

Quantitative Risk Scoring of Loan Documents using Weighted Keyword Matching and Semantic Similarity

Khushi Mantri¹, Rashi Potey², Shinosha Jain³ and Sunidhi Admane⁴

¹⁻⁴Department of Computer Science and Engineering, Ramdeobaba University, Nagpur, India
Email: mantrika@rk nec.edu, poteyr1@rk nec.edu, jainsr_3@rk nec.edu, admanesp@rk nec.edu

Abstract— LegalSutra is a rule-based system that automatically analyzes loan documents to assess legal risk. It identifies predefined risk phrases and proposes to use semantic similarity to detect variations. The system penalizes missing key clauses, such as repayment or termination terms, which may signal hidden risks. It outputs a normalized risk score—categorized as low, moderate, or high—and includes encryption and summarization features for secure, scalable, and consistent legal risk evaluation without needing large labeled datasets.

Index Terms— loan documents, legal risk, semantic similarity, rule-based scoring, omission penalty, legal document analysis, document encryption.

I. INTRODUCTION

Legal and financial contracts, especially loan agreements, often contain clauses that introduce regulatory or financial risks if poorly drafted. Repayment terms, default conditions, arbitration, and compliance requirements can affect both lenders and borrowers.

Legal experts currently spend significant time reviewing these documents, but manual review is slow, inconsistent, and difficult to scale for organizations handling large volumes.

Recent studies have applied AI and NLP to contract analysis. IEEE ICONSTEM 2024 discusses clause extraction systems, while IJECE (2021) surveys rule-based and machine learning methods, mostly for semi-structured data. Other work has explored RAG-based legal chatbots (IEEE I-SMAC 2024), topic modeling in judiciary automation (IEEE ICSSAS 2024), and deep learning for semantic understanding (IEEE TNN, 2023). These efforts reduce review effort but lack real-time, clause-level risk scoring.

To address this gap, we present LegalSutra, a hybrid NLP system for automated risk evaluation in loan agreements. It combines weighted keyword matching with semantic similarity detection to capture both explicit and implied risks. spaCy's PhraseMatcher identifies exact terms, while using Sentence-BERT, we propose to handle semantically related phrases. The system also assigns omission penalties when critical clauses—such as repayment or dispute resolution—are absent. The output is a normalized risk percentage, classifying documents as low, moderate, or high risk.

LegalSutra offers a rule-based scoring framework that avoids reliance on large annotated datasets. It integrates semantic similarity to improve recall in identifying risky clauses. It also accounts for risks arising from missing provisions through a penalty mechanism. Together, these elements provide a practical and interpretable approach for contract risk assessment, supporting more consistent and scalable legal review.

II. LITERATURE REVIEW

Legal document automation research has evolved through several strands in the last ten years, reflecting both the promise and the limitations of technology in this domain. Early initiatives by Smith et al. [1] explored the creation of structured contract templates, which streamlined drafting tasks and improved uniformity but offered no systematic way of assessing legal or financial risks embedded in the text. Similarly, Johnson [2] focused on clause extraction through rule-based parsing, which provided greater segmentation accuracy and formal structure but did not extend to interpretive analysis or any quantitative risk-scoring framework. As the field matured, attention increasingly shifted toward NLP-driven methods capable of handling unstructured data. Patel and Mehra [3], for example, demonstrated how machine learning could automate the classification of loan-related documents, while Brown and Taylor [4] highlighted the role of semantic similarity techniques in flagging potentially high-risk clauses.

Despite these contributions, such efforts tended to remain confined to classification and detection, without addressing the broader challenge of building an integrated risk analysis framework that considers not only explicit terms but also omissions or implicit contractual weaknesses. Parallel streams of research have focused on supporting technologies that complement, but do not fully solve, the problem of legal risk analysis. Williams [5] demonstrated how reliable PDF-to-text pipelines can reduce preprocessing overhead, an essential prerequisite for downstream NLP tasks.

Chen and Kumar [6] advanced clause extraction with deep learning techniques, reporting accuracy gains but leaving open questions around explainability and risk evaluation. Ahmed and Gupta [7] developed legal text summarization systems that effectively condensed lengthy case materials into more manageable outputs; however, these summaries were descriptive and not connected to risk scoring or decision support. Ontology-based frameworks such as those of Nakamura and Lee [8] contributed valuable interpretability and rule-based reasoning but struggled with flexibility when applied to unstructured financial contracts, which often involve ambiguous language and exceptions.

Retrieval-based work by Rossi and Bianchi [9] enhanced access to vast corpora of juridical texts and improved search efficiency, yet stopped short of providing clause-level evaluation or actionable insights for practitioners. More recently, Zhang and Wang [10] introduced AI-powered compliance monitoring systems capable of tracking adherence to regulations at the organizational scale.

