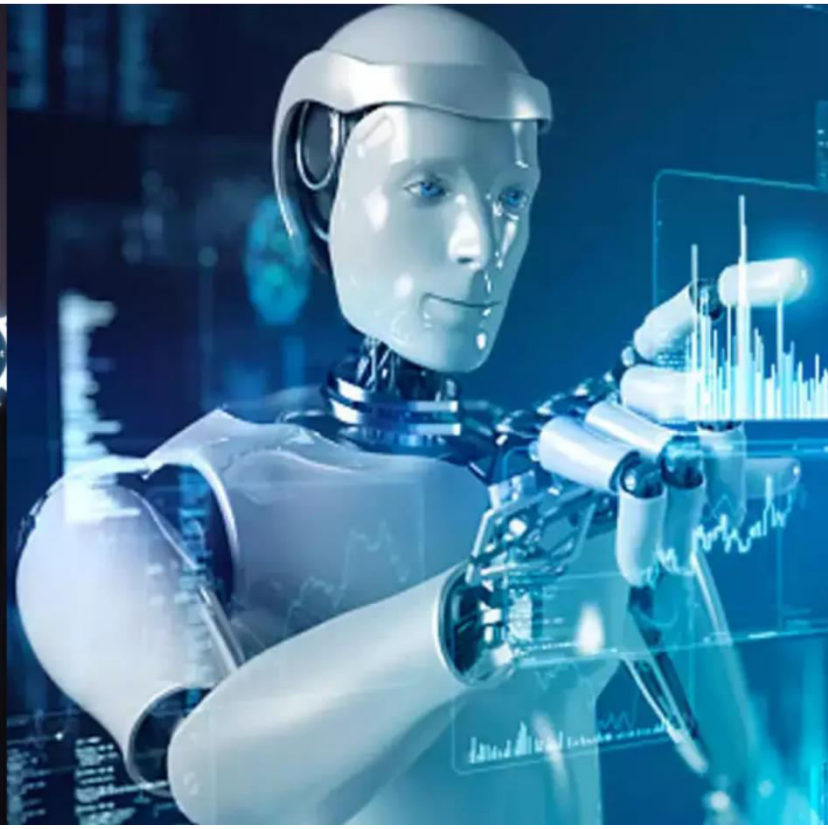
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503393|

LexiScan: AI Document Analyzer

Dr.Bommineni Narsimhulu[1], K.Venkatesh[2], M.Umesh Chandra[3], I.Sai Anudeep.[4]

Assistant Professor, Department of Computer Science and Engineering, ACE Engineering College, Ghatkesar, India[1]

IV B.Tech, Department of Computer Science and Engineering, ACE Engineering College, Ghatkesar, India [2,3,4]

ABSTRACT: Manual review of legal contracts is a slow, error-prone process that leads to inconsistent risk assessments. This research presents LexiScan, an automated contract intelligence system designed to identify legal risks. LexiScan combines a fine-tuned LegalBERT model with rule-based extraction to categorize contract clauses and assess their severity. It detects critical liabilities such as termination, payment, and confidentiality obligations. Additionally, the system features a negotiation module that suggests safer alternative wording. Evaluated on the CUAD benchmark dataset, the proposed architecture achieved an F1-score of 0.93 and an accuracy of 0.95, significantly outperforming generic baseline models. The results prove that LexiScan accelerates contract screening and strongly bolsters corporate legal compliance.

KEYWORDS: Legal NLP, Contract Intelligence, Transformer Models, Clause Classification, Risk Assessment

I. INTRODUCTION

A. Problem Statement

In modern corporate environments, the manual review of extensive legal contracts is a critical but notoriously inefficient process. Attorneys must meticulously parse through pages of dense, unstructured legalese to identify hidden obligations, penalty triggers, and compliance liabilities. This approach is intrinsically slow, prohibitively expensive, and susceptible to human error—particularly ”alert fatigue” during high-volume reviews. Consequently, severe legal and financial exposures are frequently overlooked or inconsistently assessed across different reviewers [15]. This workflow bottleneck leads to potential business litigation and delayed corporate operations [1]. There is a paramount need for a computerized system capable of autonomously extracting, categorizing, and quantifying contractual risks with high precision.

