International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2054-2060
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0275

Legal Insights: Document Summarization and Risk Analyzing Assistant

Kalpana .S [1], Dhanushree V [2] , Kiruthi.R [3] , Srinithi.A[4] , Thamarai AJ[5]
[12345]Artificial Intelligence and Data Science, Jai Shriram Engineering College, Avinashipalayam,Tirupur.
Email ID: *kalpanatamil74@gmail.com[1], dhanushree051@gmail.com[2], kiruthir2609@gmail.com [3], srinithisri00@gmail.com [4] , thamaraiarunachalamm@gmail.com[5].*

Abstract
Even non-experts are not always able to comprehend legal texts like contracts, agreements and notices due to the fact that they are in other times lengthy, complex and highly technical. Manual review of such texts is the obstacle to their access and scalability as it is time consuming, error prone and in most cases, because it is manual, is reliant upon its manuality. In order to automate the intake process, legal documents intake, interpretation, and simplification process, the paper presents the design and implementation of a Legal Document Summarization and Risk Analysis System, which is an AI based system. The suggested system is a rule-based risk detector to detect large clauses, obligations and potential legal issues and an optical character recognition (OCR) to read documents on scans and natural language processing (NLP) to discover plain-language summaries.The architecture focuses on practicalities of deployment, understandability and modularity, rather than theoretical innovation, such that users can know the legal implications of major consequences without necessarily having to research that aspect of the law on a regular basis.

Keywords: *Legal documents Analysis, text summarization, risk identification, natural language processing, optical character recognition, explainable artificial intelligence, web-based artificial intelligence systems.*

1. Introduction

In order to manage the person, business and institutional operation, there should be legal documents to limit the activity. The rights, liabilities and obligations of parties are outlined in contracts, agreements and in legal notices. These materials though are subject to very keen interpretation and experience within the field to be interpreted. The non-legal professionals might find it difficult to have the legal documents deciphered as they have complex structures of language, terminologies, and legal threats that they propose. Because of this fact, people and small-scale enterprises are inclined to take the services of attorneys even to the routine analysis of the documents that raises the prices, time, and access to timely legal help. The need to facilitate the process of comprehending legal documents through automated systems with the ability to help in such an understanding continues to grow without necessarily affecting the accuracy or the interpretability of such documents because of the massive legal document digitalization. The old methods of searching the keyword and rule extraction systems are not of much assistance as they lack the contextual understanding that will help in getting the semantic meaning of the legal provisions. Long form legal contracts may be under the condition of hidden risks, omissions and dangers that these methods fail to detect. This causes them to be less practical in the study of law. The recent progress in the Natural Language Processing(NLP) and Artificial Intelligence (AI) has shown to be incredibly hopeful with regards to the automation of semantic analysis, information extraction and text summary. The systems are capable of tackling the processing of high volumes of legal content and conserving the contextual senses of meaning of the content because of the AI-based methodologies. However, most of the currently offered solutions pay an excessive amount of attention to the functioning

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2054-2060
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0275

of theoretical models and the accuracy of benchmarks, typically disregarding the applicability of such considerations as end-to-end system usability, meaningfulness of results and practical deployment challenges. To mitigate such issues, the work proposes an implemented AI-driven system, along with clause-level risk detection, text aggregation, document ingestion, as a single web-based system. The system suggested will focus on plain-English summary generation and recognition of important clauses and legal risks that may occur in the language understandable to humans. The system is implemented in a manner that, it supports the real life processes of legal documents and enable nonexpert users to make informed decisions due to a general focus on practicality, accessibility and explainability.

