# CodifiedCant: Enhancing Legal Document Accessibility Using NLP and Longformer for Secure and Efficient Compliance

Jayapradha J[1], Su-Cheng Haw[2]*, Naveen Palanichamy[3], Nilanjana Bhattacharya[4], Aayushi Agarwal[5], Senthil Kumar T[6]

Department of Computing Technologies-School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu 603203, India[1, 4, 5, 6]

Faculty of Computing and Informatics, Multimedia University, Jalan Multimedia, 63100 Cyberjaya, Malaysia[1, 2, 3]

***Abstract*—CodifiedCant is a new idea that employs Natural Language Processing to simplify company guidelines and legal documents. Legal texts are extensive, complicated and hard for non-experts to understand. To tackle the above problem, this research incorporates the Longformer model because it functions as a transformer-based deep learning system designed to work effectively with extensive legal documents. Longformer enables the system to handle extensive documents by keeping better track of context, which results in transforming complex legal text into easily readable formats. To enhance the search and retrieval speed, this research investigates the nuances of transforming unstructured data, like tabular data from PDFs, to vectors. This revolution supports quicker, cognisant semantic routing inside the document. Further, it assists in data arrangement and detection across massive sources of legitimate and business information. Data security is also a major priority for the platform, which utilizes network encryption to protect data and privacy. CodifiedCant is a scalable, secure and intelligent solution for better employee access to legal news, greater company transparency and reinforces better compliance in the organization. Table extraction and document simplification performance of the model are validated on Cornell LII and Kaggle evaluation datasets, respectively. CodifiedCant associates the variance relating to legitimate terminology and user knowledge.**

***Keywords*—Natural language processing; transformer-based deep learning system; long former; semantic routing; network encryption; legal document; unstructured data; and data security**

## I. Introduction

Legal documents are a fundamental aspect of everyday lives, influencing everything from business transactions and government policy to personal contracts such as wills and agreements. They create a clear definition of rights and obligations, which avoids misunderstandings and settles disagreements when they occur. Whether a contract between two businesses, legislation enacted by a government, or a court judgment, these documents guarantee that laws are obeyed, and promises are kept. But legal documents tend to be drafted in language that is arcane to the common individual. They depend upon sophisticated vocabulary, or legalese, and rigid formatting conventions that are country- and institution-specific. Although this degree of specificity ensures consistency and enforceability, it can also render legal documents off-putting, particularly to non-lawyers. Consequently, most individuals find it difficult to navigate through contracts, terms of service, or even simple legal agreements without the assistance of professionals. As legal and corporate documents expand in complexity, so does the need for tools to analyze, summarize, and extract key information [1]. For people or organizations, legal documents are tedious and complex for many people, full of legal terms and complex language - an issue so common that the phrase 'The person has read and agrees with the terms and conditions' is in fact, the most abused statement on the Internet. Since every application and every page has its own Terms of Services, Privacy Policies and End User Licence Agreements, a survey suggests that to read such material, on average, every user would need more than 200 hours on a yearly basis, which makes most of them only click the «agree» button without bothering to read.

### A. CodifiedCant: A Deep Learning Solution for Legal Document Analysis

This widespread issue is being solved with the development of the idea's work, that is, a deep learning based advanced chatbot modified from the Longformer model that is prepared to understand long texts such as legal and corporate documents. It is able to condense long paragraphs into short ones, express thoughts, insights and reasons, with pointers highlighting critical aspects and cross examinations, making it easier to cope with complex words, rules and systems. The chatbot gives the user access to such information, making it easier for them to understand and act upon. One of CodifiedCant 's primary functionalities is to ease the difficult challenge of transforming unstructured data [2], e.g., tables in PDF files, into vector representations to enhance search and retrieval. This feature enables access and interpretation of tabular data, frequently serving as a core component of legal documents. This work proposes a new way of analyzing a document using Natural Language Processing (NLP) tools [3], by applying the Longformer model [4]. Due to its capability to quickly process lengthy documents, classic NLP models often have trouble processing, and Longformer has been particularly pertinent for legal and corporate texts. This distinguishing functionality fetches the information from large pieces of text and offers a deep understanding of contract wording, interests and terms. To enhance its work, strict measures have been taken to protect private data during the implementation of the system. Line encryption [5] protects data in motion to ensure the

confidentiality and integrity of every interaction. In addition, Amazon Web Service (AWS[1] Identity and Access Management (IAM) [6] allows fine-grained access control, dictating who can access the chatbot resources. Furthermore, AWS Web Application Firewall (WAF)[2] offers anti-DDoS [7] protection, helping to protect the chatbot against common web threats [8] and making it resilient against attacks[3]. The proposed work empirically evaluated the performance of CodifiedCant on extracted datasets created from actual court documents, including those from Cornell LII [4] and Kaggle[5], mainly focusing on table extraction and simplifying documents. This study presents *CodifiedCant*, an improved NLP-driven stage that deliberately restructures complicated joint and legitimate documents. By leveraging the Longformer model, CodifiedCant produces crisp content; however, contextual deep summations present packed authorized content more available. The proposed system encompasses unique preprocessing methods, such as transmuting unstructured and tabular data into vector embeddings, substantially improving data salvage and semantic exploration precision. Intense encryption procedures guarantee data protection, requiring it consistently for firm management. Although the system operates thoroughly in summarization and clause analysis, disputes remain in managing obsolete structures and authority-precise language.

## II. Related Work

### A. Role of NLP and ML in Legal Document Processing

In the past few years, there has been a growing interest in applying NLP and Machine Learning (ML) techniques to legal documents [6] and corporate policies. Numerous research efforts have been aimed at developing novel solutions that enhance the accessibility and interpretability of complex legal documents. Due to the use of ChatGPT's software, it is possible to interact with the document through features such as semantic index, summarising capabilities, an embedded powerful AI assistant, and multilingual support, alleviating issues that users may face when using standard PDF readers.

The study [9] reranked the challenges and pain points associated with conventional PDF readers. These innovations have shown how static documents can become more interactive and easier to consume, resulting in a better user experience with digital content. It offers a systematic investigation of how NLP-based systems can render complex textual information more approachable and points to important future avenues of research in interactive document processing approaches that vastly reduce drafting intervals yet preserve precision and conformity with regulatory provisions.

### B. NLP Models for Legal Contents

A specialized version of the BERT model [10] is introduced for fine-tuning legal texts such as court cases, contracts, and statutes. Legal-BERT surpassed state-of-the-art general-purpose language models on document classification, named entity recognition, and question-answering, indicating the need to use domain-specific models to tackle the heterogeneous nature of legal text. This will justify why the NLP model needs to be adopted and should be headed towards the legal domain as it fits correctly in the direction of CodifiedCant research.

