Enhancing Risk Assessment in Legal Documents Through Advanced Machine Learning

MSc Research Project
MSc Data Analytics

Sai Teja Pusarla
X23193638

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sai Teja Pusarla		
Student ID:	X23193638		
Programme:	Data Analytics	**Year:**	2024
Module:	MSc Research Project		
Supervisor:	Jorge Basilio		
Submission Due Date:	12/12/2024		
Project Title:	Enhancing Risk Assessment in Legal Documents Through Advanced Machine Learning		

Word Count:8779 **Page Count:** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Sai Teja Pusarla ……………………………………………………………………………………………

Date: 12-12-2024 ……………………………………………………………………………………………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Risk Assessment in Legal Documents Through Advanced Machine Learning

Sai Teja Pusarla
X23193638

Abstract

In this thesis, we will conduct research to examine the use of BERT, Legal-BERT and Isolation Forest models to strengthen the risk assessment and anomaly detection of these legal documents. It shows that Legal-BERT, a BERT model pretrained on legal texts can outperform a general BERT model well beyond chance level with a test accuracy of 76.47%, and excellent specificity in precision, recall and F1 score measures. In addition, the Isolation Forest Algorithm, an unsupervised learning model, was able to find 975 anomalies among 19,501 clauses, and we can say that it is able to find deviants from normal threads of law. Most notably, real practice was the proving ground for these models, as they were tested in real-time on new legal texts and their applicability was confirmed. These findings support the opportunity advanced machine learning models provide for automation, increased accuracy, and scale in the analysis of legal documents and constitute a meaningful step toward automated risk assessments applied to real-world legal scenarios within the field of legal informatics.

Keywords: BERT, Legal BERT, Clause, Risky, Not Risky, Isolation Forest, Anomaly

1. Introduction

1.1 Background

Business contracts are vital in organizations due to their role in defining business relations, standard procedures, and compliance with the law. In this case, it becomes a challenging task for organizations to analyze such contracts to prevent the risk of any legal or financial issues going unnoticed, which is always risky for any organizational firm. In the past, it was done through a process involving legal practitioners which is costly and very time-consuming. Due to recent innovations in Natural Language Processing (NLP) and machine learning (ML), this process has the potential to be automated, which will make the process faster, more accurate, and capable of handling significantly larger volumes of data. The BERT and legal BERT model is used to analyze huge volumes of texts and categorize documents in terms of risks. This work applies these technologies to mitigate the difficulties associated with risk evaluation and recognize unusual patterns in legal contracts to contribute a new resolution to the legal and corporate management problem.

1.2Aims and Objectives

1.2.1 Aims

The main objective of this study is to provide a reliable approach to the assessment of risks and anomalies in legal texts by employing modern machine learning tools.

1.2.2 Objectives

- To recognize and assess high-risk clauses within legal contracts with NLP-based models.
- To analyze legal documents and identify the presence of anomaly results through the concept of unsupervised learning.
- To evaluate how much working with general and Legal-BERT models is beneficial compared to each other.
- To offer a conceptual and realistic application that would classify potential risks and other abnormality indicators found within legal documents, in real time.

1.3 Research Questions

This study aims to answer the following questions:

1. How can BERT and Legal-BERT utilized to assess risks in legal documents, and what challenges can arise from applying these models to real-time legal text analysis?

2. How effective do unsupervised learning approaches in detecting anomalies within legal texts, and what role do these detections play in amplifying the precision and comprehensiveness of legal document analysis?

1.4 Problem Statement

The objective examination of the legal contracts’ text experiences several problems, namely, the presence of a high cost, low efficiency, and the impossibility of increasing the effectiveness of manual examination several times due to the scale of the activity and the propensity for errors. Therefore, during the augmentation in legal document complexity and number, the challenges of providing a thorough and precise review keep rising. Today’s methods sometimes struggle to correctly analyze legal texts and recognize the specific risks or unusual patterns. This lack is behind the need for an automated, friendly, and accurate method adapted for legal document analysis. Legal contract analysis presents several main challenges which include, scalability is a major problem where varying numbers of contracts are executed in large organizations necessitating the use of manual methods which are time-consuming and therefore slow down the workflow. The legal language is highly technical and specialized and, for the most part, is shielded with several context-influenced words and phrases requiring legal professionals' input to interpret. Data confidentiality brings extra concern, as organizations are wary of uploading sensitive papers to external sites, thus requiring locked-down, in-house solutions.