2. Related Work

The growing volume of digital legal information and its growing complexity have also generated additional research effort on the analysis of legal texts. The first methods of information extraction and search were rule-based information extraction methods and key-word search to find the relevant sentences and legal words. These were the comparatively weak means of carrying out a thorough legal study because they had little semantic awareness and could not determine contextual relations among clauses even though they had primitive document search abilities. Natural Language Processing (NLP) technology has been studied in numerous studies concerning the processing of legal documents. These methods are dependency parsing, named entity recognition, tokenization, and part-of-speech tagging methods.These methods enabled the legal agents like the date, obligations, parties, and monetary values to be obtained. These systems were however not readily scaled to a large variety of legal documents as well as nations due to the large amount of custom rule generating and domain specific modification required. Machine learning has also come up with models of supervised and unsupervised learning to classify and summarize legal documents. Statistical and neural-based techniques of summaries have left behind to generate brief presentation of lengthy legal texts.Even though this enhanced the quality of the summaries, majority of them only addressed text compression without regard to the clauses that were of legal significance or hazards that were of interest to the end-users. The recent works in the area of legal document understanding using deep learning and transformer-based models with contextual embeddings to classify the clauses and conduct semantic similarity analysis were investigated. These frameworks had the capability of depicting long-range dependencies and complicated legal semantics.This is, however, not the case in real- life legal situations, need explanations, and consequently this kind of solution can be costly in terms of annotated datasets, computational facilities, and black box decision-making. There is also the case of risk detection and obligation extraction which have been addressed through hybrid techniques that integrate NLP techniques with rule-based reasoning. Although these systems made it easy to identify high-risk clauses, they were normally done in disjointed segments as compared to end-to-end solutions. Moreover, much of the literature available highlights accuracy measures at the model level, and is not focusing on sufficiently practical deployment, interpretability and user interaction. The suggested system is centered on building usable and comprehensible AI-based system as opposed to the current methods that are geared towards applying risk detection and document input on a case-by-case basis and summarizing the outcomes in just a single web-based solution. The system will facilitate the gap between the academic research and legal work process of documents processing by integrating the OCR, the NLP-based summarization, and the rule-based risk analysis and presenting the unprofessional users with the simplified legal information [1].

3. Proposed System

The solution being suggested will be an AI- powered system of Legal Document Summarization and Risk Analysis that will presumably help non-legal users in their interpretation of complex legal documents. The

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2054-2060
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0275

system is targeted at the automation of the processing of legal documents through translating the legal texts to simplified and easy to understand products without the loss of legal content and identification of the potential hazards. The approach proposed is founded on practicality, explainability and ease of implementation as opposed to the introduction of new theoretical models. Legally formatted documents scanned or generated digitally can be inputted into the system. The machine readable text is lacked by the application of Optical Character Recognition (OCR) that is implemented in scanned documents. In order to obtain it and undergo more analysis this extracted text is first processed using Natural Language Processing (NLP) tools such as text normalization, sentence partitioning and tokenization. As part of the system, the NLP-based summary methods are used to provide brief plain-language summaries of long legal texts after the preprocessing phase; the summarizing component is aimed at providing an account of the crucial information, including the purpose of the document, major points and requirements. The reason is that it enables users to be familiar with the general purpose and structure of the document soon. The solution suggested in the paper is a rule-based way of synthesizing risk analysis and summarization at clause-level.Legal terms are reviewed to find the requirements, sanctions, termination, rules of compliance and other risky factors and the system is developed as a web-based one presenting the information about the risk and the condensed material. The system is designed so that one does not have to have any legal knowledge in order to use it since a user could upload a document, view them and condensed summaries and highlight the risk areas. The architecture can be designed in a modular format since it is possible to make single components, like risk identification, summarization, and OCR, developed or updated separately. On the whole, the given solution will minimize the amount of manual work to interpret legal documents, enhance comprehension of documents and make informed decisions.

4. System Architecture

The legal Document Summarization and Risk Analysis System proposed is designed to be scalable, interpreted, and efficient in both processing legal documents, owing to the modular and tiered design. The architecture modules are such that they guarantee the smooth flow of data, such as the entry of documents to the display of the results and each of these modules is tailored to do a particular job [2].