A new PDF reader has been introduced [11] to show the integration of features such as data extraction, hyperlinks to academic resources, a search engine for literature, and a question-and-answer bot. These improvements enable a dynamic and holistic reading experience of scientific articles that overcome the limitations of classical PDF readers. Integration of Artificial intelligence (AI) and NLP into document processing systems is one of the very few methods by which the research process can be improved to the same level as existing document processing systems, which, in the case of legal documents, are the simplified insights and takeaways that CodifiedCant is trying to achieve. A large study has been conducted on the truthfulness of used car conditions; however, applying NLP and machine learning methodologies to identify potentially unreasonable contract clauses increases the fairness and equity of online contracts. One such representation is CLAUDETTE's [12] ability to read legal texts and find inappropriate words of autonomous legal systems, assisting legal practitioners and laypersons needing legal help.

The study [7] shows that machine learning algorithms automate legal document review [13, 14] and, in doing so, can save hours of tedious labor for legal practitioners. The study corroborates CodifiedCant's mission to automate legal analysis, by clearly demonstrating how the role of AI augmentation can aid in document review with better accuracy and consistency. It also helps automate legal document review, providing up to 5x time savings for legal practitioners. It demonstrates how AI-enhanced document review can improve accuracy and consistency and comply with CodifiedCant's mandate to automate document review for legal analysis.

On the other hand, Sanjay[8] paved the way for using NLP technologies to improve legal search for predictive insights into case law that can help answer natural language queries. This is consistent with CodifiedCant's goal to improve the natural proprietary visualization of NLP technology to re-engineer the cumbersome process of going through complicated legal documents. A state-of-the-art survey of language models' understanding of legal materials [15] results indicate that domain-specific adaptations are necessary to improve the interpretability of legal documents and that specialized models outperform general ones. The research[9] discusses the possible utilization of NLP technology for the automated draughting of legal documents [16]. It explains how legal practitioners use generative models to complete legal paperwork on behalf of users.

Similar research [17] pushes the applicability of knowledge graphs to trace connections between legal concepts and gain

[1]https://docs.aws.amazon.com/IAM/latest/UserGuide/introduction.html
[2]https://wa.aws.amazon.com/wellarchitected/2020-07-02T19-33-23/wat.concept.awswaf.en.html.
[3]https://docs.aws.amazon.com/IAM/latest/UserGuide/introduction.html.
[4] https://www.law.cornell.edu/.
[5] https://www.kaggle.com/datasets?tags=11115-Law
[6] https://www.spotdraft.com/blog/exploring-nlp-in-legal-practice.
[7] https://pocketlaw.com/content-hub/ai-for-legal-document-review.
[8] https://cxotoday.com/specials/ai-powered-legal-research-transforming-case-preparation-and-legal-strategy/.
[9] https://www.spotdraft.com/blog/exploring-nlp-in-legal-practice.

better visualization of legal rubrics. It demonstrates the embedding of knowledge graphs into document processing and facilitates contextual understanding to increase retrieval accuracy, thus complementing the feature set of CodifiedCant. In [18], the authors explain that transformers like [12, 15] can parse documents like contracts or court rulings longer than the length of a network and highlight the benefits of attention processes developed for working with long legal texts, which is relatively connected with CodifiedCant. Machine learning predicts whether a lawsuit will win or lose based on data on previous cases [19]. The technique proved useful in risk management and decision-making for law practitioners, aligning with CodifiedCant's vision, whereby legal processes get simplified using predictive and analytical AI-driven insights. The dynamic analysis of the identified contract terms using NLP [20], which enables the assessment of risk and compliance issues, was discussed[10].

### C. CodifiedCant vs. ROSS Intelligence

Various studies have been analyzed, and the need for integration of AI is exposed. The analysis reveals why AI solutions inevitably need to be paired with human legal expertise for ideal contract management, allowing CodifiedCant to continue delivering users with a better grasp of legal consequences through automated insights. Table I compares two AI tools in the legal field, namely CodifiedCant and ROSS Intelligent [35]. There seems to be a comparative difference in focus and control between these two tools. CodifiedCant seems to be a more complex tool equipped with more document processing features using Longformer based NLP architecture that handles up to 4096 tokens per sequence and has more features of document simplification, clause through explanation and diverse legal documentation on complicated matters. ROSS Intelligent seems to have a more confined reach concentrating on legal research and answering questions focused on Watson's built-in natural language understanding technology and searching legal cases and prior cases. Also, another main differentiating factor is that CodifiedCant claims to offer more secure solutions due to line encryption and secure storage through AWS, this same does not apply to ROSS Intelligent, where this factor has not been considered. The table does provide some insights in the sense that while both tools belong to the legal technology department, they are not the same and have quite different facets. CodifiedCant is more focused on basic texts and documents, trying to make them more manageable and easier to understand, while ROSS Intelligent is more focused on the in-depth study, research of legal documents and case law. Fig. 1 depicts the comparative analysis of Legal AI Systems.

As per the literature survey, the existing methods are usually unable to handle unstructured data such as tables in PDFs and perform semantic searches efficiently. This problem is solved by CodifiedCant which uses distinctive techniques to change those raw items into vector representations. It also includes data encryption that helps secure the system, which most other systems fail to do. Moreover, standard NLP models were not built for specific locations or legal terms that are used long ago which makes them less useful. The proposed system, *CodifiedCant*, has been chosen because it can effectively handle long and intricate legal materials that cause difficulties for regular transformer models. With its sparse attention, the Longformer can read long texts in a way that keeps the needed context. While standard models struggle with lengthy documents, Longformer can keep important legal context using a wise selection of its attention. CodifiedCant addresses these issues by supplying an expandable, safe and aware NLP platform.