1.5 Summary

This paper focuses on the interest and implications of machine learning and NLP in reducing the problems associated with contract legal analysis. As a result, the study is intended to be pragmatic, feasible, and accurate in risk assessment and anomaly detection compared to prior machine learning legal research. The difficulties and opportunities of the present study, this work aims at progressing the analysis of legal documents through automation regarding safety, speed, and credibility.

1.6 Research structure

Chapter 2: Related Work
This survey paper discusses the literature on the application of NLP and ML in legal contract analysis and discusses rules based and trend-based approach for risk assessment and identifies the essence and issues with respect to detection of anomalies in the area of interest.

Chapter 3: Research Methodology
It explains the data collection, data preprocessing, and model used to analyse the legal text, whether in risk assessment or anomalies detection context.

Chapter 4: Design Specification
Specifies the system architecture and components for risk assessment and anomaly detection, and describes how specific models and technologies will be implemented

Chapter 5: Implementation
Describes key technical aspects of BERT, Legal BERT, and Isolation Forest models including data preparation and training, with discussion of challenges met.

Chapter 6: Evaluation
Provides both quantitative and qualitative evaluations of models, including model performance on risk assessment and anomaly detections and whether they are applicable in real world scenarios.

Chapter 7: Conclusion and Future Work
Provides a summary of the results of the study, explores potential impacts on legal text mining, and recommends potential next steps to improve model performance.

2. Related work

The technique of analyzing legal contracts using NLP and ML is still considered a growth area of research. This section provides a survey of previous work in the domain, assesses the advantages and disadvantages of prior methods, and positions this work within that context. The discussion is structured into two subsections two of the issues that remain essential in legal documents and texts including Risk Assessment in legal documents and Anomaly Detection in legal texts.

2.1 Risk Assessment in Legal Documents

The primary meaning of risk assessment in legal documents is an evaluation of the legal text to identify clauses or certain language patterns suggesting risks or compliance hazards. Many works describe the progress and the deficit in this field. One important direction investigated how one can employ transformer-based models in document classification in law. This paper verified how it efficiently detects legal topics and how helpful pre-trained transformers can be. However, it is limited by its ability to not only lack interpretability to comprehend contract risk. The second experiment where transformer models fine-tuned on legal corpora in one of the domains outperformed general-purpose models in tasks that specifically require knowledge of law can be listed among them.

Some works were aimed at identifying the clauses that can be potentially dangerous when used in legal contracts, and then the authors used both rule-based and knowledge-based approaches to achieve this goal. (Sharma et al. 2023), despite the proposed systems giving high precision in identifying the predefined risk types, flexibility for non-standard kinds of clauses remained a challenge, affirming the research question’s need for enhanced context comprehension. Other studies used the integration of NLP techniques with statistical models for the scoring of financial contract risk.

2.2 Anomaly Detection in Texts

Anomaly detection is significant for defining the points in legal documents significantly different from the standard regulatory provisions and indicating errors, fraud, or violation of legislation. Other papers have focused on the use of unsupervised and supervised, learning techniques for anomaly detection (YAO et al., 2024). One common example of an unsupervised model identified the anomalies from the patterns. It has been found effective, especially with high-dimensional data. However, for text data, especially legal documents, its applicability involves a lot of preprocessing or feature extraction.

Another framework proposing autoencoders for anomaly detection in legal texts outperforms and opts to identify complex anomalies by reconstructing the input data and comparing them with the deviation. However, its ability to only learn complex features from large, high-quality training datasets seriously constrained its deployment.

