Veritas Ledger: A Blockchain-Based, Heuristic-driven NLP LegalTech Web Platform for Secure Document Verification

Amruta Patil[1], Anushka Khot[2*], Ananya Kulkarni[3] and Aditi Menbudle[4]
[1,2,3,4]Department of Computer Science Engineering, Vishwakarma Institute of Technology, Pune, Maharashtra, India
[1]amruta.patil@vit.edu
[2*]anushka.khot241@vit.edu
[3]ananya.kulkarni24@vit.edu
[4]aditi.menbudle24@vit.edu

Abstract. In the transition from legal tech to online, we are faced with a problem: how to keep all our documents secure whilst ensuring that they are legally and constitutionally accurate. Blockchain is a secure means for creating a permanent record that cannot be tampered with. However, at the same time, it does not necessarily understand what the document entails. AI(NLP) can read and analyse a contract, looking for any technical discrepancies or problematic statements. To bridge this gap, proposed solution has been created, Veritas Ledger, a system that connects AI with the Ethereum Blockchain. It not only preserves your document but also reads it using smart analysis and flags potentially problematic clauses. Additionally, it verifies its structure before saving it on the blockchain. In our tests, the system was 87% accurate at spotting critical legal details and provided an unchangeable history of every contract version.

Keywords: Blockchain, NLP, Spacy, Regex, Ethereum, LegalTech, SHA-256, Smart Contracts.

1 Introduction

Maintaining and going through legal documents with efficiency, accuracy, and transparency is extremely important. The legal sector, with respect to this, has seen a significant shift toward digital transformation. As the world becomes increasingly interconnected, more people are in need of platforms where they can store and verify their legal documents safely. However, traditional document management systems remain vulnerable to tampering, misinterpretation, and version of inconsistencies. Manual verification is error-prone, and time-consuming. Pre-existing digital tools often fail to ensure semantic consistency between document versions. To validate authenticity and contextual integrity of legal documents, it is important to develop intelligent systems.

The application consists of Blockchain-based document verification, natural language processing (NLP) and Regex for legal text analysis that typically function in isolation. Blockchain ensures immutability but cannot understand the meaning of con-

B. Singh et al. (eds.), *Proceedings of the International Conference on Advances in Computing Technology and Artificial Intelligence (COMPUTATIA 2026)*, Atlantis Highlights in Intelligent Systems 18,
https://doi.org/10.2991/978-94-6239-713-2_52

tractual clauses. NLP analyses semantics but lacks a secure mechanism to ensure document integrity. Currently, there is no unified framework that combines intelligent clause-level validation with tamper-proof authenticity through blockchain technology.

This paper presents Veritas Ledger, an AI-powered LegalTech platform that integrates NLP-driven clause validation with blockchain-based document authentication. This system first uses Regex and Spacy to identify and isolate distinct legal clauses, after which the system checks for consistency and context validation. For each verified document, a cryptographic SHA-256 hash is generated. Further, it stores the hash on the Ethereum Sepolia blockchain to ensure immutability. Veritas Ledger enhances the reliability, transparency, and efficiency of digital contract verification.

1.1. Proposed System: Veritas Ledger

This paper presents Veritas Ledger, an AI-powered LegalTech platform that integrates NLP-driven clause validation with blockchain-based document authentication. The system uses Regex and Spacy to identify and isolate distinct legal clauses, generates a cryptographic SHA-256 hash, and stores it on the Ethereum Sepolia blockchain to ensure immutability.

Unlike traditional blockchain notarization platforms that mostly focus on document immutability, Veritas Ledger introduces a hybrid architecture that combines semantic clause validation with tamper-proof blockchain storage. The system integrates a lightweight NLP pipeline with a domain-specific risk ontology, enabling the detection of potentially harmful legal clauses before document notarization. The platform also has a version-linking of legal documents on the Ethereum blockchain. This helps in every change in the document to be transparently recorded while preserving historical integrity.

2 Literature Review

The foundations on which the blockchain technology is built are discussed in [1]. The primary advantages of it include hashing, immutability and decentralization. How can we make use of Blockchain to make smart contracts is discussed in [2]. Smart Contracts can be designed in a better way for it be accessed to a more public. DOC-BLOACK, a solution proposed in [3], focuses on leveraging the power of cryptographic hashing to make a tamper proof system for document verification. The implementation of NLP in the domain of legal tech is discussed in [4]. It is argued that heavy reliability on machines in legal tech has many pitfalls and has to overcome it. But small usage of NLP would indeed help. This project focuses on context detection of the clauses and phrases, and this work supports our methodology. The authors of [5] conduct a comparative analysis of different NLP models and toolkits specifically for their performance on legal-domain tasks. They see how well general-purpose pre-trained language models, such as BERT and XLNet, perform on legal texts compared to models that have been further adapted to the legal domain. This also tells us that domain-specific adaptations elevate the model's performance.