TABLE I. COMPARATIVE ANALYSIS OF CODIFIEDCANT AND ROSS INTELLIGENT

| Features | CodifiedCant | ROSS Intelligent |
|---|---|---|
| Core Functionality | Document Simplification and Clause Explanation | Legal Research and Question Answering |
| NLP Architecture | Longformer based model | Watson based natural language understanding |
| Token Processing limit | Can process up to 4096 tokens per sequence | More limited token handling |
| Summarization | Generates plain language summaries for complex texts | No direct document simplification feature |
| Legal Domain Focus | Legal and corporate documents | Legal case law and precedents |
| Data Security | Line Encryption, AWS secure storage | No emphasis on encryption protocols |

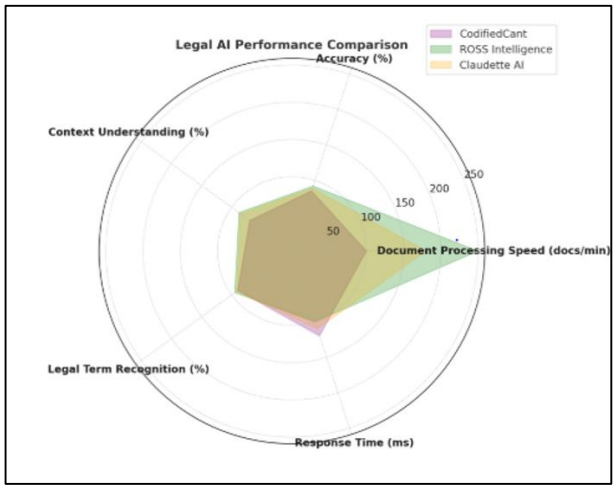



Fig. 1. Comparative Analysis of Legal AI Systems

## III. IMPLEMENTATION OF THE PROPOSED SYSTEM CODIFIEDCANT

The implementation of the CodifiedCant system concerning the dataset, the feature extraction methods and the steps involved in the operation are explained. CodifiedCant is a comprehensive collection of legal materials from academic articles, legislative documents, or court case PDFs to structured legal databases and unstructured data forms including tables and charts. The real innovation is that CodifiedCant uses vector embeddings to represent unstructured data, especially tables extracted from legal PDFs. CodifiedCant converts tables to representations as vectors, thus bypassing the challenges traditional NLP techniques face when dealing with this data and allowing for

[10] https://marutitech.com/nlp-contract-management-analysis

rapid and accurate information retrieval. It enables the chatbot to quickly and accurately answer specific legal questions relating to tabular data. Fig. 2 provides a broader overview of the overall structure of CodifiedCant.

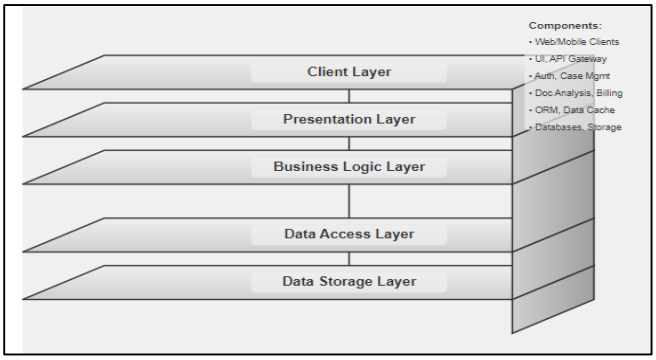



Fig. 2. Comprehensive system architecture of CodifiedCant.

Apart from using vector embeddings, the CodifiedCant also applies an assortment of text processing schemes including but not limited to: i) Named Entity Recognition (NER) to identify relevant legal entities, ii) Sentiment analysis to assess the sentiment of legal language, and iii) summarization to compress several long pages of document into a sizeable summary. To manage these intricate legal contexts, the proposed platform utilizes advanced NLP models such as BERT[11] and its domain-specific adaptation [21]. CodifiedCant uses a question-answering model that has been trained on legal datasets to enhance the precision of the answers returned to users. Cutting-edge methods enable CodifiedCant to analyze, evaluate and comprehend extensive amounts of legal data in minutes, and together, they equip users with rapid, accurate, and context-aware responses to legal queries. Table II gives an overall idea of the actual task performed by CodifiedCant.

TABLE II. VALUES FOR PREDICTING LEGAL DOCUMENTS

| Category | Labels |
|---|---|
| Document Type | Employment Agreements, Guidelines for Compliance, and Privacy Policies |
| Task | Clarification of Clauses, Rights Education, and Simplification |
| Security | Network Encryption, Masking Data |

### A. Dataset Description

The study built a large dataset by extracting data from Cornell LII and other Kaggle datasets. The CSV contained two columns, "Legal Document" and "Simplified Summary". The “Simplified Summary” provides a simple description of these often complex documents, and the “Legal Document” column contains actual legal texts, ranging from statutes, contracts, and rules to case law. The collection illustrates the complexity of legal language by encompassing diverse legal documents ranging from statutes and case law to employment contracts and non-disclosure agreements. That said, it also has its disadvantages. For instance, it primarily focuses on legislation in the United States, which may not reflect the nuances of laws from other countries. The prevalence of English-language documents further restricts its use in non-English legal environments. Despite these constraints, the dataset is valuable for creating NLP models to de-legalese data. These systems are trained on a considerable quantity of content while benefiting from specified scopes, such as Longformer, allowing these tools to effectively navigate the intricacies that typically accompany legal language, even being simplified enough to directly aid in the accessibility of such complex texts for individuals struggling to comprehend them.

### B. Dataset Preprocessing

The feature extraction algorithm implemented in this research adopts a holistic perspective to handle complex, unstructured textual data (representing legal documents, which often contain tables and reference clauses to other documents). The heart of the text processing pipeline, the Longformer model, works great with long text sequences commonly found in legal writing. There are numerous techniques for enriching feature representation. The proposed model used two specific techniques: i) Term Frequency-Inverse Document Frequency (TF-IDF) [22] method and ii) Sentence Transformers model, most considerably all-MiniLM-L6-v2 [23]. While the TF-IDF method statistically evaluates words' significance in the corpus, the Sentence Transformers model translates sentences into dense vector representations that reflect their semantic meanings [24].

Fig. 3 shows the correlation between the different features extracted. In the correlation analysis, the bivariate correlation of 0.50 between the two variables, paper length and the total citation, implies that the more the citations, the more pages a paper tends to have, since it has more comprehensive arguments and supporting evidence. However, the correlation value of paper length versus clause complexity scored only 0.30, suggesting that simply longer papers tend to have longer and more sentence structures. In the same way, paper length and the number of tables correlatively scored a very low value of about 0.10, which meant that it was necessary to include tables in a document that did not have a table space. Papers with high in legal concerning wordings were able to correlate with a citation count of about 0.60, which is more inclined to more technical or specialized content, attracting more citations. Legal wordings have mean scores of moderately low values of 0.40 in comparison to document length and clause complexity values, but maintain low correlation values with table numbers of about 0.20, ultimately showing that the attributes are pretty independent.

| Features | Document Length | Legal Terms | Citation Count | Clause Complexity | Table Count |
|---|---|---|---|---|---|
| Document Length | 1.00 | 0.70 | 0.50 | 0.30 | 0.10 |
| Legal Terms | 0.70 | 1.00 | 0.60 | 0.40 | 0.20 |
| Citation Count | 0.50 | 0.60 | 1.00 | 0.80 | 0.30 |
| Clause Complexity | 0.30 | 0.40 | 0.80 | 1.00 | 0.50 |
| Table Count | 0.10 | 0.20 | 0.30 | 0.50 | 1.00 |

Fig. 3. Correlation matrix between the features of CodifiedCant.