2.3 Comparative Analysis of Related Work

The analysis of the state-of-the-art reveals that machine learning and NLP have found a host of applications across industries. In the line with the above perspectives, the anomaly detection in the oil and gas contracts has been explored by (Cardona et al., 2024) study with bibliometric analysis, PRISMA, and PCA in identifying the leading countries and institution that works on the machine learning of this aspect. In another study, (Krishnan et al., 2022) used neural networks and other hybrid approaches of learning in an effort to improve the efficiency organizations have in identifying financial fraud such as insider trading. Using NLP for cybersecurity, (Sharma et al. 2023) used Named Entity Recognition, Topic Modeling, and Document Classification methods in order to detect cybersecurity swellings like the zero-day attacks without triggering other false signals. In the domain of fraud detection, (Boulieris et al., 2024) have proposed the FraudNLP dataset and given insight into how characteristics including applicability of NLP-based features to fraud detection in online banking while preserving privacy enhanced the classification result. Finally, (Bancale, 2022) proposed a Medical Device Cyber-Risk Assessment Model to use Named Entity Recognition, Deep Learning and Statistical methods to construct an adaptive real time and preventive cyber-risk index to improve Health care risk evaluation. All of these studies together establish the increasing significance of applying machine learning and NLP to address varied difficulties in various fields and highlight the prospects for additional developments and utilization.

2.4 Legal Clause Identification

Clause identification is one of the core levels within the field of legal NLP because it translates documents such as contracts into sections that can be analyzed. Conventional approaches to this problem have been rule-based systems primarily because the domain under consideration here has a rather formal grammar. There are random approaches that use sets of rules to characterize and find the legal clauses of the documents.

Due to deep learning algorithms in conjunction with transformer architectures, such as BERT or legal BERT in the same context, the precise identification of clauses has altered (Sharma et al., 2023). These models operate within high-rank spaces and their functionality is based on the context and semantics held by textual information. As one can derive from BERT training the general and legal data, the model is successful in identifying clauses and the connection between the legal terms.

2.5 Legal Document Summarization

The process of finding extracts of legal texts, including contracts and agreements, is gradually becoming more critical to minimize the human factor and time spent on reading large volumes of documents. Extractive methods, which are used most often, allow one to choose the most important parts of the document and present it separating them into sections according to the structure of the original text. Traditional approaches are text summarization is quite rigid and does not provide much flexibility. Dual against extractive, strengthening state-of-art transform models like GPT, BERT, and abstractive summarization can create a summarization output in natural language which is a condensed re-writing of the main content of a document (Kasman, 2024). Integrating information of these models is also able to generate results that are more human-comprehensible than models that simply extract key information from comprehensive texts.

2.6 Technological Integration and Legal Implications

BERT and Legal-BERT: Use & Impact of BERT & other kinds In the Field of Law The incorporation of BERT and Legal-BERT has provided high capabilities with accuracy in analyzing complicated legal texts which also creates substantive legal and ethical issues. Such models trained in legal language enhance the complex comprehension of legal jargon but also present a grey area of data privacy and security and bias. Those can then reflect historical biases in the training data that may distort the fairness of the decisions made when applying the law in the AI. In addition, there are a few issues regarding adherence to data protection legislation such as what level of compliance is required under the GDPR, and the definition of accountability concerning decisions made by humans in the course of their work being assisted by AI. This will require new legal frameworks to deal with things like intellectual property rights and AI-generated content, with these AI technologies being deployed in legal settings. According to (Jones & Taylor, 2023) in their study on the legal and ethical challenges introduced by enhanced AI in law, developing guidelines for responsible and ethical AI use is essential to maximising the benefits of these technologies whilst minimizing the associated risks.

2.7 Findings and Justification for Research

From the literature review, it is found that there are significant developments in the application of machine learning and NLP to legal document process analysis, however, there remain key deficiencies particularly related to the specificity and adaptability of risk assessment models such as BERT and Legal-BERT, which are more general for various text mining applications, and the anomaly detection method. Though transformer-based models have achieved progress in comprehending intricate sentences, their implementation on legal documents frequently lacks the specificity needed for proper risk assessments, not sufficiently adapting to the specialized legal vocabularies and systematically nuanced nature of legal language. At the same time, existing anomaly detection methods, including unsupervised models such as Isolation Forest, have not been sufficiently optimized to accommodate the nuances of legal texts, especially under-represented legal-tailored features governing their characteristics and legal saliency of detected anomalous observations. In addition, the usage of NLP technologies in law brings serious ethical and compliance issues, particularly in terms of data privacy and the amplification of biases that current models do not mitigate. We will bridge these gaps by tailoring Legal-BERT to a cost-effective legal text interpretation and a contextualized isolation forest-based solutions for legal documents ensuring that developers not only provide a cutting-edge but are also compliant with groundbreaking ethical paradigms and legal frameworks like the GDPR. Such dual methods offer a new avenue of research, which we believe will make a substantial contribution to legal informatics; they can greatly enhance both accuracy of legal risk assessment and efficiency of anomaly detection.