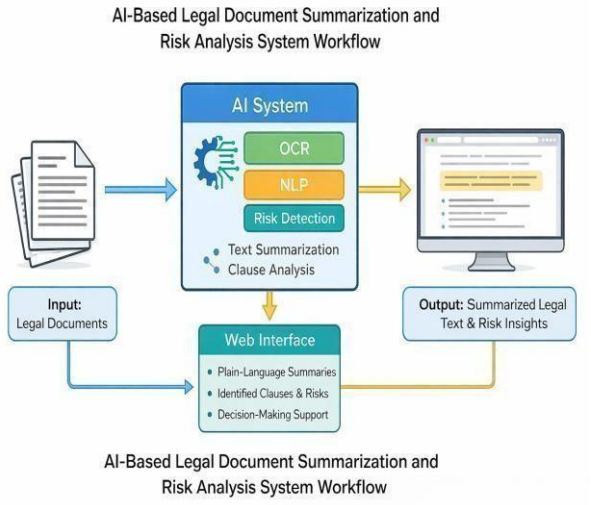

Figure1 : AI based system

4.1.Document Input Layer

The point of entry point of the system is Document Input Layer. A user can also post legal documents including agreement, contracts and legal notices using an online interface. The system is dynamic when it comes to processing of various types of documents since it accommodates scanned documents and digitally generated documents.

4.2.Big Data Character recognition (OCR) Module

Scanned documents and image based documents shall be converted to machine readable texts using OCR module. This is an Optical module that preserves the structural integrity of original document as it retrieves the textual contents. The text that was retrieved is then received in later modules to be further analyzed.

4.3.Text Preprocessing Module

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2054-2060
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0275

The Text Preprocessing Module is employed to clean and normalize the text taken with the help of the algorithms of the Natural Language Processing. This is text normalization, noise suppression, text segmentation and tokenization. Preprocessing will result in the further accuracy and consistency of operations such as risk examination and downstream recapitulation.

4.4.Summarization Module

Summarization module,Matters the Summarization module is applied to give a summary of lengthy legal documents in simple English. Although it also has the original meaning as it is used in the legal practice, it also highlights key issues such as the intent of the document, some of the key provisions and the pre-requisites. In addition, it is easy to learn what is contained in the document without reading as much as this module can enable one to navigate easily [3].

4.5.Risk Analysis Module

The Risk Analysis Module considers all the provisions with the consideration of determining the legal risks. The combination of rule-based analysis and semantic understanding is known as the module and recognizes high-risk items, including obligations, penalties, termination, compliance, and so on. The identified risks will be classified in order to assist the user in focused reading.

4.6.User Interface Module

The Web-Based User Interface Module reflects the condensed information and the outcomes of the risk analysis in a simple fashion. Insights can be explained,highlighting of risk clauses and summaries can be given to the users. The interface should be easy to use and understand by the non- legal users.

5. Methodology

To transform unstructured litigations into intelligible form of risk analysis and brief summaries that can be understood by common individuals, the proposed AI-based Legal Document Summarization and Risk Analysis System will follow the systematic steps. It is very accurate, easy to use and efficient in dealing with digitized and scanned legal files.

5.1.Collection and Data preparation

The quality of the legal papers collected all contain a good quality of the contracts, agreements, legal notices, which are available on the trusted sources of laws, and in the free databases. With the Optical Character Recognition (OCR) technology, it has been possible to convert documents scanned into machine-readable documents.In order to increase the quality of the data derived, further purification of the resultant text is done by elimination of redundant symbols, stop words and noise.