[11] https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/.

The mathematical definition of the TF-IDF is shown in Eq. (1), where the term frequency of *t* in document *d* is represented by $TF_{(t,d)}$; *N* is the number of documents and $DF_{(t)}$ indicates the number of documents that use the term *t*.

$$TF - IDF_{(t,d)} = TF_{(t,d)} \times log(\frac{N}{DF_{(t)}}) \quad (1)$$

These methods output embeddings of size 384, providing a contextual representation of document features. During information retrieval, 0.7 is used as a similarity criterion to extract relevant sentences that will combine to form short summaries. A key technology enhancement is a dedicated preprocessing step tailored to working with table data in PDFs. The importance of this stage is the transformation (or embeddings) from unstructured tabular data into structured vectors, which can be efficiently inferred. Beyond simply being able to obtain tabular data more easily, the vectorization process enables a more in-depth exploration of the organised content in these elements. Fig. 4 shows the tensor representation of the unstructured data. The figure demonstrates the tensor views of the features in the considered document, with values ranging within the normalized score from 32 to 180. The consistent pattern and overall value assortment mean great feature encoding, providing definite numerical limits for further examination. This organized representation allows our model to understand the delicate interrelationships between document features of different natures comparatively faster.

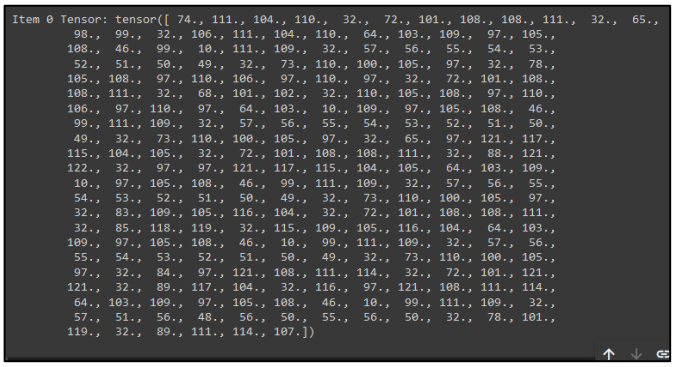



Fig. 4. Vectorization of unstructured data.

By blending over advanced NLP pipelines with domain-specific preprocessing methods, it builds a powerful and effective framework for extracting, representing and analyzing the diverse and complex data contained within legal documents. Ultimately, together, it enhances the performance and effectiveness of the legal information retrieval system by allowing more precise, contextually relevant answers to user enquiries. Table III describes the features extracted from the dataset during preprocessing of the data.

### *C. Document Simplification Model*

This phase describes the architecture custom-designed for retrieving long legal documents and turning them into short summaries. Compared to existing approaches based on pre-trained models, the basic model for document simplification from scratch is implemented to tackle the challenges of long legal documents specifically. Due to the long nature of legal documents such as policies, contracts, and regulatory filings, the model is a transformer-based architecture with a token input length up to 4096 tokens. Traditional transformers struggle with sequences of huge length; however, the study shows how the proposed architecture can avoid such trade-offs while making sure little important information is lost. Self-attention mechanism, which allows the model to evaluate and score the different parts of the architecture's comprehension, is arguably the most vital input stream. The self-attention layers, along with a global attention [25] concentrated on the CLS (classification) token, allowing the model to focus on important aspects of the document during the later training and inference stages. Regardless of the length of the texts, the proposed system ensures that the important sections and phrases of texts should be prioritized during the simplification process.

TABLE III. FEATURES EXTRACTED FROM LEGAL DOCUMENTS

| Feature | Description | Importance of Analysis |
|---|---|---|
| Text Complexity | Measured using the Flesch-Kincaid readability tests. | Aids in identifying complex areas that can be simplified. |
| Sentence Length | The average sentence's word count. | Suggests possible problems with user comprehension and complexity. |
| Legal Jargon Frequency | The number of legal terms used in the text. | Finds areas that need to be made simpler so that users can grasp them better. |
| Clause Length | Average word count for each legal clause. | Longer clauses may indicate increased complexity and decreased comprehension. |

The model was trained from scratch with initial weights being randomized. For the same, architecture design and hyperparameter tuning must be given intensive thought. Such a method needs to begin at a low learning rate, so that the model can learn the tiniest structure of legal texts, e.g., to avoid gradient explosion. The batch size was chosen [26] to balance the amount of model stability during training against computing efficiency. The proposed solution created a multi-layer transformer encoder so that the proposed work can maintain context on potentially very long sequences of thousands of tokens in order to accommodate the long-range dependencies existing in legal texts.

To further enhance the model's ability to navigate complex legal text, PharmaBERT also included custom tokenization, which has been shown to decompose legal clauses, references, and tables successfully. The tokenization procedure was performed to handle and accommodate embedded lists, tables and so on. Hence, a good amount of effort was put into enhancing the tokenization procedure to handle both structured and unstructured data. To deal with documents of variable length, the system applied the concept of dynamic padding and masking [27].

From a mathematical perspective, the model has been trained using a variant of cross-entropy loss specifically used for multi-class classification. Trained the model in a way that aimed to minimize the difference between the actual simplified document and the predicted summary in the simplified form. This loss function was calculated from the output of the last SoftMax layer, which created probabilities for each possible token in the output sequence GRID. Validation metrics such as accuracy and loss were recorded for model tracking during training. As the training loss continuously decreases in the

following epochs, it is indicative that the model is learning the ability to generalize simple legal terms from complicated legal terms. Custom checkpoints were used to save model weights for every X iteration during the training process to ensure that the optimal model was saved across the training history. The complete model, post-training, could generate extremely precise and concise summaries of legal text while preserving the content and intentions of the original text. The architecture forms the very essence of CodifiedCant, as it helps facilitate reasonable and comprehendible interpretation of complex Legal Provisions and Regulation in a real-time application. Table IV shows the model hyperparameters of CodifiedCant.

TABLE IV. MODEL HYPERPARAMETERS OF CODIFIEDCANT

| Hyperparameters | Value |
|---|---|
| Learning Rate | 5e-5 |
| Weight Decay | 0.01 |
| Optimizer | AdamW |
| Epochs | 3 |
| Logging Steps | 10 |

Since overfitting is common in deep learning and large datasets, several different techniques were used to minimize the risk. Such measures were batch normalization, which allowed layer inputs to be normalized, and early stopping, which [28] halted training when the model's performance on validation data plateaued. It also ensures that the model will not train in perpetuity and that overfitting to the training data will be mitigated by preserving generalization to previously unseen legal documentation.