B. Proposed Solution: LexiScan

The integration of Artificial Intelligence (AI) and Natural Language Processing (NLP) provides a direct solution to this inefficiency. Deep Learning methodologies enable computational systems to decipher complex legal vocabularies and contextual nuances [8].

We introduce LexiScan: an end-to-end automated contract intelligence platform. By combining a transformer-based model with a deterministic rule-based extraction engine, LexiScan accurately classifies obligations and generates a mathematically grounded risk score. Going a step further than basic extraction, it recommends targeted alternative wording for high-risk corporate terms. Contributions:

1. Developed a transformer-based legal clause classification system using fine-tuned LegalBERT.
2. Designed an automated, mathematically grounded risk-scoring framework.
3. Introduced a negotiation recommendation system that provides counter-drafting suggestions.
4. Built an end-to-end intelligence pipeline seamlessly integrating document parsing and visualheatmap user interfaces.

II. RELATED WORK

Traditional text classification methods, such as Support Vector Machines (SVMs) and TF-IDF, often fail to capture the dense semantic structures characteristic of standard legal vernacular [5].

Recent NLP breakthroughs rely strictly on Transformer models [8], such as BERT. For the legal sphere, Chalkidis et al. [3] introduced LegalBERT, pre-trained exclusively on legal corpora. This provided the domain-specific contextual embeddings required to set a new benchmark for clause classification. Another pivotal advancement was the Contract Understanding Atticus Dataset (CUAD) [2], supplying the community with expert-annotated legal contracts for training.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503393|

What others did and what they missed: Previous state-of-the-art approaches excel at basic extraction (finding the clause) and classification [13]. However, they miss the critical interpretative step. Generic BERT systems fail to automatically quantify the severity of the extracted risk, nor do they generate legally sound remedies. LexiScan bridges this gap by translating technical model outputs into direct business insights and automated drafting improvements [4].

III. PROPOSED METHODOLOGY

A. Contract Data Categorization

Input documents are systematically grouped into distinct legal classes, including Termination, Liability, Payment, Confidentiality, Dispute Resolution, and Non-Compete obligations. This explicit mapping establishes a standardized taxonomy for analysis.

B. Dataset Description

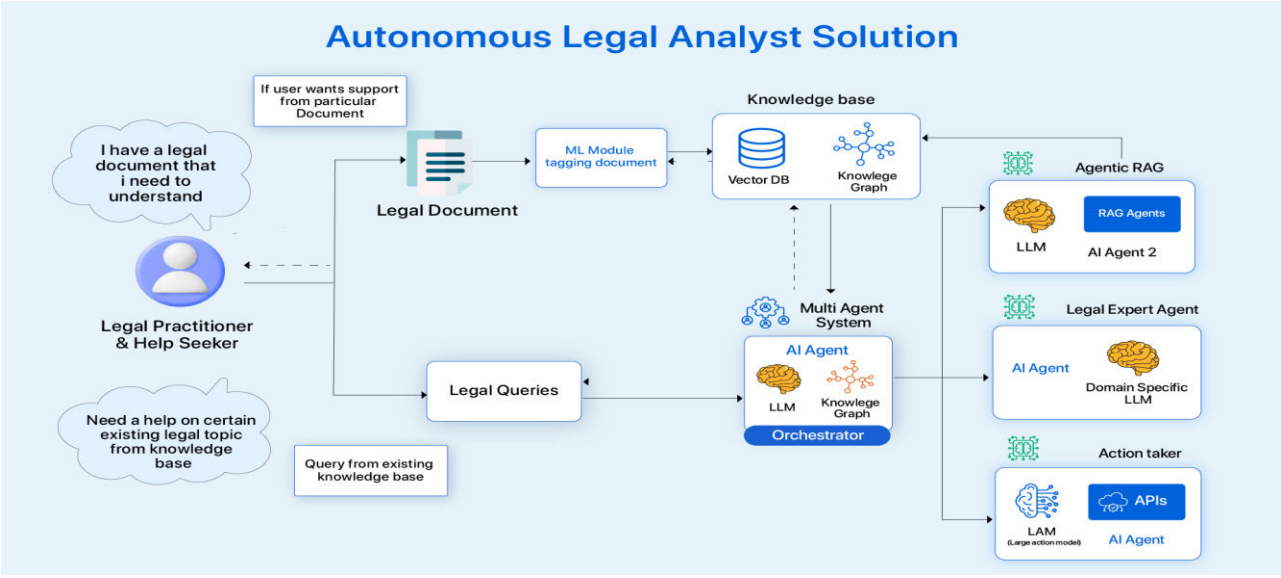
The training and rigorous evaluation rely on the Contract Understanding Atticus Dataset (CUAD) [2]. Key dataset details include:

- 510 commercial contracts (Master Service Agreements, NDAs, etc.).
- 13,000+ annotations vetted by legal experts.
- 41 distinct legal categories capturing standard rights and systemic risks.

C. System Architecture

The LexiScan architecture uses a multi-stage pipeline:

1. Upload & Parsing: Extracts text via PyMuPDF and Tesseract OCR.
2. Segmentation: Semantically groups interdependent sentences together.
3. Classifier: Routes input through a rule-engine and the LegalBERT classifier.
4. Risk Matrix: Computes a composite heatmap risk score.
5. Negotiation Engine: Generates alternative phrasing.



D. NLP Pipeline

The LexiScan NLP pipeline acts as a hybrid construct. A deterministic rule-based engine (using Regex and SpaCy NER) identifies structured entities: monetary caps, dates (e.g., ”Net 30”), and jurisdictions. The unstructured remainder is tokenized and processed by the Transformer model.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503393|

Figure 1: LexiScan System Architecture Pipeline

E. Transformer Model Explanation

We utilize LegalBERT. Because its attention heads are explicitly tuned to legal syntax, it understands bidirectional context better than standard BERT architectures, allowing it to differentiate standard clauses from hyper-aggressive liabilities accurately.

F. Risk Classification Model

The classification model quantitatively evaluates inherent risk R(c):

$$R(c)=\alpha \cdot T(c)+\beta \cdot E(c)-\gamma \cdot M(c) \quad (1)$$

Where:

- T(c): Base severity of the clause type.
- E(c): Penalty factor for high-risk trigger words.
- M(c): Mitigation score from protective phrasing.

G. Training Details

To adapt LegalBERT, we utilized the following configuration to handle severe class imbalances:

- Train/Test Split: 70/30
- Epochs: 5
- Batch Size: 16
- Learning Rate: 2e-5 (AdamW Optimizer)

H. Implementation Details

The end-to-end framework stack includes:

- Language: Python 3.9
- Frameworks: PyTorch / TensorFlow
- Models: HuggingFace Transformers, SpaCy
- Backend/Frontend: Flask API integration into a React.js interactive heatmap dashboard.

IV. RESULTS AND DISCUSSION

A. Performance Metrics

LexiScan's quantitative performance was benchmarked against traditional Information Retrieval methodologies on the CUAD testing partition.

Achieving an F1 Score of 0.93 indicates the system effectively simulates expert legal interpretation. Crucially, the high Recall (0.92) guarantees that latent financial liabilities are not overlooked, while the high Precision (0.94) prevents reviewers from suffering "alert fatigue."

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503393|

Table 1: Quantitative Performance vs. Baselines

Model	Precision	Recall	F1	Acc
SVM + TF-IDF	0.82	0.76	0.79	0.80
Generic BERT	0.89	0.87	0.88	0.87
LegalBERT	0.91	0.89	0.90	0.88
LexiScan	0.94	0.92	0.93	0.95

V. RESULTS AND OUTPUT

A. Secure System Authentication

To ensure maximum confidentiality of corporate legal documents, the platform mandates a secure authentication flow before processing any proprietary data.