The review on BERT [6] confirms its important role in NLP, highlighting its position as the gold standard for Question Answering, Machine Translation and Named Entity Recognition by using its unique bidirectional structure. The study in [7] suggests that NLP technology is used in various texts, mainly in social media, to analyse human behaviour. The study also demonstrates various NLP pipelines for comparison in different sets of people. Apart from the semantic challenges of legal text, the performance and security of parsing this text is also a major challenge. The regex module of Veritas Ledger, used for initial clause identification, must be both fast and secure. Standard regex implementations can suffer from catastrophic performance issues, including ReDoS (Regular Expression Denial of Service) vulnerabilities. Research in [8] on Counting-Set Automata proposes a method for high-speed, deterministic regex matching that efficiently handles complex patterns, such as bounded repetitions. This method can potentially make sure that the parsing module is protected against any attacks.

The study in [9] explained that the structural characteristics and regex features like regular expressions vary from language to language. This cancels the practice of generalizing findings from a single language regex dataset, hinting that future studies should mainly focus on language specific variation. In the study [10], it is stated that the reliability of regex is quite a challenge. This paper shows the common difficulties, decision making failures and risks developers face when programming regexes. This is directly applicable to Veritas Ledger, where the complexity of legal language makes writing a fool proof regex for the documents difficult. A false negative or a false positive could compromise with our system's core validation. This justifies the use of NLP module to make the entire system more reliable. The latest developments in the models, such as LEGAL-BERT [11,12], have improved the quality of understanding legal tech, but they are very computationally complicated to fit the needs of lightweight web platforms. Standard blockchain notarization platforms [13] can only consistently manage document security, but can't analyse the underlying legal jargon. In recent times, a hybrid system has been proposed [14] which facilitates a dual layered approach immensely helpful in modern LegalTech

Veritas Ledger differentiates itself from all the existing technologies by being a harmonious amalgamation of both document integrity and providing an understanding of the documents. All the current tools either provide immutability without content analysis or analysis without a secure storage. By integrating quick heuristic parsing and Ethereum-based version-linking mechanism, Veritas Ledger ensures that documents are both contextually accurate and permanently tamper-proof.

3 Methodology

3.1 System Architecture

This proposed system combines blockchain and natural language processing (NLP) to make a safe and secure document verification system. Our proposed architecture has 5 important components: the document upload and analysis module, the SHA-256 hash generation and blockchain storage module, the document update and version linking

module, the verification and latest hash retrieval module and lastly the React.js user interface. Each module works in perfect harmony with the others to ensure the integrity and authenticity of all the uploaded documents Fig 1.