2.8 Method Validation

The comprehensive approach to method validation for BERT and Legal-BERT in legal contexts will require accounting for accuracy, robustness, and fairness. When evaluating how accurate a classification job is we talk about practical measurements such as precision, recall, and F1-scores that are important when we will be working with class imbalances as we will certainly see when we do a legal dataset. (Jackson and Tredoux, 2023) have already pointed out that cross-validation methods can play an important role in assessing generalizability to other legal documents to avoid model overfitting. It is critical to test for bias, and to conduct this testing, we propose techniques such as Adversarial Testing (Zhou et al., 2024). We conducted robustness checks in various aspects: the model outputs consistently against the formats and words of the document variates which is important as it formulates the fundamentals of abstraction of a diverse set of laws. It is essential to comply with legal standards (especially when dealing with sensitive data, as in our case, choose algorithms and programs that comply with GDPR regulations) and conduct ethical audits, as proposed by (Smith and Kumar, 2020), to ensure that the use of a given model corresponds to the values of society. Regular monitoring of models with feedback mechanisms guarantees that the models remain accurate and in line with ethical standards as legal interpretations evolve over time and better data becomes available.

3. Research Methodology

The research approach for this study on risk assessment and anomaly detection in legal documents is an attempt to enhance the application of machine learning as well as natural language processing considering the emerging challenges of processing text data from legal documents. Adopting this systematic analysis with guidelines on data collection, preprocessing, model selection, evaluation, and result interpretation produces an efficient methodology for legal document analysis with quantitative and qualitative results. The methodology is divided into several stages: The steps involved are data acquisition and data pre-processing, Model selection, training, evaluation, and lastly result interpretation.

3.1 Research Approach

This research adopts an advanced anomaly detection framework to address the complexities of analysing legal clauses. Legal clauses, essential elements of contracts, often contain nuanced and context-sensitive language where deviations from standard patterns may indicate risks, compliance issues, or negotiation opportunities. By focusing on legal clauses as the unit of analysis, this study directly addresses practical challenges in contract review, compliance assessments, and risk minimization.

The research employs state-of-the-art natural language processing (NLP) techniques, specifically leveraging BERT and Legal BERT models. While BERT provides a strong foundation for understanding textual data, Legal BERT, a domain-specific model fine-tuned for legal text, offers superior semantic understanding of legal terminology and conventions. These models enable accurate anomaly detection by identifying non-standard clauses, such as overly restrictive indemnity terms or the absence of key provisions like confidentiality clauses, which could pose significant risks.

3.2 Data Collection and Sources

3.2.1 CUAD Dataset

The Contracts Understanding Atticus Dataset (CUAD)[1] aims uniquely at containing labeled data for legal contract classification. This dataset is employed for risk assessment as a part of developing the risk assessment model, the purpose of which is to rank the contract clause as to the extent of the risk implied. The second representation of the data can be used to feed models for risk classification since the dataset contains structured data necessary for classifying a clause as a high risk or a low risk to a legal entity.

3.2.2 Legal Clauses Dataset

This dataset[2] contains a vast library of typical legal clauses used in contracts for industries of legal clause extraction, anomaly, as well as contract benchmarking. The following dataset enables a structured approach to the study of clauses and contracts to find possible discrepancies, mistakes, or variations in contract language among the contacts. It is about the rationale of utilizing the introduced structure as a dataset for training models that can gradually reduce the human engagement in the legal documents review and verification. It is possible to employ this resource to refine detecting of anomaly models, in particular for search of clauses that contain error, inconsistency or any other deviation from normative-legal language.

3.3 Data Preprocessing

The raw legal text data of both datasets need a certain amount of data processing to make them compatible with machine learning algorithms.

3.3.1 Text Normalization

The standardization process at this step involves converting all text to lowercase, removal of all punctuation and special characters, and those words which provide little semantic meaning e.g. (‘and’, ‘the’, etc.). Legal language is filled with structures that include a great number of

[1] https://www.atticusprojectai.org/cuad

[2] www.kaggle.com/datasets/mohammedalrashidan/contracts-clauses-datasets

words, and it is necessary to delete all the words that do not make any reference to the legal field for the models to take into consideration the legal terms only.