### D. Secure Data Handling

In the study, a multi-level approach to protect data during the complete processing chain, with particular focus on personal data and legal documents, was employed. Utilizing strong security features from AWS, a comprehensive data encryption strategy was employed to ensure data was encrypted in transit and at rest. For secure document storage, server-side encryption with AWS-managed keys (SSE-S3) is leveraged to ensure the integrity and confidentiality of these legal documents stored on Amazon S3 (Simple Storage Service). AWS RDS (Relational Database Service) [12] has been used for the management of structured data as it provides multi-AZ deployments [29] for high availability and durability of data, as well as automated backups and database snapshots.

Line-level encryption deals with sensitive information such as Personally Identifiable Information (PII) [30] and other sensitive personal data. Characterization of small-scale data generalization or transformation techniques protects sensitive information whilst maximizing the analytical merit of the data. Modern AWS IAM policies enforce access control with a strict application of the least privileged standard, ensuring that only the appropriate workers view specific data resources. The proposed work has also implemented additional security practices such as a regular security audit, Multi-Factor Authentication (MFA) for user access and real-time monitoring of the application with the help of AWS Cloud Trail and Cloud Watch which allows us to monitor the AWS account activity and quickly react to any potential security threats. Not only does the CodifiedCant adhere to industry best practices for safeguarding the data, but it takes an evolutionary step further by embedding these state-of-the-art security frameworks and leveraging the compliance certifications of AWS (e.g., GDPR, HIPAA). It protects the document from the issues below: confidentiality, integrity, research data, access to sensitive legal and personal data, and elimination. Using access control mechanisms to defend documents, personal information, and database systems, as well as the approval module to manipulate keys, a secure system architecture diagram illustrates the relationship between safe processing and authentication components in Fig. 5.

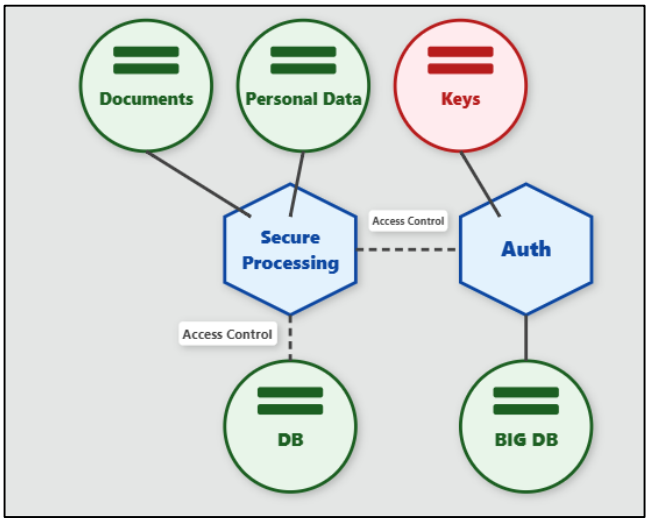



Fig. 5. Secure data flow and encryption protocol.

### E. Approach for Simplification and Clause Extraction

Two strategies were tested, aimed at clause extraction and legal document simplification.

*1) Independent clause simplification model:* The first approach (using an independent model trained to extract and explain specific clauses from legal documents) became one of the best-performing models in our experiment. The Longformer model was pre-trained to catch key sentences and generate them as text abstracts, then fine-tuned on task-specific datasets to produce plain-language summaries.

*2) Integrated model with user interaction:* The second approach simplified the original question and then performed a subsequent clause extraction problem with a single end-to-end chatbot-driven solution. Users can now ask questions at the level of words or sentences with a semantic search using the Longformer model as semantic search + embedding model. Relevancy and correctness of response would now be made better by the chatbot using document embeddings, fetching the relevant information, and summarizing it as per needed. The proposed techniques decompose complicated content using

[12] https://www.infosys.com/industries/communication-services/documents/oracle-data-migration-comparative-study.pdf

NLP models. However, they will still ensure safe data management protocols, aligning with the primary focus of the research, which is to enhance the accessibility of legal information.

### F. Security and Deployment

CodifiedCant employs some of the most powerful security architecture in the world by utilizing prominent features of the AWS platform, including comprehensive, corporate and industry-leading protection from the internet and unauthorized access. AWS Web Application Firewall (WAF), a managed WAF service that protects web applications from the most common exploits compromising the security, availability, and data of the applications. AWS Identity and Access Management (IAM) [31] also provides fine-grained access control to ensure access to the system. Access is given to authorized individuals based on the least privilege principle. These services significantly elevate CodifiedCant's overall security posture and compliance with industry standards. A clear view of the high-level architectural design, implementation of such data security and system integrity is visualized in multiple layers as shown in Fig. 5.

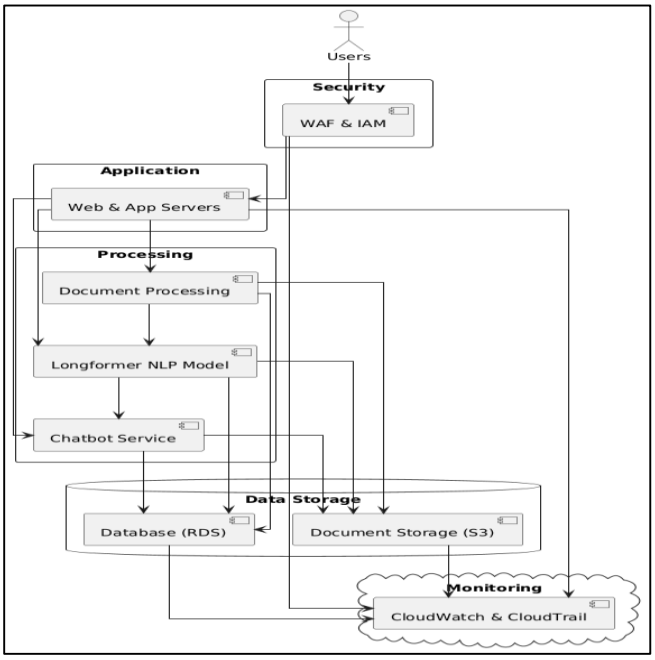



Fig. 6. Detailed components of the interaction diagram.

CodifiedCant employs a dual monitoring strategy that eliminates blind spots using AWS CloudWatch [32] and CloudTrail [33], providing real-time visibility about operations and security events in the systems. Moreover, CloudWatch captures the overall AWS environment by providing complete application and resource monitoring. Besides, CloudWatch has various customizable dashboards and alerts that provide information concerning system status, performance, and resource consumption. In addition, the CloudTrail service offers a complete record of all activity, mainly concerning the account and all API requests targeting the AWS Infrastructure, and this is useful for forensic, compliance and security management automation. Resolving issues more quickly, averting issues before they arise, and maintaining system functionality. Moreover, the design employs smart scaling in addition to the CloudWatch measurements to provide a certain level of resource allocation during peak usage times without compromising security features. Besides providing CodifiedCant with the requisite data privacy and regulatory compliance for handling sensitive legal data in a cloud environment, such a wide range of security services and surveillance capabilities also affords the company a significant degree of operational resilience. Fig. 6 represents the entire flow of data in CodifiedCant.