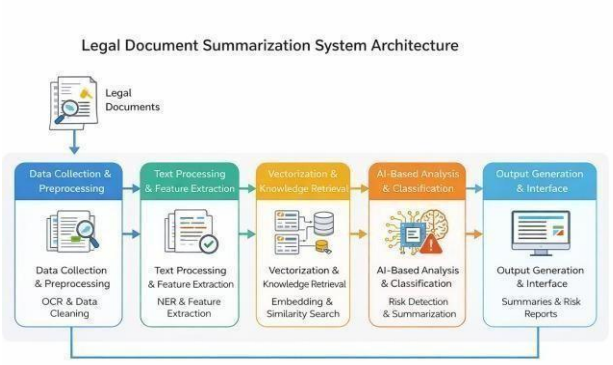

Figure 2: legal document

5.2.Processing and Feature Extraction Texts

The process that is taken to process the preprocessed text is known as Natural Language Processing (NLP). Important legal aspects, like clauses, duties, dates, and parties are identified with the assistance of tokenization, syntactic parsing, and named entity recognition (NER). These characteristics will contribute to the interpretation of legal texts and their meaning.

5.3. Retrieval and Vectorization of Knowledge.

Embedding techniques are a form of transformation of the processed text into numerical representations as vectors. These vectors give it the ability to effectively execute similarity searches and access the appropriate legal information in the knowledge base. This step assists in generating the right summary and risk identification.

5.4.Analysis and Classification of the Ai

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2054-2060
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0275

The models of machine learning and deep learning assist in categorizing legal terms and identifying possible hazards. The system examines the patterns of texts in order to detect dangerous phrases, clauses that have been eliminated, and compliance problems. Document summarizing is expected to make briefs that are useful and short.

5.5.Generation of output and interface

Report on risk analysis and summaries that are organized are some of the end products. The online or application interface of the results is user friendly allowing the user to comprehend important legal information and make informed decisions within very short time.

6. Experimental Model And Measurement

This section outlines the setting of the experiment and the procedures to be used to evaluate the usefulness, dependability, and efficiency of the presented AI-based Legal Document Summarization and Risk Analysis System.

6.1. Experimental Design:

The experiments conducted were performed upon a collection of various legal documents (contracts, agreements and legal notices). The strength of the system was tested using digitally generated and scanned documents. The documents were of different length and different structure in order to reflect the practice situation of the legal document in the real world. The system was executed in a controlled environment and was tested as the final application OCR scan tests on scanned documents were conducted to find out the accuracy of the text extraction. It was extracted and preprocessed the text and utilized to test the summarization and risk analysis modules. The usability of the system and the consistency of the response were measured using the web-based interface.

6.2.Data description

The data set involving documents of the majority of the regions like employment agreements, service contracts, rental contracts, etc. as a result of such diversity, the system was tested in the diversity of the legal settings. Documents were anonymized in order to avoid information leakage which is sensitive.

6.3.Accuracy of OCR

This approach measures how accurate the text extracted after scanning documents is compared to ground truth text. Summarization Quality is quantified by readability, coherence and coverage of summaries.

6.4.Risk Detection Accuracy

This is a measure of the ability of the system to detect and to indicate high risk terms like commitment, fines, and termination terms. With the assistance of a comparison between the successful detection of risk clauses and the false positives and false negatives, the precision and the recall are used to measure the effectiveness of the risk detection at the clause level.

6.5.Processing Time

This is a measure of system efficiency and this is through the amount of time that it consumes when processing documents in an upload to result production situation.

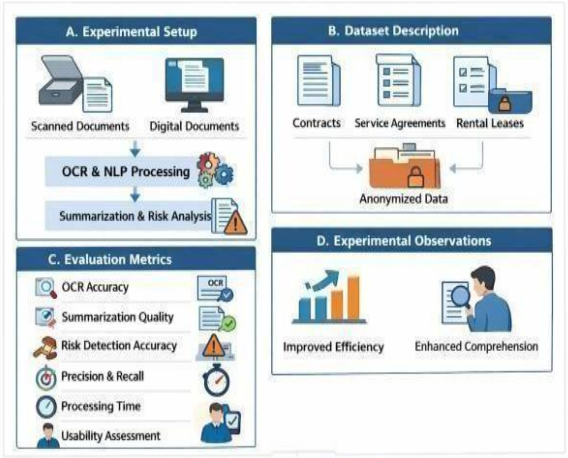

Figure 3: Experimental Setup and Evaluation **Metrics**

6.6.Usability Assessment

Qualitative was gauged upon the output clarity, usability and readability of the risk indicators and summaries in the eyes of the user.

