-stage Framework

The clause level dataset contain noticeable imbalance high risk underrepresented compared to low risk. Also linguistic overlap between medium and high risk classification caused interclass confusion. These led to misclassification for direct three class classification approach reducing reliability. To address these challenges we thought to use two stage hierarchical method.

In the first stage binary classification for low vs risky to maximise recall for risky clauses. These also solved a problem of being underrepresented making size comparable for low vs risky (medium and high) classification.
In second stage clauses that predicted as risky are further classified into medium risk and high risk. This decomposition simplified the decision boundary and reduce confusion between classes.

3.5 Decision Support Output with Explanation

This system is designed to focus on interpretability and usability by non-expert people. Associating each clause with risk level and explanation it provides actionable insights to user to take decisions about what he should do and what to avoid. Explanation component is added to translate complex legal language into simplified, user understandable summary. Unlike prompt driven LLM model approach, this model relies on trained classifiers. This ensures consistent results.

Overall, this work tries to bridge the gap between automated legal document processing and practical decision support by introducing clause level risk evaluation.

3. Dataset and Preprocessing

4.1 Data Collection

The dataset required for model training was constructed from publicly available legal contract documents which include rental and service agreements, obtained from open access repositories. These documents were selected to reflect real world contracts. Each contract was converted to machine readable text. Manual review done to extract individual clauses.

A total of 76 contracts were collected. From these set of documents about 750 clauses were manually extracted for annotation and explanation. On average 9-12 clauses extracted per document. An 80–20 split was used to divide the dataset into training and testing sets.

4.2 Annotation Process

Extracted clause were manually labeled with risk severity levels: low, medium and high along with explanation why it was considered as assigned level. Potential legal and financial impact of clauses were considered to mark their severity risk. Obligations, penalties, termination rights and liability related clauses were typically considered as medium or high risk based on conditions mentioned in contracts. Labeling criteria is defined as follows:

- Low risk: Informational or clauses with minimal legal or financial exposure.
- Medium risk: Clauses that include conditional obligations, moderate penalties.
- High risk: Clauses that contain high financial penalties, unilateral termination, strict liabilities.

Distribution of annotated was imbalanced. Majority of clauses are low risk followed by medium. The dataset contained around 400 low risk, 200 medium risk and 150 high risk clauses. This created a major challenge for direct three class classification.

4.3 Preprocessing

Standard preprocessing steps, that includes normalization and tokenization, were applied prior to model training to ensure consistency. Stop words removal was avoided to preserve the legal significance. Words such are ‘shall’, ‘may’ and ‘must’ plays crucial role in conditional obligations. These processed clauses were used as input for TF-IDF based and transformer based models.

EXPERIMENTAL RESULTS AND DISCUSSION

In this section experimental evaluation of the proposed framework is discussed. The experiments were conducted in step by step manner. We explored multiple approach to do risk analysis refine them based on observed metrics. This progression led to final two stage transformer based approach.

5.1 Experimental Setup

The dataset includes contract clauses annotated with risk levels and simple explanation. All experiments were evaluated using standard metrics that include accuracy, precision, recall and f1 score. Dataset was split into training dataset and testing dataset to ensure evaluation

5.2 TF-IDF Approach

Initial TF-IDF approach establish a baseline for risk classification. This approach is purely based on lexical features and term importance.

1) TF-IDF with Machine Learning Classifier: At first TF-IDF vectors were directly applied for clause level risk prediction. This approach gave reasonable performance for some classes. It failed to capture contextual dependencies in legal language. This resulted in frequent misclassification of medium and high risk clauses.

2) TF-IDF with Rule-Based Enhancement: To improve the performance, rule-based logic were applied with TF-IDF. These rules targets commonly occurring legal patterns. This hybrid approach improved recall for certain risk categories.

5.3 Transformer-Based Approach

TF-IDF was unable to cover the contextual meaning of clauses. There was also a dataset size constraint. So we decided to use pretrained transformers model. These models are already trained on large legal corpora. So they have better understanding of context of legal clauses.

1) Fine-Tuned Transformer attempt: In the initial attempt we try to fine tune hugging face's LegalBERT which understand contract structure, legal terminology. We did 5 epoch training. Due to limited data and class imbalance model showed unstable performance. After third epoch model began to overfit the majority class. High-risk clauses were classified as low and medium risk as their low representation in dataset. Metric score were poorer than TH-IDF based approach. It failed to achieve the performance.

2) Transformer Embeddings without Fine-Tunning: In this we directly used pre-trained model for feature extractors. Clause embedding were generated using CLS token representation and mean pooling over token embedding. In this approached we noticed improved understanding. The results were more stable than fine-tuned model. Performance was quite similar to TF-IDF approach.