3.3.2 Tokenization

Tokenization divides the text into smaller parts or smallest units usually ranging from a word and the smallest part of a word, called SUBWORD. This process is very important, especially for the new generation of models such as Transformers BERT, and legal BERT where text complexity is best handled well. It also allows the models to cope with multiple structures of sentences and word forms characteristic of writing in the legal field.

3.3.3 Named Entity Recognition (NER)

The legal documents involve the use of specific entities such as date, parties, jurisdiction, and regulation. These are particular words and phrases, which NER facilitates to recognize and classify as appropriate for recognizing the context of legal clauses. For example, the model should be able to identify the relation 'parties' and 'law!' and 'jurisdiction' to comprehend under which conditions a certain clause is to be used.

3.3.4 Feature Engineering

Besides general transformations, more functions were extracted from such legal documents as samples. In terms of risk assessment, the following variables were applied, namely governing laws, number of contracts, and clause frequencies. For anomaly-detecting features related to the number of clauses, the frequency of words outside of the corpus, and the standard deviation of the number of sentences were used.

3.3.5 Handling Missing Data

The preprocessing involved checking for missing values in the datasets; for the case where these either have been dropped the records or used statistical methods to impute missing values such as mean or median.

3.4 Data Visualization

The bar chart **(Figure 1)** called "Frequency of Different Answers for Governing Law" shows the number of answers by jurisdiction. Again, New York was the most often cited, followed by California States and Delaware States, while other States that were mentioned include Texas, Florida States, and international jurisdictions such as the People's Republic of China-Singapore, England, and Ontario in Canada. Many entries are rare, like Switzerland, Japan, or India while some include certain California, Hong Kong, and the type of clause like California, the state where the breach occurs. This is due to different preferences in the legal areas.

Figure 1: Top 20 Governing Laws in Contracts

Figure 2: Confusion matrix

This is heatmap **(Figure 2)** is the count of agreements according to governing law and year for different jurisdictions. Along the y-axis, there are listed jurisdictions, such as California, Delaware, Florida, and the like; along the x-axis, there are two-fold of the years starting from 1993 to 2023. The frequency varies within the range from blue, as a sign of low activity, to red as the signal of high activity, based on the number of agreements made, while the highest values were identified for New York in recent years. The title at the top gives background information of the information presented in the visualization.