6.7. Experimental Observations

The experimental research was conducted as regards functional correctness, output clarity and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2054-2060
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0275

system stability rather than the theoretical standards of performance. The results indicate that the proposed system is effective in the source of labor cuts concerning the manual review, and the interpretation of the legal documents in the circles of users that will not seek the assistance of a lawyer.

7. Outcomes And Conversation

Both digitally and scanned legal papers were used to test the proposed AI-based risk analysis and legal document summarizing. The system was able to extract text successfully in a variety of documents and produce succinct summaries and highlight legal risks that were critical. It was enabled by the OCR integration to be able to deal with the scanned documents and generated readable text that could undergo further NLP processing.The summary module was positioned to successfully narrow down the document without losing. Important provisions, duties, and conditions, which were legally binding. This rendered the long legal texts simple to perceive the overall idea of these texts in a considerably brief time frame without missing the critical information.The logical consistency of summaries generated with the original data and situational concreteness allowed the entire paper to be more understandable, and coordinated work of the risk analysis, NLP, and OCR modules was secured by the integrated backend architecture. The web interface was further simplified as the results were displayed in an organized and comprehensible manner. Overall, the findings have validated the feasibility of the proposed approach as it is decipherable and appropriate to the study of legal texts in real life scenarios wherein purely black-box models might not be the best option to use.

Conclusion

In this paper, a proposal was made to create an AI-based Legal Document Summarization and Risk Analysis System that will help reduce the complex legal documents and the level to which laypersons can read and comprehend. Legal texts are ever lengthy, complicated and hard to comprehend as they involve use of professional terminology and legal hazards. The suggested solution will solve these problems by surrounding document import and summarizing and recognizing risks on the clause level in a single and comprehensible format. The system provides the use of rule-based semantic analysis to recognize potential legal hazards, optical character recognition (OCR) to document scanners, and natural language processing (NLP) to perform basic text pretreatment and summarization.The system helps significantly to reduce the amount of work necessary to be done manually and to improve the comprehension of documents by summarizing them in plain language and highlighting meaningful sentences. The web-based interface and modular system are also enhanced to meet the scalability, usability and an easy deployment. According to the analysis of the experiment, the suggested approach is efficient to provide significant summaries and identify the clauses that are of high risk in various legal writings. When comparing it to the manual review and simple search according to the key words, the results show a greater efficiency, clarity and reliability. It is the system that is unusually applicable to the real life legal document processing due to the fact that it is more concerned with the interpretability and practical use as compared to the complexity of the model. To conclude, the suggested AI-based system is an effective decision support system that can help non-expert users to understand intricate legal texts. It might help people and small businesses to make the correct decisions and minimize the amount of the necessity to visit the professionals on the daily basis and analyse the document.

References

[1]. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin published the paper titled “Attention Is All You Need” in 2017 at the Neural Information Processing Systems (NeurIPS) conference.

[2]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova published “BERT:

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 04 April 2026
Page No: 2054-2060
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0275

Pre-training of Deep Bidirectional Transformers for Language Understanding" in 2019 at the NAACL-HLT conference.

[3]. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu published "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" in 2020 in the Journal of Machine Learning Research.

[4]. Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu published "PEGASUS: Pre-training with Extracted Gap-Sentences for Abstractive Summarization" in 2020 at the International Conference on Machine Learning (ICML).

[5]. Abigail See, Peter J. Liu, and Christopher D. Manning published "Get To The Point: Summarization with Pointer-Generator Networks" in 2017 at the Association for Computational Linguistics (ACL).

[6]. Rada Mihalcea and Paul Tarau published "TextRank: Bringing Order into Texts" in 2004 at the Conference on Empirical Methods in Natural Language Processing (EMNLP).

[7]. Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos published "Extreme Summarization of Legal Texts" in 2019 at the Association for Computational Linguistics (ACL).