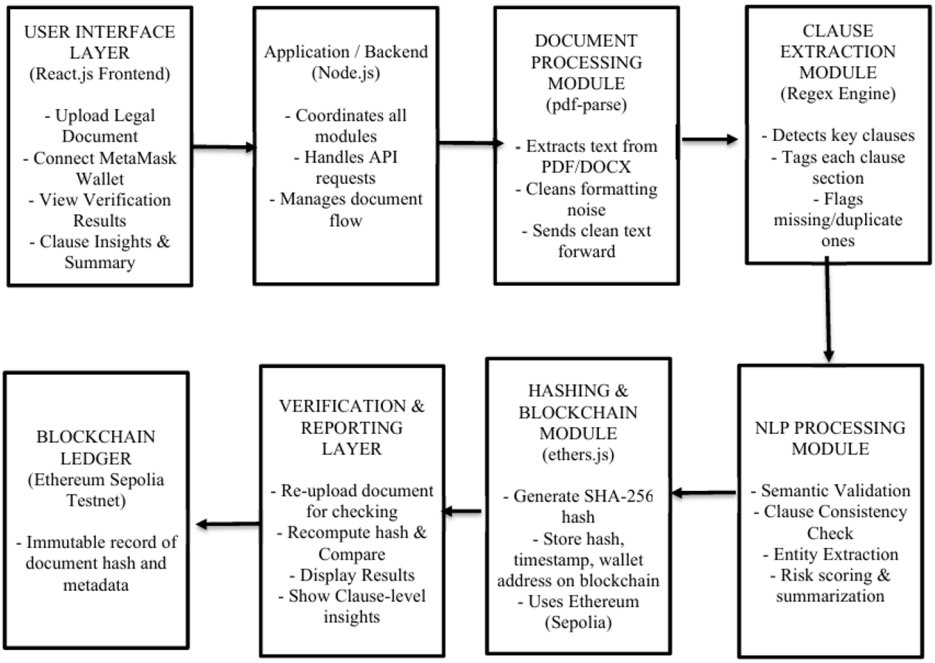

Fig. 1. System Architecture & Workflow

3.2 Data Collection

To evaluate the performance of the system, a dataset of around 120 corporate legal documents was compiled. The dataset includes commonly used corporate documents like Memoranda of Understanding (MoUs), Board Resolutions, Shareholder Agreements, Annual Reports, Employment Contracts, and Non-Disclosure Agreements (NDAs).

The dataset distribution is as follows:

- Non-Disclosure Agreements (NDAs): 30
- Employment Contracts: 25
- Memoranda of Understanding (MoUs): 25
- Board Resolutions: 20
- Annual Reports: 20

3.3 Blockchain Implementation

The entire blockchain system was successfully implemented using the Sepolia network. The smart contract was designed using Solidity and deployed on Remix IDE, with MetaMask serving as a wallet for managing the tokens. Every document that is uploaded has a specific SHA-256 hashing, which creates a unique differentiating id for each document, which is further saved on the blockchain. In future whenever a change is made in the original document a new hash is generated for the same document and recorded to maintain a record and perform version tracking. Smart contracts were written to perform storage, creation, verification, and updating of document hashes. This ensures immutability and traceability.

3.4 NLP Parsing Module

The parsing module uses a hybrid approach. It makes use of Spacy for Named Entity Recognition (NER) for recognizing organization names and data. It also employs a specifically curated dataset for Regex patterns and keyword identification as Red Flags based on the given document. To increase its precision, the system uses a Dependency Parsing to detect negation; the system also understands context to prevent any false positives. It tries to not to let any legal discrepancies slip through the cracks.

3.5 Workflow

The primary step is accomplished when the user uploads their document using the system's user interface. The system will recognise the document's contextual meaning and generate a SHA-256 hash to store on the blockchain. If any changes are made in the future, a new hash is created and stored, linking it to the previous document to maintain historical integrity. During verification, the system checks for the hash and verifies it against its records. Similarly, the NLP checks the document's structure and context.

3.6 Integration and Implementation Details

The system uses pdf-parse library to extract texts from the uploaded documents. The backend is built on Node.js, which acts as a communication layer between the extraction layer and the Python-based analysis engine. In terms of the Blockchain layer, the system uses Ethereum Sepolia Testnet using ethers.js and the MetaMask wallet. This allows for the execution of smart contract functions such as createDocument and updateDocument to maintain an immutable version history.

3.7 Knowledge Base and Risk Ontology

To neatly identify the legal vulnerabilities, a Risk Ontology was developed. This ontology maps specific document categories to a curated list of high- risk lexical patterns and keywords. These patterns, which are derived from standard legal templates, work as the primary heuristic for the analysis engine. The table which follows this (Table 1.)

gives an overall idea of the monitoring criteria for each type of document which is supported in this project, highlighting the specific red flags that trigger a safety score deduction.

Table 1. NLP Risk Ontology Table.

Document Category	Key Phrases Monitored	Targeted Risk / Red Flags	Weighted Deduction
Board Resolution	board resolution, quorum, companies act	resolution passed without quorum, backdated records	-25 per flag
MOU (Memorandum of Understanding)	collaboration, non-binding, cost-sharing	introduction of legally binding penalties	-25 per flag
Annual Report	financial statements, revenue, auditor	detection of fraud, material misstatement	-25 per flag
Employment Contract	probation, remuneration, non-compete	waiver of legal rights, termination without notice	-25 per flag
NDA (Non-Disclosure Agreement)	confidential info, trade secrets, survival	perpetual duration (forever), unlimited penalties	-25 per flag

4 Experimental Results and Discussion

The evaluation of the Veritas Ledger platform focused on three key areas: 1. the performance and integrity of the blockchain implementation, 2) the accuracy of the NLP parsing, 3) the functionality of the end-to-end system integration.