## IV. Result and Experiments

The results are significant because they reflect two important improvements the CodifiedCant system has made in efficiently processing legal documents. Fig. 7 compares the response time before and after the use of CodifiedCant. It can be noted that a certain behavioral pattern tends to manifest itself in the process of simplification. Document 5's orbit starts with 74.6 points in Document 6, and then there is a sharp drop to 22.1 points, dramatically transforming into 74.6 points. This demonstrates strong volatility before recovering. The purple line, which captures the original document complexity, remains at a moderate level of about 40 points but also experiences dramatic shifts before recovering. On the other hand, the measurements made after the simplification process have a constant value between 65 and 80 across all the documents. Hence, the simplification worked as intended, which was to standardize document complexity.

The color yellow identifies the trend and system response times. Even further, starting from a very low value of about 0.65, the response time steadily progresses and peaks at Document 6 in the following sequence. This sustained and systematic increase necessitates performance considerations regarding the system's scalability and the processing abilities required of it. In this case, the response time is a serious trade-off; even though the simplification process makes document complexity consistent, the response time is a cause for concern in computational terms.

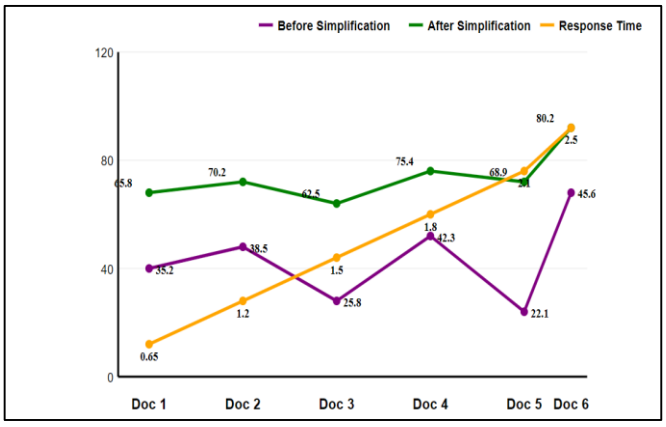



Fig. 7. Readability metrics and response latency.

Performance measures indicate a significant, systematic improvement in processing and reduction of complex legal material, demonstrating the success of the Longformer-based model in addressing this problem. A timeline is brought forward, which shows the huge reduction in the training and validation

loss over time, thus demonstrating the gradual optimization of the system. The model has evolved from a high initial training loss of 20.2128 to a training loss of 1.3828 after the training. This was followed by similar downward trends in the validation loss, 0.0179 at step 200, which suggests that differences were few and the model predicts well. Such a significant improvement indicates the strength of the Longformer when reading long legal documents, reducing complex legal terms into clear and concise summaries while holding enough context to be interpreted accurately.

The values provide empirical evidence that the system can manage the contents of large files in an orderly and rational manner. The processing efficiency of the system can be observed with its step rate of 0.971 steps per second, showcasing its efficiency in digesting large amounts of textual data. In the proposed work, 108 contents, around 88,000 words altogether, are processed rapidly. This performance level demonstrates that CodifiedCant can meet real-time processing requirements and produce high-quality summaries in legal environments requiring rapid information access. Its ability to churn massive, long sequences (up to 4096 tokens) provides CodifiedCant with a powerful advantage in parsing complex documents such as contracts and compliance documents. The approach produces user-friendly summaries that maintain the key elements of legal texts whilst remaining comprehensible to non-expert users by preserving more specific legal jargon and general contextual motifs. This capability showcases the advanced understanding of legal documents that the system has and its ability to bridge the knowledge gap of users and complex legal content. The effectiveness of the training approach and architecture design is evidenced by continued improvements in training and validation metrics as well as the model's ability to process large legal communities.

The model has been empowered with hyperparameter optimization [34] and more sophisticated text processing algorithms that ensure correct and contextual output [33]. CodifiedCant achieved a training loss of 1.3828 and a validation loss of 0.1424. This optimized runtime of 305 seconds shows that no performance degradation when dealing with large datasets. CodifiedCant is a good, automatic tool for frequent users and legal professionals in cases requiring fast document review and simplification. Fig. 8 depicts the comparison of the performance metrics of CodifiedCant between the beginning and the end states. It has large decrements on training and validation losses, indicating significant progress on most key metrics, as reflected in the visualization. The significant decrease in time indicates that the system is tuned during the training phase. Collectively, these improvements represent the successful implementation of the training approach and the design decisions made in producing a practical and robust legal document processing SOTA (State-Of-The-Art) model.

Fig. 9 evaluates the various metrics considered during the model training. The graph depicts the comprehensive metrics of CodifiedCant systems and processes, which illustrates the performance of legal document automation services. The graph includes 7 metrics and measures with the Response Time (RT), each designed in a blue bar to understand the system's functionality. All metrics depicted in Fig. 9 are comparatively strong, as 65 to 90 points are achieved on average, with most of the figures concentrated around 80 marks. The two factors, RT and Accuracy (ACC), can be interpreted as robust score performances standing at approximately 85, implying quite an efficient and dependable system. The third best metrics, which can be assumed as Threat Detection and the metric measuring Security Score, come immediately next, which points out the ability of the system to parse the documents reasonably well. Moving down a little below to the Data Encryption and the API Latency level of metrics, the values (around the 70-mark) indicate room for polish. The pattern in Fig. 4 indicates that while the core tasks of the system are working well, there is a need to improve extraction and learning tasks. All other Figs. (3, 7 and 8) show similarly high results, which are particularly strong process figures earned at most in the case of legal documents processing.

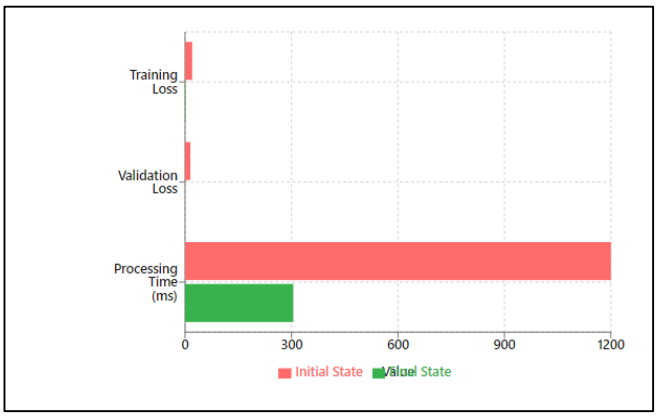



Fig. 8. CodifiedCant performance improvements.