5.4 Two-Stage Transformer-Based Classification

The final approach we thought of two adopts hierarchical two-stage transformer based classification. This method designed to improve both recall and class discrimination.

In the first stage we categorised the risk in broadly 2 types low risk and risky (combining medium and high risk). Primary objective to do this is improve recall for risky clauses. Therefore, instead of using default decision threshold of 0.5, we performed threshold tunning on validation set. The decision threshold was varied between 0.20 to 0.55, and recall was evaluated for each level. A threshold of 0.30 achieved the highest recall (0.875) while maintaining precision. It ensures that no high risks are classified as low risk. It simplifies decision boundary converting into binary problem. Thus first stage become our risk filter.

In second stage, clauses that were classified as risky in stage 1 were taken for further classification. These risky clauses were classified into medium and high risk. In legal documents medium and high risky clauses overlaps in linguistic pattern. It is hard form the machine to distinguish between medium and high risk. That caused the stage two classification accuracy dropped to 60%. This stage focuses on reducing inter class confusion between medium and high risk.

This hierarchical decomposition transform three-class problem into two simpler sub problems.
1) Risk detection (Low vs Risky)
2) Risk severity detection (Medium vs High)
This approach significantly improved accuracy precision and recall for the clause risk severity detection.

5.5 Performance Comparison

Table I. Comparison of Evaluated Approaches

Approach	Accuracy	Precision	Recall	F1-score
TF-IDF only	66%	0.64%	0.66%	0.65%
TF-IDF + rules	67%	66%	67%	66%
Transformers Fine tunned	47%	42%	47%	41%
Transformer (CLS)	61%	62%	61%	62%
Transformer (Mean pooling)	75%	76%	75%	75%

Table II. Stage 1 Risk Detection Performance

Stage 1 model (Low vs risky)	Accuracy	Precision	Recall	F1-score
Transformer binary	81%	82%	82%	82%

Table III. Stage 2 Severity Classification Performance

Stage 2 model (Medium vs High)	Accuracy	Precision	Recall	F1-score
Transformer binary	60%	62%	60%	60%

Table IV. Two-Stage Architecture Performance

Stage 2 model (Medium vs High)	Accuracy	Precision	Recall	F1-score
Transformer binary	83%	77%	78%	77%

DISCUSSION

This experimental result shows that lexical approaches like TF-IDF are not sufficient to capture semantic complexity of legal document. TF-IDF with rule offer slight improvement, but they lack adaptability as each time we might have to change the rule base. Transformer-based embeddings provide a significant improvement in context understanding without requiring fine tuning as it is already trained on large legal corpora. The proposed two stage transformer-based approach enhances risk severity discrimination making its application well suited for practical contract analysis and decision support.

CONCLUSION AND FUTURE SCOPE

This paper presents a two-stage clause risk classification framework to assist non-expert users in understanding the complex and potential unfavourable provisions mentioned in any legal contracts. Unlike the existing systems that works on entire document and general classification, our framework performs fine grained risk analysis at clause level. This study systematically evaluated multiple methods including TF-IDF based, rule enhanced, transformer based. These results showed that lexical approaches were unable to catch semantic meaning. Transformer model improved semantic understanding while suffered three class classification due to class imbalanced and inter class confusion.

To address this, two-stage strategy was introduced. First stage prioritizing recall distinguishing between low risk and risky using threshold tunning. While second stage further classifying into medium and high risk. This decomposition reduced class ambiguity, improved discrimination performance. Overall, the proposed framework demonstrates strong potential for clause risk detection and practical use in contract review systems.

Despite improved results, the study had some limitations. The dataset size is relatively small and limited to some specific contracts that may affect generalization. Future work will focus on expanding the annotated dataset and variety of contracts. Domain specific fine tuning transformers on large corpora may enhance the performance. Integrating multilingual contract analysis may increase usability for users.