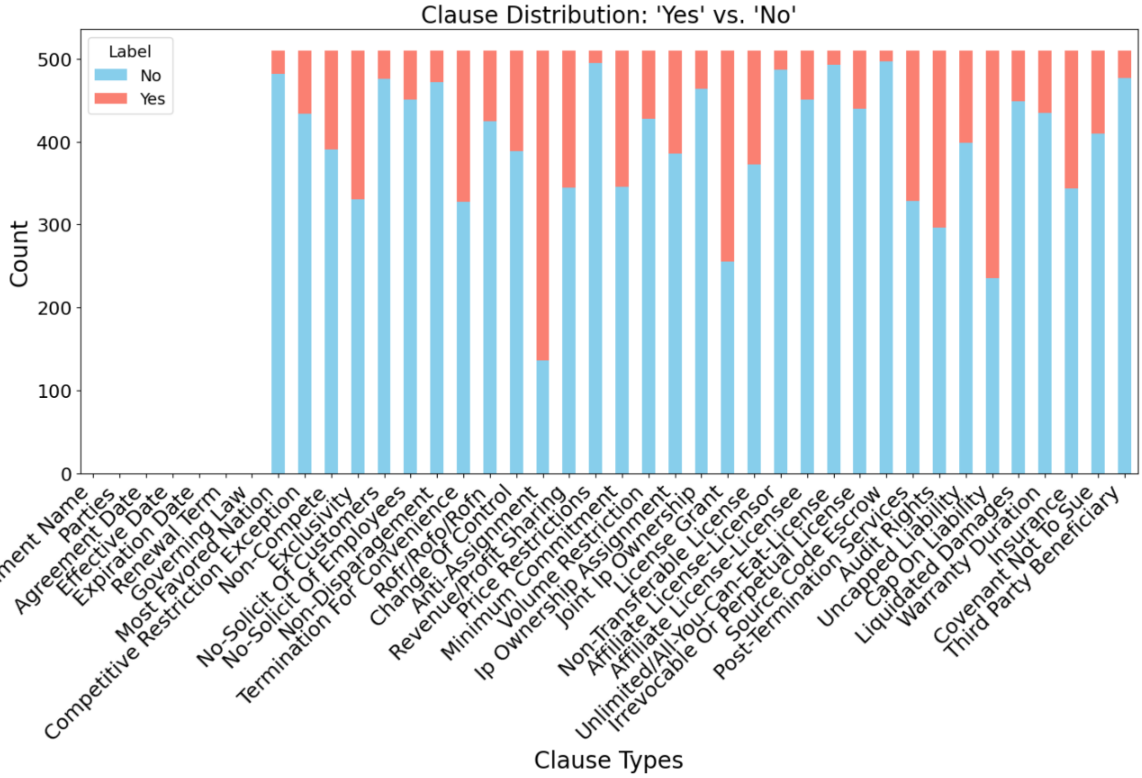


Figure 3: Clause Distribution

The following bar chart **(Figure 3)** shows the proportions of different kinds of legal clauses (attributed as Yes (risk) or No (no risk) in the CUAD dataset, which have been used in training BERT and Legal-BERT for a risk assessment project. The length of each bar corresponds to a certain clause type in legal documents (including Change of Control, Termination for Convenience, Indemnification, etc.). The blue bars show how many clauses these do not consider to be risky, and the redden the number of clauses to be deemed risky. The graph visually summarizes the data used to train and tune machine learning models to automatically detect risk in contracts and has to be able to show both the prevalence, as well as the variation across different types of clauses, of risk assessments. This is important when you have automated analysis, to dry run and add more attention on clauses that are usually riskier.

This bar chart**(Figure 4)** depicts the average word count per clause type within a dataset of legal clauses, illustrating the complexity of legal language. Clauses such as “Assignment,” “Confidentiality,” and “Vesting” have a higher word count, indicating more elaborate terms. By contrast, the clauses such as Miscellaneous and Notices are generally concise.