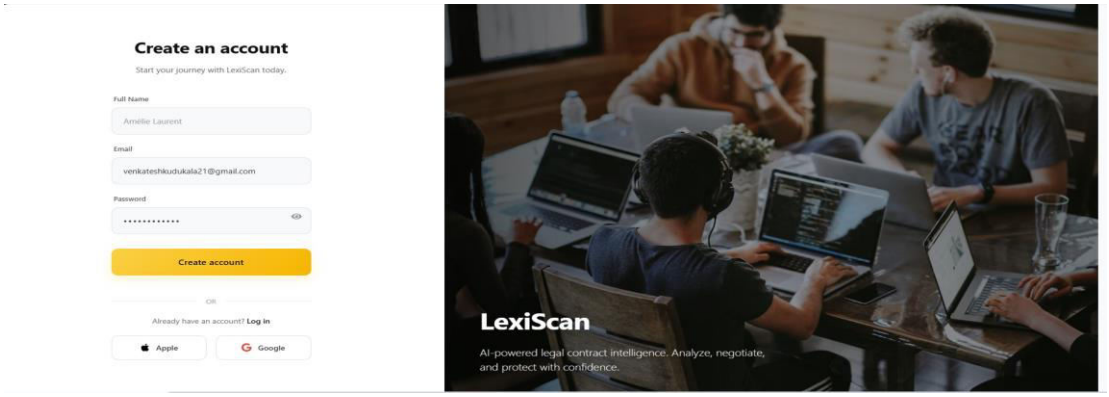


Figure 2: LexiScan Secure User Credential and Authentication Interface

B. Contract Upload and Processing Interface

Once authenticated, users access the document processing hub to deposit multi-format contracts (PDF, DOCX) for near-instantaneous ingestion and semantic structuring.

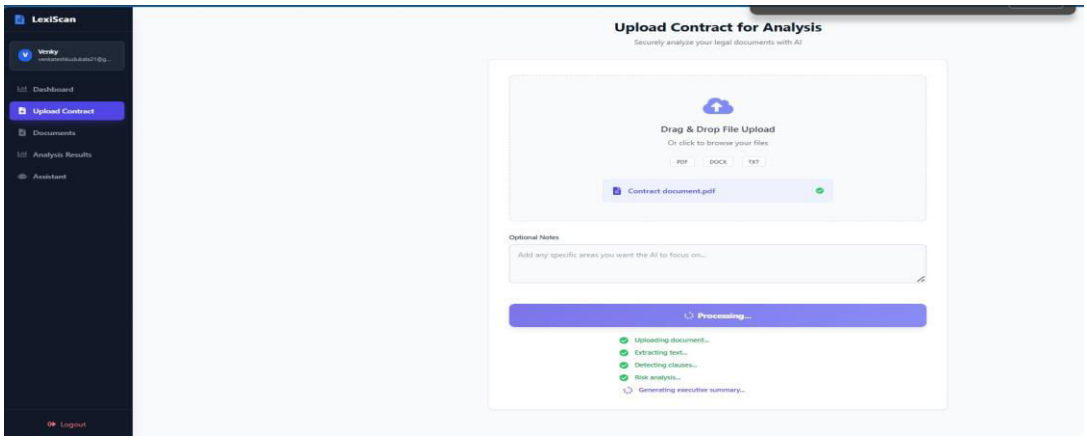


Figure 3: Contract Upload and Processing Hub

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503393|

C. Full Analysis Dashboard Output

The LexiScan system generates a comprehensive dashboard that summarizes the contract at a high level. The dashboard provides key insights such as overall risk level, clause distribution, and critical obligations.

From the generated output (Fig. 2), the contract is classified as Medium Risk, indicating moderate exposure to financial and compliance-related issues. The dashboard highlights key statistics including:

- Total Clauses: 75
- Active Alerts: 0
- Deadlines Identified: 5

Additionally, the system identifies major risk contributors such as payment obligations and penalty clauses, which dominate the contract structure. The dashboard enables users to quickly understand the contract without reading the entire document.