REFERENCES

1. G. S., R. R., S. M. R., J. M., J. E. N. and S. K. L., “AI Legal Documentation Assistant,” in Proc. 2024 Int. Conf. Smart Technologies for Sustainable Development Goals (ICSTSDG), Chennai, India, 2024, pp. 1–5, doi: 10.1109/ICSTSDG61998.2024.11026581.
2. A. K. R., A. V. R., S. V., S. N. and P. R., “Revolutionizing Legal Workflows: Advanced AI Techniques for Document Summarization, Legal Translation, and Conversational Assistance,” in Proc. 2025 Int. Conf. Advanced Computing Technologies (ICoACT), Sivakasi, India, 2025, pp. 1–4, doi: 10.1109/ICoACT63339.2025.11004791.
3. R. K., P. Gupta, G. Suthar, K. S. Sidhu, R. Sarkar and P. Satya Narayana, “Natural Language Processing for AI-Powered Legal Document Analysis,” in Proc. 2025 Int. Conf. Computing Technologies & Data Communication (ICCTDC), Hassan, India, 2025, pp. 1–5, doi: 10.1109/ICCTDC64446.2025.11159030.
4. P. G. Thirumagal, M. M. A. Raj, S. J. Naser, N. A. Hussien, J. K. Abbas and S. Vinayagam, “Efficient Contract Analysis and Management through AI-Powered Tool: Time Savings and Error Reduction in Legal Document Review,” in Proc. 2024 Ninth Int. Conf. Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2024, pp. 1–6, doi: 10.1109/ICONSTEM60960.2024.10568823.
5. A. Kasar, S. Matade, D. Rasal and S. Shinde, “Enhancing Summarization of Legal Text Documents using Pre-trained Models,” in Proc. 2025 Int. Conf. Emerging Systems and Intelli gent Computing (ESIC), Bhubaneswar, India, 2025, pp. 58–61, doi: 10.1109/ESIC64052.2025.10962732.
6. M. Nithya, H. S., K. S. and S. K., “AI-Driven Legal Automation to Enhance Legal Processes with Natural Language Processing,” in Proc. 2024 Int. Conf. IoT Based Control Networks and Intelligent Systems (ICICNIS), Bengaluru, India, 2024, pp. 1246–1253, doi: 10.1109/ICIC NIS64247.2024.10823316.
7. A. Rao, A. Halgekar, D. Khankhoje, I. Khetan and K. Bhowmick, “Legal Document Clustering and Summarization,” in Proc. 2022 6th Int. Conf. Computing, Communication, Control and Automa-tion (ICCUBEA), Pune, India, 2022, pp. 1–4, doi: 10.1109/IC CUBEA54992.2022.10010585.

8. H. Irfan, S. Peerzada and M. Mansoor, “Transforming Legal Workflows: A Deep Dive into NLP Solutions for Legal Challenges,” in Proc. 2024 19th Int. Conf. Emerging Technologies (ICET), Topi, Pakistan, 2024, pp. 1–6, doi: 10.1109/ICET63392.2024.10935072.
9. D. Chakrabarti et al., “Use of Artificial Intelligence to Analyse Risk in Legal Documents for a Better Decision Support,” in Proc. TENCON 2018– IEEE Region 10 Conf., Jeju, South Korea, 2018, pp. 683–688, doi: 10.1109/TENCON.2018.8650382.
10. M. Fthima, D. P. Dhinakaran, T. Thirumalaikumari, S. R. Devi, B. M. R. and S. P., “Effectual Contract Management and Analysis with AI-Powered Technology: Reducing Errors and Saving Time in Legal Document,” in Proc. 2024 Ninth Int. Conf. Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2024, pp. 1–6, doi: 10.1109/ICONSTEM60960.2024.10568733.
11. A. Radhika, N. K. Bhasin, S. R., Y. R. Raju, K. N. V. Satyanarayana and I. I. Raj, “Optimization of Natural Language Processing Models for Multilingual Legal Document Analysis,” in Proc. 2024 Third Int. Conf. Intelligent Techniques in Control, Optimization and Signal Pro cessing (INCOS), Krishnankoil, India, 2024, pp. 1–6, doi: 10.1109/IN COS59338.2024.10527598.
12. B. S. Reddy, A. Balavivekanandhan, P. Lakhera, R. Changala, R. Sabareesh and A. Balakumar, “Optimization of BERT Algorithms for Deep Contextual Analysis and Automation in Legal Document Pro cessing,” in Proc. 2024 15th Int. Conf. Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 2024, pp. 1–6, doi: 10.1109/ICCCNT61001.2024.10723962.
13. R. Shinde, C. Dathasai, V. J. Aditya, S. Srinivasreddy, S. Hariharan and D. Geetha, “Understanding Legal Document using Natural Language Processing Approach,” in Proc. 2025 5th Int. Conf. Pervasive Computing and Social Networking (ICPCSN), Salem, India, 2025, pp. 1313–1317, doi: 10.1109/ICPCSN65854.2025.11035732.
14. P. A., K. V. Nagaraja and M. Venugopalan, “Legal Contract Analysis and Risk Assessment Using Pre-Trained Legal-T5 and Law-GPT,” in Proc. 2025 3rd Int. Conf. Integrated Circuits and Communication Systems (ICICACS), Raichur, India, 2025, pp. 1–8, doi: 10.1109/ICI CACS65178.2025.10968817.
15. T. Ghosh and S. Kumar, “A Survey of Legal Text Analysis Techniques for Indian Legal Documents,” in Proc. 2024 Int. Conf. Circuit, Systems and Communication (ICCSC), Fes, Morocco, 2024, pp. 1–6, doi: 10.1109/ICCSC62074.2024.10616889.