III. METHODOLOGY

A. Document Ingestion and Text Extraction

LegalSutra begins its operation when the user uploads a contract in PDF format. For backend processing, pdfplumber is used to extract text from PDF documents. Extracted text is cleaned by removing unnecessary formatting, headers, and footers to ensure consistency. The resulting plain text serves as the basis for clause-level processing.

B. Clause Segmentation

The plain text is segmented into independent clauses to enable fine-grained risk evaluation. Logical breaks are identified using sentence endings, punctuation marks, numbered formatting, and common legal connectors. Each clause is treated as a self-contained unit and stored independently for linguistic and semantic evaluation.

C. Linguistic Preprocessing Using NLP

Each clause undergoes NLP-based preprocessing through SpaCy. Tokenization is used to isolate individual words and symbols, while lemmatization standardizes word forms for more reliable matching. Part-of-speech tagging provides grammatical context, and NER is employed to extract relevant entities such as monetary values, names, durations, and organizational references. These linguistic annotations are preserved for downstream risk detection.

D. Semantic Similarity via SBERT

To identify semantically risky clauses beyond keyword matches, LegalSutra proposes to use SBERT to convert clauses into vector embeddings. These will be compared against a reference set of high-risk clauses, and those exceeding a similarity threshold proceed to risk scoring. This will ensure nuanced or paraphrased risks aren't overlooked.

E. Keyword Matching and Penalty Accumulation

It also proposes to add domain-specific keyword list identifies high-risk terms in each clause, each assigned a penalty based on legal impact. These penalties can be summed to produce a raw risk score, reflecting the clause's legal sensitivity for classification.

F. Percentage Conversion and Risk Categorization

The clause score is normalized into a percentage for better interpretability. Depending on set thresholds, the risk is categorized as low, medium, or high. For example, scores below 30% may indicate Low risk, while those above 70% may indicate High. These thresholds were calibrated during testing using domain-specific guidelines.

G. Output Generation in PDF Format

After processing, all clause results are compiled into a structured PDF report, including scores, matched keywords, similarity indicators, and risk categories. LegalSutra uses htmltopdf.js to generate the downloadable report, ensuring confidentiality and offline access. The PDF is designed for easy review, sharing, and archiving.

H. Final Risk Scoring Formula and Thresholding

LegalSutra computes a total risk score by combining the weighted sum of matched keywords and penalty scores for missing but expected legal clauses. Here's how we calculate the total score:

$$\text{Formula: Total Score} = \sum (\text{occurrence}_i \times \text{weight}_i) + \sum (\text{penalty}_j)$$

Where:

i refers to each matched risk keyword

j refers to each missing clause from a list of legally required elements

occurrence_i is the number of times keyword i appears in the clause

weight_i represents the severity score assigned to keyword i , and penalty_j is the fixed penalty applied if clause j is missing

This total score is then normalized into a percentage using a domain-calibrated threshold:

$$\text{Risk Percentage} = \min((\text{Total Score} / T) \times 100, 100)$$

Where T is the normalization threshold calculated as:

$$T = \text{Average Weight} \times \text{Expected Matches} \times \text{Number of Risk Categories}$$

Based on the risk score, each clause or document is classified as Low (0–30%), Moderate (31–60%), or High Risk (above 60%), helping users quickly identify and prioritize key concerns.

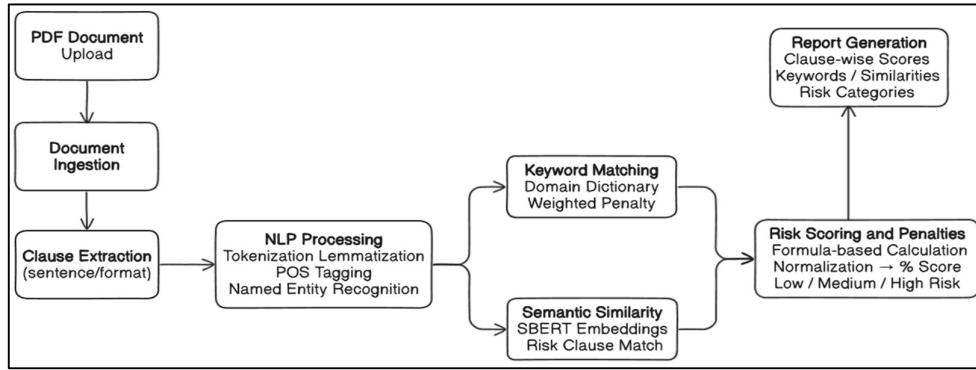


Fig 1. Architecture Diagram

IV. PROPOSED SYSTEM

The proposed system aims to automate the assessment of risks in loan contracts through the integration of rule-based keyword scoring, semantic similarity identification, and omission penalties. As opposed to generic NLP pipelines that are purely classification or extraction-oriented, LegalSutra presents a hybrid framework that is able to detect both explicit and implicit risks and quantify them into actionable scores. The architecture is modular to facilitate flexibility, scalability, and incorporation into current legal review workflows.