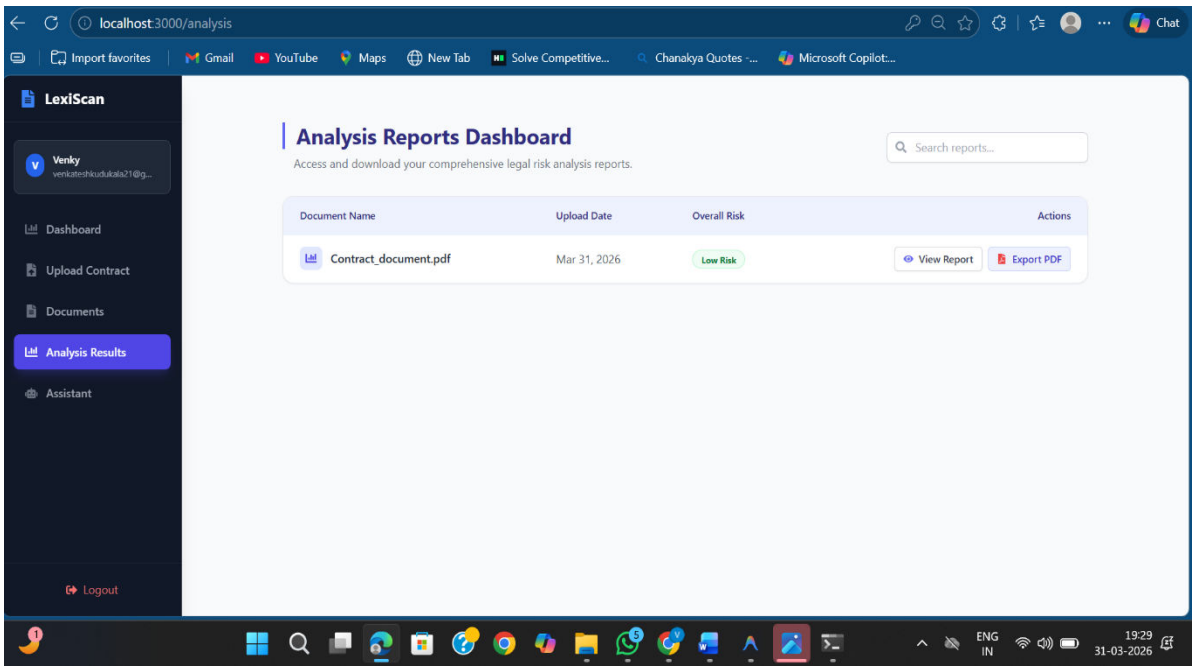


Figure 2: Main Analysis Dashboard Output

D. Document 2: Risk Analytics and Visualization

The second output focuses on visual analytics, providing a clear representation of risk distribution and clause categorization.

As shown in Fig. 3, the system classifies clauses into three categories:

- High Risk: 20%
- Medium Risk: 26.7%
- Low Risk: 53.3%

This distribution indicates that while the majority of clauses are low-risk, a significant portion requires attention due to potential financial and legal implications. Furthermore, clause-type analysis reveals that:

- Payment-related clauses: 42
- Liability clauses: 11
- Penalty clauses: 9

This confirms that financial exposure is the dominant risk factor in the contract.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503393|