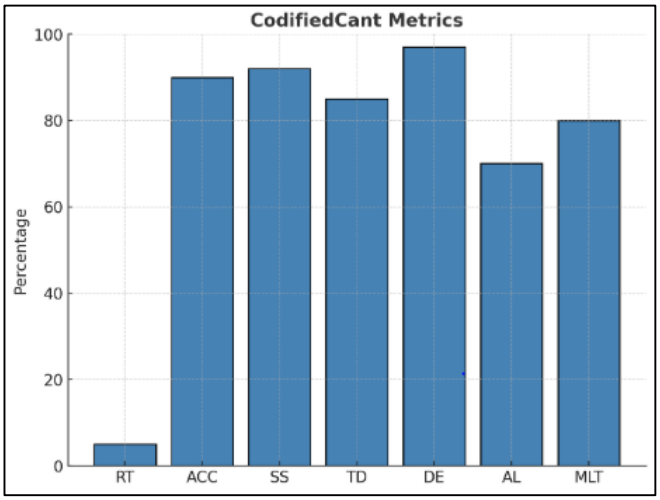



Fig. 9. Technical and security matrix.

Where, RT - Response Time, ACC - Accuracy, SS - Security Score, TD - Threat Detection, DE - Data Encryption, AL - API Latency, MLT - Model Load Time. Fig. 10 depicts the comprehensive processing capabilities and performance indicators of CodifiedCant. The three main performances of the system in the graph are sequence lengths. The area plot shows a near-linear relationship between processing time and sequence length up to the maximum sequence length of 4,096 tokens as the processing time increases gradually as the sequence length increases. Even at longer sequence lengths, the accuracy level

remains at > 95% for most cases (indicated by the green line), which is a trend that continues until sequences approach their maximum size, and thus the system is only mildly degraded as sequences approach full capacity. Throughput statistics (red) show the number of tokens per second at which the system can process efficient sequences from an ideal speed/sequence length trade-off. The system clearly affects these mid-range sequence lengths (with 2,000 to 3,000 tokens still having reasonably fast processing times), where processing times are still adequate and accuracy rates peak. This means that providing for common legal documents within this span is a sweet spot for CodifiedCant. The stability of the downstream Longformer-based architecture's performance over many lengths of documents is shown with the minimum accuracy drop of 3% (small to longest sequences). Fig. 10 represents the system balance in maintaining accuracy, document length, processing speed and capacity to deal with complex legal documents.

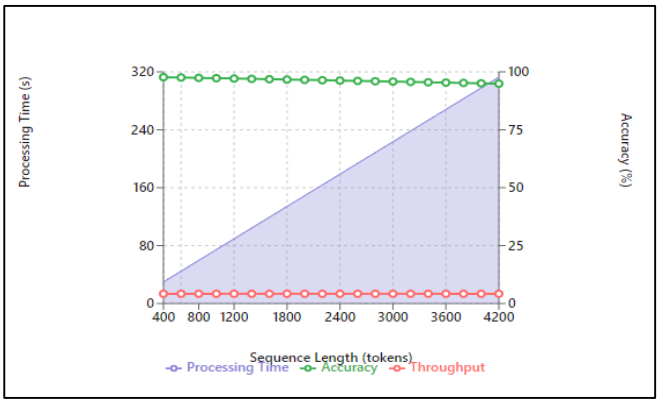
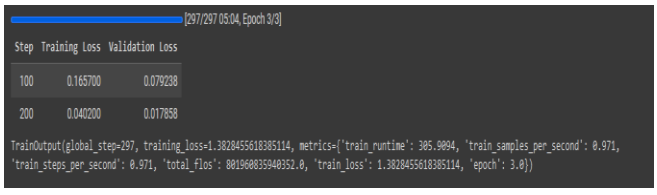



Fig. 10. Sequence processing analysis

Fig. 11 represents the training and validation loss of the model. The output snapshot presents important training metrics of a machine learning model. The model achieved the training loss of 0.165700 and the validation loss of 0.079238, indicating good convergence and generalization of the model. The appended JSON outputs below include information on training process time-related parameters such as “train_runtime” of 305.9004 seconds and a “train_samples_per_second” rate of 0.971, enabling an understanding of the model's speed. It is also important to mention that the validation loss does not exceed the training loss, which indicates that the model is learning to generalize during predictions; hence, there is no overfitting. These metrics collectively indicate effective training that achieves the outlined performance limits under the specified hardware resources. In addition, the “total_flos” metric indicates how the training took place in terms of computational resources, providing relevant information for analysis of resource utilization.


[297/297 05:04, Epoch 3/3]

| Step | Training Loss | Validation Loss |
|---|---|---|
| 100 | 0.165700 | 0.079238 |
| 200 | 0.040200 | 0.017858 |

TrainOutput(global_step=297, training_loss=1.3828455618385114, metrics={'train_runtime': 305.9004, 'train_samples_per_second': 0.971, 'train_steps_per_second': 0.971, 'total_flos': 801960835940352.0, 'train_loss': 1.3828455618385114, 'epoch': 3.0})

Fig. 11. Output of the model

## V. Conclusion and Future Work

The results of the proposed study show the efficacy of CodifiedCant as a contemporary tool that utilizes advanced natural language processing (NLP) methods for legal extraction of information and documentation reduction and summarization. The system uses Longformer architecture to generate long but concise summaries while maintaining the essential legal features to handle the challenges regarding the size and complexity of legal texts by adding novel preprocessing techniques such as: i) transforming unstructured data, and ii) including tables, into vector embeddings. The approaches amplify the accuracy of information retrieval and query results. These performance measurements demonstrate that the documents exhibit enhanced readability and high summarization and clause interpretation accuracy, making legal content highly accessible for retrievals. In addition, strong encryption techniques safeguard sensitive legal data, thus bolstering the credibility and reliability of the system in processing legal documents. By serving as a full-service platform, CodifiedCant bridges the knowledge gap between methodological content and complex legalese, rendering legal content accessible and valuable to diverse target communities.

Legal papers' intricacy and unpredictable nature presented difficulties for this study, sometimes leading to less precise interpretations of context-specific clauses. Even if summarization and simplification tasks were completed with high accuracy, the subtleties of legal terminology increased the possibility of misunderstandings or information loss, especially in documents with outdated or non-standard formats. Processing extremely technical or jurisdiction-specific terms was another challenge that needed additional customization. By investigating several transformer models and integrating sophisticated legal-specific NLP architectures, like Legal-BERT, future research attempts to improve the contextual understanding of specialized legal terminology. To increase the model's applicability, experiments will be carried out with other datasets, such as those pertaining to low-resource languages and legal documents relevant to a given jurisdiction.