A. Document Acquisition and Preprocessing

The pipeline starts with document ingestion, in which loan contracts are uploaded by users in PDF format. With pdfplumber, text is pulled out and cleaned from noise like headers, footers, or formatting artifacts. Cleaned text is tokenized into clauses based on sentence boundaries, punctuation, numbering, and legal links. A clause is processed as an independent analysis unit so that risk assessment can be carried out at a fine-grained level.

B. Linguistic and Semantic Analysis

Following segmentation, every clause is then processed linguistically with spaCy. Tokenization, lemmatization, part-of-speech tagging, and named entity recognition (NER) are applied to extract financial values, dates, organizations, and obligations. Meanwhile, clauses will also be embedded into dense vectors with Sentence-BERT (SBERT). These embeddings will then be matched against a manually curated collection of high-risk clauses so that even when risk words are paraphrased or implied, the system can recognize semantic similarity.

C. Omission Penalties and Risk Scoring

Risk assessment combines keyword-based matching and semantic detection. A risk dictionary specific to the domain gives weights to high-risk words (default, arbitration, foreclosure). The occurrences are scaled by weights to arrive at a raw score. The system also keeps a checklist of boilerplate critical clauses that would be present in generic loan contracts (repayment clauses, resolution of disputes). If the boilerplate clauses are absent, penalty points are incurred. This method ensures that overt risks and latent risks due to absence are identified.

$$\text{Formula: Total Score} = \sum (\text{occurrence}_i \times \text{weight}_i) + \sum (\text{penalty}_j)$$

D. Summarization and Output Generation

To deliver actionable insights at a glance, LegalSutra uses BART for summarizing at the clause level, such that prominent risks are emphasized in simple words. All the results are collated in a PDF report, created using htmltopdf.js, which contains: Clause-wise risk scores and classes, Highlighted keywords and semantic matches, Missing clause penalties, Document overall risk percentage

E. Scalability and Adaptability

The design of the system is scalable, allowing adaptation between various financial institutions and contract types.

With a mere modification of the keyword corpus and penalty rules, LegalSutra can be extended from loan agreements to other financial or regulatory documents. The modular NLP pipeline guarantees adaptability across different document structures without retraining big models.

F. Impact and Utilization

LegalSutra allows organizations to conduct quicker, uniform, and data-driven risk assessment of contracts. It decreases reliance on human review, reduces surveillance in high-volume contract processing, and maintains compliance by identifying missing or unclear clauses. For legal departments, it provides explainability clause-wise; for financial institutions, it improves scalable risk governance. By harmonizing rule-based interpretability with semantic breadth, LegalSutra offers a workable and dependable instrument for contemporary contract risk assessment.

V. COMPARATIVE ANALYSIS

TABLE I. COMPARISON OF SYSTEMS

Title	Authors	Year	Proposed work	Limitation	Our Solution
Effectual Contract Management and Analysis with AI-Powered Technology (ICONSTEM 2024)	M. Fathima et al.	2024	Automates contract management, named entity recognition, OCR	No real-time risk quantification or clause-level scoring	Adds dynamic risk scoring with normalization and interpretable classification
AI-Driven Contract Management (Thomson Reuters/HighQ 2023–24)	Thomson Reuters Legal Solutions	2023	Automates extraction of key data, streamlines contract triage	Focuses on administrative review, lacks explicit risk scoring	Introduces quantified risk metrics (Low/Moderate/High) and semantic-aware clause matching

- [7] S. Ahmed and R. Gupta, "Natural language processing for legal case summarization," in Proc. IEEE Int. Conf. Data Mining, 2023, pp. 789–794.
- [8] T. Nakamura and H. Lee, "Automated legal reasoning using ontology-based systems," IEEE Access, vol. 11, pp. 45678–45685, 2023.
- [9] M. Rossi and G. Bianchi, "Information retrieval techniques in legal document management," in Proc. IEEE Int. Conf. Information Retrieval, 2023, pp. 123–128.
- [10] Y. Zhang and X. Wang, "AI-powered tools for legal compliance monitoring," IEEE Trans. Comput. Soc. Syst., vol. 10, no. 1, pp. 34–42Jan. 2024, doi: 10.1109/TCSS.2024.12345