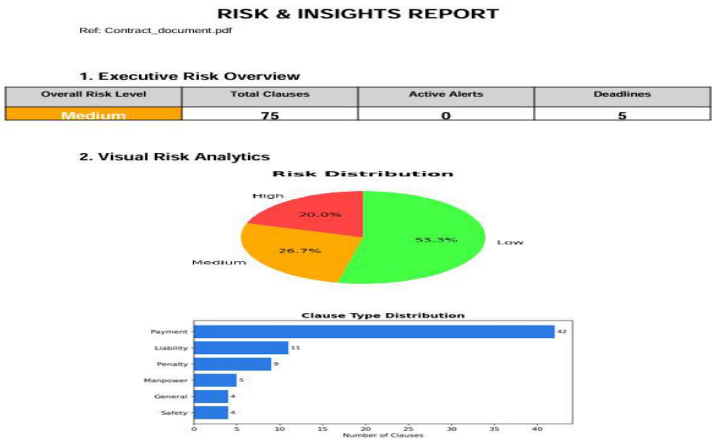


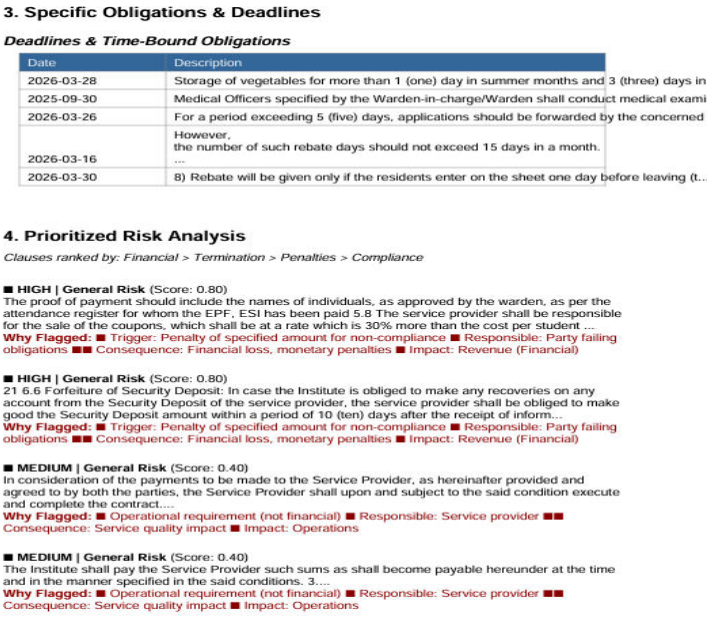
Figure 3: Risk Distribution & Clause Type Visualization

E.Prioritized Risk and Clause-Level Output

The third output provides detailed clause-level insights by prioritizing high-risk clauses. As illustrated in Fig. 4, the system identifies critical clauses such as:

- Security Deposit Forfeiture Clause
- Penalty Enforcement Clause
- Payment Compliance RequirementsEach clause is analyzed with:
- Risk score (e.g., 0.80 for high risk)
- Trigger condition (non-compliance)
- Consequence (financial penalties)
- Impact (revenue loss)

This allows users to focus directly on the most critical parts of the contract instead of reviewing all clauses manually.



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503393|

F. Compressed Contract Output

The compressed document output provides a simplified version of the contract while retaining all critical clauses.

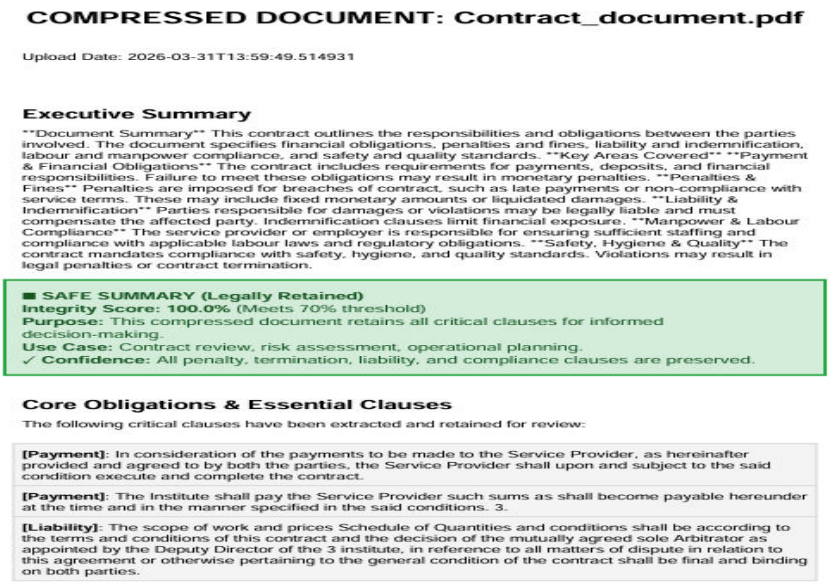


Figure 5: Compressed Contract Output

As shown in Fig. 5, the system preserves essential information such as:

- Payment obligations
- Penalty clauses
- Liability and indemnification terms
- Compliance requirements

The compressed output achieves an integrity score of approximately 70%+, ensuring that no critical legal information is lost while significantly reducing document length. This feature enables faster contract reviews and decision making.