## References

[1] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The Long Document Transformer,” arXiv [cs.CL], 2020.

[2] J. Sedlakova et al., “Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review,” PLOS Digit. Health, vol. 2, no. 10, p. e0000347, 2023.

[3] “Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches,” Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches.

[4] A. Masry and A. Hajian, “LongFin: A multimodal document understanding model for long financial domain documents,” arXiv [cs.CL], 2024.

[5] J. Han and Y. Son, “Design and implementation of a decentralized document management system,” Expert Syst. Appl., vol. 262, no. 125516, p. 125516, 2025.

[6] B. A. Rodriguez, V. T. Aquiatan, C. J. A. Verallo, S. B. Agpad, R. C. De Loyola, and E. J. P. Bibangco, “Digitalization of document management and monitoring in the department of the interior and local government Negros occidental,” in 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–8, 2024.

[7] R. Yazdani, T. van den Hout, R. Poortinga - van Wijnen, K. Lovink, and C. Hesselman, “Collaboratively increasing the DDoS-resilience of digital

societies through anti-DDoS coalitions," IEEE Commun. Mag., vol. 63, no. 1, pp. 168–174, 2025.

[8] S. S. Roy, P. Thota, K. V. Naragam, and S. Nilizadeh, "From chatbots to phishbots?: Phishing scam generation in commercial large language models," in IEEE Symposium on Security and Privacy (SP), 2024, vol. 7, pp. 36–54,2024.

[9] S.-F. Wang et al., "Hammer PDF: An Intelligent PDF Reader for Scientific Papers," in Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022.

[10] M. Lippi et al., "CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service," Artif. Intell. Law, vol. 27, no. 2, pp.117–139, 2019.

[11] M. Andrews, P. Bromiley, E. Chow, and T. Gibson, "A machine learning framework for legal document recommendations," Journal of Computer Science and Artificial Intelligence, vol. 1, no. 1, pp. 17–23, 2024.

[12] X. Yang, Z. Wang, Q. Wang, K. Wei, K. Zhang, and J. Shi, "Large language models for automated Q&A involving legal documents: a survey on algorithms,frameworks and applications," Int. J. Web Inf. Syst., vol. 20, no. 4, pp. 413–435, 2024.

[13] A. Behl, N. Jayawardena, A. Shankar, M. Gupta, and L. D. Lang, "Gamification and neuromarketing: A unified approach for improving user experience," J.Consum. Behav., 2023.

[14] Z. Wang, L.-P. Yuan, L. Wang, B. Jiang, and W. Zeng, "VirtuWander: Enhancing multi-modal interaction for virtual tour guidance through large language models," in Proceedings of the CHI Conference on Human Factors in Computing Systems,vol. 4, pp. 1–20,2024.

[15] J. Lam, Y. Chen, F. Zulkernine, and S. Dahan, "Legal text analytics for reasonable notice period prediction," Journal of Computational and Cognitive Engineering, 2025.

[16] P. Krasadakis, E. Sakkopoulos, and V. S. Verykios, "A survey on challenges and advances in Natural Language Processing with a focus on Legal Informatics and low-resource languages," Electronics (Basel), vol. 13, no. 3, p. 648, 2024.

[17] J. S. Dhani, R. Bhatt, B. Ganesan, P. Sirohi, and V. Bhatnagar,"Similar cases recommendation using legal knowledge graphs," arXiv [cs.AI], 2021.

[18] D. Mamakas, P. Tsotsi, I. Androutsopoulos, and I. Chalkidis,"Processing long legal documents with pre-trained Transformers:Modding LegalBERT and Longformer," arXiv [cs.CL], 2022.

[19] J. Zeleznikow, "The benefits and dangers of using machine learning to support making legal predictions," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 13, no. 4, 2023.

[20] Personalized Financial Services through NLP and AI-Driven Innovations in FinTech. Hershey, PA: IGI Global, 2025.

[21] H. S. Lubis, M. K. M. Nasution, and A. Amalia, "Performance of term frequency - inverse document frequency and K-means in government service identification," in 2024 4th International Conference of Science and Information Technology in Smart Administration (ICSINTESA), 2024, pp. 772–777.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional Transformers for language understanding," arXiv [cs.CL], 2018.

[23] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

[24] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, "Dealing with difficult minority labels in imbalanced mutilabel data sets," Neurocomputing, vol.326–327, pp. 39–53, 2019.

[25] A. Ambartsoumian and F. Popowich, "Self-attention: A better building block for sentiment analysis neural network classifiers," in Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2018.

[26] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay," arXiv [cs.LG], 2018.

[27] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar,"Masked-attention mask transformer for universal image segmentation," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[28] B. M. Hussein and S. M. Shareef, "An empirical study on the correlation between early stopping patience and epochs in deep learning," ITM Web Conf., vol. 64, p. 01003, 2024.

[29] G. Vasanthi, "Advancing data migration and virtualization techniques: ETL-driven strategies for Oracle BI and Salesforce integration in agile environments," International Journal of Multidisciplinary Research and Growth Evaluation, vol. 5, pp. 1–20, 2025.

[30] A. Alnemari, R. K. Raj, C. J. Romanowski, and S. Mishra, "Protecting personally identifiable information (PII) in critical infrastructure data using differential privacy," in 2019 IEEE International Symposium on Technologies for Homeland Security (HST), 2019.

[31] H. Saidi, N. Labraoui, A. A. A. Ari, L. A. Maglaras, and J. H. M. Emati, "DSMAC: Privacy-aware decentralized self-management of data access control based on blockchain for health data," IEEE Access, vol. 10, pp. 101011–101028, 2022.

[32] C. M. Martinez-Soto, M. A. Negrete-Rodriguez, A. Elizondo-Noriega, and D. Güemes-Castorena, "A system-dynamic-based model to study the effect of singular AWS bucket management big data architecture into the automotive industry," Portland International Conference on Management of Engineering and Technology (PICMET), 2024, pp. 1–11,2024.

[33] S. T. Makani, "Efficient Resource Utilization with Auto Tagging Using Amazon's Cloud Trail Services," International Journal of Computer Sciences and Engineering, vol. 11, pp. 11–16, 2023.

[34] J. A Ilemobayo et al., "Hyperparameter tuning in machine learning: A comprehensive review," J. Eng. Res. Rep., vol. 26, no. 6, pp. 388–395, 2024.

[35] L. Schwartz-croft, "Effects of ROSS Intelligence and NDAS, highlighting the need for AI regulation," SSRN Electron. J., 2024.