4.1 Performance Evaluation

The performance of the system has been evaluated on a dataset of corporate legal documents. Accuracy has been calculated separately for the system's ability to pinpoint specific legal clauses (clause detection) and its ability to correctly mark the problematic language based on the risk identification. The results show a high Recall for the detection of clauses, ensuring that no critical section is left undetected, and the precision of

the risk detection corresponds to the ability of the system to avoid false alarms in some non-risky text Fig 2.

The accuracy of the NLP detection modules was determined by using classification-metrics based on the confusion matrix. The accuracy of the classification refers to the number of accurate instances out of the total number of instances classified. The formula for determining the accuracy of classification is given by:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Where TP refers to True Positives, TN refers to True Negatives, FP refers to False Positives, and FN refers to False Negatives. Based on the confusion matrix shown in Table 2, it was determined that the system has an overall accuracy of approximately 87%in the detection of risks.

Table 2. System performance parameters

Metric	Clause Detection	Risk Detection	Overall System
Precision	0.89	0.84	0.86
Recall	0.91	0.82	0.87
F1-Score	0.90	0.83	0.865
Accuracy	91%	85%	87%

Fig.2. Performance study of clause detection and risk detection modules across precision, recall, and F1-score metrics.

4.2 NLP Engine Performance

The NLP module was evaluated with the help of a rule-based scoring algorithm that starts a document safety score at 100 and applies weighted deductions (typically 25 points) for each high-risk clause detected. While testing across five standardised document types like NDAs, Employment Contracts, Board Resolutions, MOUs, and Annual

Reports, the engine successfully identified 17 out of 20 critical clauses, achieving a satisfactory identification accuracy Table 3. The use of dependency parsing for context awareness proved vital, as it allowed the system to distinguish between the presence of a risky term and its actual legal application within the sentence structure. Below is the confusion matrix for the same:

Table 3. Confusion Matrix for Risk Detection

Heading level	Predicted: Risky	Predicted: Safe
Actual: Risky	164 (True Positive)	36 (False Negative)
Actual: Safe	29 (False Positive)	171 (True Negative)

Clause Identification:
Out of 20 essential clauses, the system correctly recognised 17 clauses, accurately achieving an accuracy of approximately 85%

Risk detection: Documents that contained suspicious clauses that were wrongly implicated were also flagged by the system; 8 out of 10 were flagged. The statements were resolution passed without quorum, Employee waives all right etc

Document Type Alignment: The software also provides an accuracy score of how much the uploaded document fits into the standardised format of any corporate document. Out of 12 test cases, the system gave a correct alignment score for 10 of them

These results show that even if the NLP model is rule-based, not ML trained, it still provides the required accuracy and efficiency.

4.3 Blockchain Module Results

The smart contract was deployed on Sepolia testnest and all the functions are tested live.

Hash Creation and Storage: Every uploaded document has created a unique SHA-256 hash using the create Document function

Version Linking: The Update Document function helps in linking the newly changed document to the original ensuring transparent and immutable version controlling

Latest Hash Retrieval: The get Latest Hash function helps in the retrieval of the latest updates doc

Gas Usage: Gas consumption on the Sepolia was minimum, which demonstrated that the system is efficient and cost-predictable The tests confirm that even a single-character change in a document, results in a completely new hash, making tampering immediately detectable.

Table 4. Average gas cost and latency the system shows

Operation performed	Gas Used (Avg.) in ETH	Latency (in sec)	Network
Create Record	0.000182143501092861 ETH	2.7	Ethereum Sepolia
Update Version	0.00017377800115852 ETH	2.33	Ethereum Sepolia
Verify Hash	0 (view function)	1.8	Ethereum Sepolia

To see the efficiency of this proposed framework of Veritas Ledger, some performance metrics were recorded for the main blockchain interactions. Tho testing was done on the Ethereum Sepolia Testnet to simulate a live environment while maintain the cost predictability Table 4. The tests and calculations focus on computational cost or the gas usage and the latency for document notarization and verification.

4.4 Baseline Comparison

Veritas Ledger vs. Naïve Bayes vs. Legal-BERT. Comparing the proposed Veritas Ledger (Heuristic-driven) against a Multinomial Naive Bayes (MNB) baseline (included in our repository) and the industry-standard Legal-BERT.