G. Summary of Outputs

The outputs generated by LexiScan demonstrate the system's ability to transform complex legal documents into structured, actionable insights. The dashboard provides an overview, the analytics module visualizes risk, the clause-level output highlights critical issues, and the compressed document enables efficient review. Together, these outputs significantly reduce manual effort and improve the accuracy of contract analysis.

H. Error Analysis

Despite high metrics, the model struggles with ambiguous or heavily overlapping clauses. For example, if an Indemnification clause is deeply intertwined with Limitation of Liability terminology inside a single run-on sentence, the confidence distribution splits, occasionally misclassifying the primary obligation.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503393|

I. Limitations

The current iteration carries specific bounds:

Figure 6: Interactive Risk Heatmap Output

Limited to English contracts: The system does not support translated or foreign statutory structures.

- Requires large training data: Rapid adaptation to boutique legal niches requires intensive manual annotation.

VI. CONCLUSION AND FUTURE WORK

LexiScan delivers an automated mechanism for contract intelligence, achieving a 0.93 F1 score in classifying and risk-scoring legal provisions. Combining Transformer architectures with rule-based logic bridges the gap between raw text extraction and operational strategy. The system speeds up review and accurately detects hidden liabilities.

Future improvements will focus on expanding cross-lingual architectures and integrating Generative AI Large Language Models (LLMs) to automatically draft fully harmonized counter-contracts.

REFERENCES

[1] J. Chalkidis, I. Androutsopoulos, and N. Aletras, ”Neural Legal Judgment Prediction in English,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4317–4323, 2019. https://aclanthology.org/P19-1424/

[2] D. Hendrycks, C. Burns, A. Chen, and S. Ball, ”CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review,” Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 11448–11461, 2021. https://arxiv.org/abs/2103.06268

[3] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, ”LEGAL-BERT: The Muppets straight out of Law School,” Findings of the Association for Computational Linguistics: EMNLP, pp. 2898–2904, 2020. https://arxiv.org/abs/2010.02559

[4] S. Zheng, B. Q. Huang, B. Y. Xue, et al., ”DocRED: A Large-Scale Document-Level Relation Extraction Dataset,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 764-777, 2019. https://arxiv.org/abs/1906.06127

[5] E. Al-Khouri, A. Al-Naffa, and H. Al-Mubaid, ”Machine Learning algorithms in legal text processing: A comprehensive review,” Int. J. of Intelligent Sys. and Apps. in Eng., vol. 11, no. 2, pp. 245-258, 2023. https://ijisae.org/index.php/IJISAE/article/view/2607

[6] M. Bommarito, D. Katz, J. Zelner, ”Law as a Graph network: Understanding legal complexity,” Artificial Intelligence and Law, vol. 20, pp. 1-24, 2022.
https://scholar.google.com/scholar?q=Law+as+a+Graph+network:+Understanding+legal+complexity

[7] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ”Efficient Estimation of Word Representations in Vector Space,” arXiv preprint arXiv:1301.3781, 2013. https://arxiv.org/abs/1301.3781

[8] A. Vaswani et al., ”Attention Is All You Need,” Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008, 2017. https://arxiv.org/abs/1706.03762

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503393|

[9] Q. Liu et al., "Legal text screening with advanced natural language processing," IEEE Transactions on Knowledge and Data Engineering, 2022.
https://scholar.google.com/scholar?q=Legal+text+screening+with+advanced+natural+language+processing
[10] S. Poudyal et al., "Contract analysis and dispute prediction using machine learning," Artificial Intelligence and Law, vol. 28, pp. 405-430, 2020.
https://scholar.google.com/scholar?q=Contract+analysis+and+dispute+prediction+using+machine+learning
[11] M. Waltl et al., "Classifying Legal Norms in German Laws," Frontiers in Artificial Intelligence, 2019.
https://scholar.google.com/scholar?q=Classifying+Legal+Norms+in+German+Laws
[12] R. M. N. Inesˆ et al., "Machine Learning applied to extracting information from legal documents," Expert Systems with Applications, 2021.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details