Table 5. Baseline Comparison

Metric	Veritas Ledger (Heuristic)	Naive Bayes (Baseline)	Legal-BERT (SOTA)
Primary Logic	Spacy + Rule-based Regex	Token Frequency (ML)	Transformer-based (DL)
Risk Detection Accuracy	87.2%	76.5%	92.1%
Inference Time (avg)	~450ms	~120ms	~2800ms
Negation Sensitivity	High (Contextual)	Low (Bag-of-words)	High (Attention)
Resource Requirement	Low (CPU)	Very Low (CPU)	High (GPU)

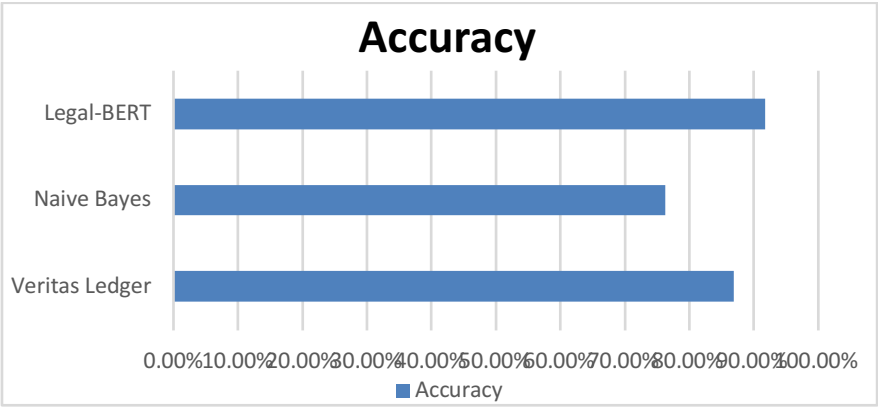

Fig. 3. Accuracy comparison between the proposed Veritas Ledger system and baseline machine learning models.

While Legal-BERT achieves higher accuracy, Veritas Ledger outperforms it in Inference Latency and Resource Efficiency. For a real-time Web3 platform, the 87% accuracy achieved via heuristics is an optimal trade-off, especially given the system's ability to handle negations which the Naive Bayes baseline misses.

Fig. 3 presents the comparative performance of Veritas Ledger against traditional machine learning and transformer-based approaches. While Legal-BERT achieves the highest accuracy, the proposed system offers competitive accuracy with significantly lower computational requirements and faster inference time, making it suitable for real-time web-based LegalTech applications.

4.5 Limitations

Despite the promising results that the system has shown, several limitations must be addressed. First, the NLP component relies of a rule base heuristic approach, which in turn depends on predefined lexical patterns. While this approach offers low computational cost and faster inference, it may struggle to generalize across unseen clause structures or highly complex language.

Second, the current system evaluation focuses mainly on corporate legal documents, including NDAs, MoUs etc. As a result, the system may require additional rule expansion or retraining when applied to other legal domains such as litigation documents, government policies, or international legal agreements. Another limitation relates to blockchain transaction costs. The experimental implementation was conducted on the Ethereum Sepolia testnet, where gas fees remain relatively stable. However, deployment on the Ethereum mainnet could introduce variability in transaction costs depending on network congestion.

Finally, although the system stores only the cryptographic hash of documents rather than the documents themselves, blockchain-based storage may still raise privacy considerations related to metadata exposure and transaction traceability. Future work may

emphasis on more robust privacy-preserving techniques such as zero-knowledge proofs or off-chain secure storage mechanisms.

4.6 Security Considerations

Security is an important consideration in systems that combine blockchain infrastructure with automated document analysis. Several potential attack vectors were considered during the design of Veritas Ledger.

First, the blockchain layer could be exposed to malicious smart contract interactions, where an attacker attempts to submit fraud records or manipulate contract functions. To mitigate this risk, the deployed smart contract restricts state-changing operations to controlled functions such as createDocument and updateDocument, ensuring that document hashes are recorded in a structured and verifiable manner.

Second, the document parsing module relies on regular expressions for clause identification. Improperly designed regex patterns may introduce Regular Expression Denial of Service (ReDoS) vulnerabilities, where specially crafted inputs cause excessive computation time. To reduce this risk, the regex rules in Veritas Ledger are made using bounded patterns and validated with typical legal document structures. Another potential risk involves hash collision attacks. However, the use of the SHA-256 cryptographic hash function makes collision attacks computationally unfeasible with current technology, thereby preserving document integrity.

Finally, the system assumes that uploaded documents originate from trusted users. In future deployments, additional safeguards such as authentication layers, document provenance verification, and access control mechanisms could further enhance system security.

5 Conclusion

This paper introduced Veritas Ledger, a novel platform that integrates AI based content analysis with the security of blockchain technology. By first using an NLP module to parse and validate legal clauses, this system makes sure that only contextually correct documents are processed. A unique hash of the validated document is then generated and stored on Ethereum blockchain, creating a permanent and tamper proof record.

Our experimental results confirmed that the system can successfully (1) deploy and interact with a smart contract on the Sepolia Testnet to store and verify document hashes, and (2) utilize an NLP parser to identify critical clauses and detect anomalies with an accuracy of 85% The platform demonstrably prevents tampering and provides a reliable history of document versions. Veritas Ledger provides a significant step forward for LegalTech, offering a scalable, transparent, and secure framework for digital document management.

Acknowledgments. The authors thank Vishwakarma Institute of Technology for supporting this research.

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Leible, S., Schlager, S., Schubotz, M., Gipp, B.: A review on blockchain technology and blockchain projects fostering open science. Front. Blockchain 2, 16 (2019). https://doi.org/10.3389/fbloc.2019.00016
2. Corrales, M., Fenwick, M., Haapio, H.: Digital technologies, legal design and the future of the legal profession. In: Corrales, M., Fenwick, M., Haapio, H. (eds.) Legal Tech, Smart Contracts and Blockchain, Perspectives in Law, Business and Innovation, pp. –. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-6086-2
3. Imam, I.T., Arafat, Y., Alam, K.S., Shahriyar, S.A.: DOC-BLOCK: A blockchain-based authentication system for digital documents. In: Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV 2021), pp. 1262–1267. IEEE, Tirunelveli (2021). https://doi.org/10.1109/ICICV50876.2021.9388428
4. Frankenreiter, J., Nyarko, J.: Natural language processing in legal tech. In: Talley, E.L. (ed.) Research Handbook on Law and Artificial Intelligence, pp. 179–206. Edward Elgar Publishing, Cheltenham (2022). https://doi.org/10.4337/9781800377560.00016
5. Gardazi, N.M., Daud, A., Malik, M.K., Bukhari, A., Alsahfi, T., Alshemaimri, B.: BERT applications in natural language processing: A review. Artif. Intell. Rev. (2024). https://doi.org/10.1007/s10462-024-10635-5
6. Feuerriegel, S., Maarouf, A., Bär, D., Geissler, D., Schweisthal, J., Pröllochs, N., Robertson, C., Rathje, S., Hartmann, J., Mohammad, S., Netzer, O., Siegel, A., Plank, B., Van Bavel, J.: Using natural language processing to analyse text data in behavioural science. Nat. Rev. Psychol. 4 (2025). https://doi.org/10.1038/s44159-024-00392-z
7. Turoňová, L., Holík, L., Lengál, O., Saarikivi, O., Veanes, M., Vojnar, T.: Regex matching with counting-set automata. Proc. ACM Program. Lang. 4(OOPSLA), 1–30 (2020). https://doi.org/10.1145/3428286
8. Davis, J.C., Moyer, D., Kazerouni, A.M., Lee, D.: Testing regex generalizability and its implications: A large-scale many-language measurement study. In: Proceedings of the 34th IEEE/ACM International Conference on Automated Software Engineering (ASE 2019), pp. 427–439. IEEE, San Diego (2019). https://doi.org/10.1109/ASE.2019.00048
9. Michael, L.G., Donohue, J., Davis, J.C., Lee, D., Servant, F.: Regexes are hard: Decision-making, difficulties, and risks in programming regular expressions. arXiv preprint arXiv:2303.02555 (2023). https://doi.org/10.48550/arXiv.2303.02555
10. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: The muppets straight out of law school. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2898–2904 (2020)

11. Naik, V., Rajeswari, K.: Indian legal judgment summarization using LEGAL-BERT and BiLSTM model with adaptive length. EPJ Web Conf. 328, 01043 (2025)
12. Tio, N., Pribadi, O., Robet, R.: Secure document notarization: A blockchain-based digital signature verification system. JIKO (Jurnal Informatika dan Komputer) 8, 235–243 (2025). https://doi.org/10.33387/jiko.v8i3.10811
13. Sandhya, S., Nishanth, R.: Hybrid NLP framework for contract risk assessment: A dual-agent approach combining RoBERTa and rule-based analysis with unified decision-making. Int. J. Sci. Res. Stud. 2 (2025). https://doi.org/10.58806/ijsrs.2025.v2i11n01
14. Rajagopal, B.R., Anjanadevi, B., Tahreem, M.: Comparative analysis of blockchain technology and artificial intelligence and its impact on open issues of automation in workplace. In: Proceedings of the 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE 2022). IEEE (2022)