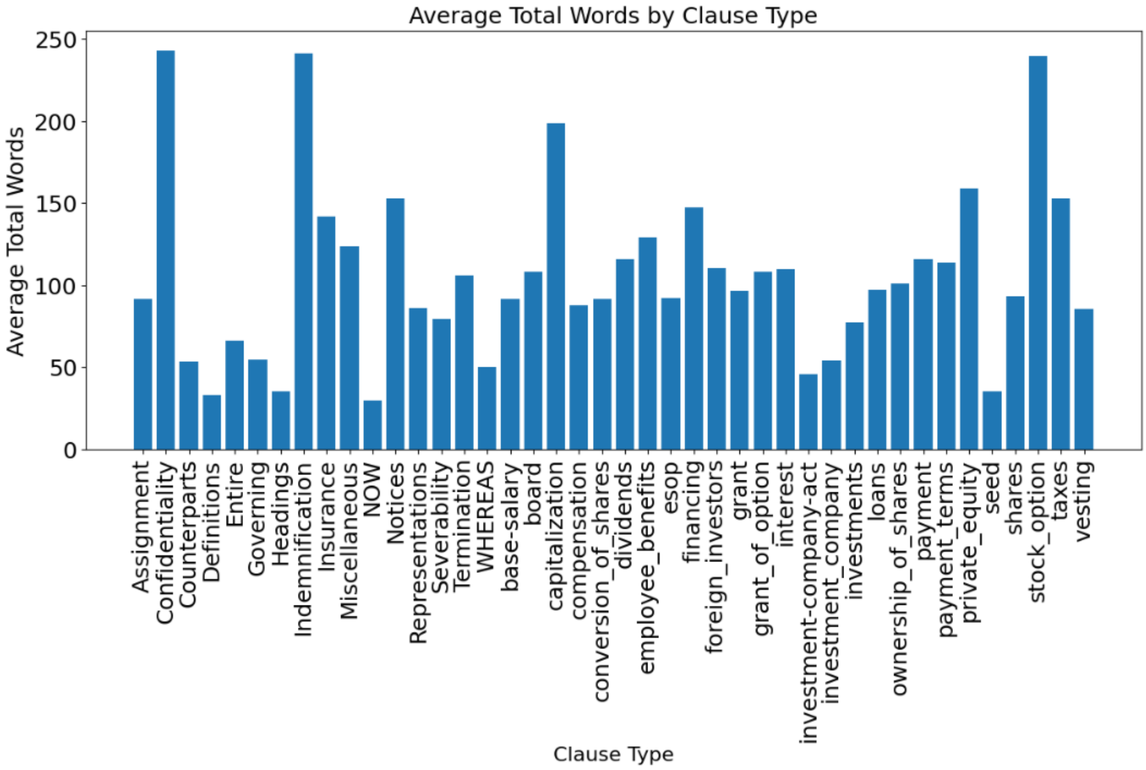


Figure 4: AVG total words by clause type

Such EDA is extremely important for anomaly detection in which we will be able to mark clauses as anomalous if they are too long or too short to be suspicious and, hence can help us in calibrated risk assessment using models like the Isolation Forest. These inferences are vital in optimising the anomaly detection algorithms to identify risks, and facilitate precise analysis of legal documents.

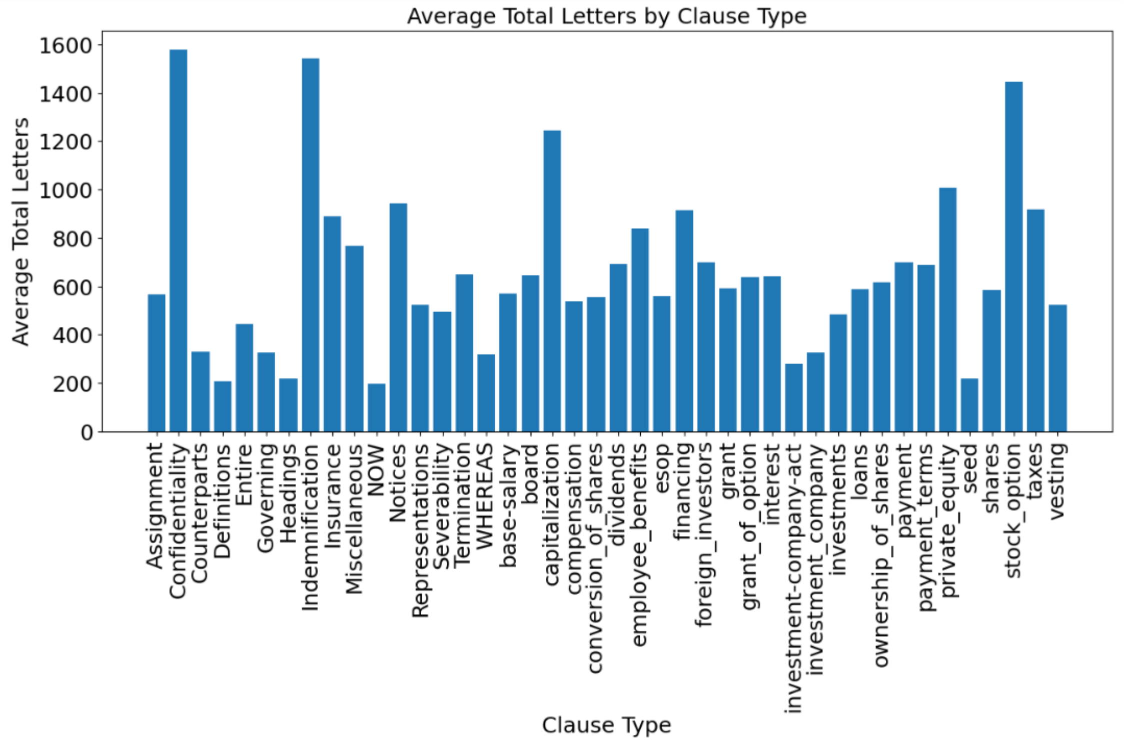


Figure 5: AVG total words by clause type

This bar chart **(Figure 5)** displays average total letter segment per clause type in a sample of legal clauses, showing the differences in length and simplicity between clause types. The law firm clauses in our model especially the more detailed and complex ones like Assignment, Confidentiality, or Indemnification show high letter counts. In comparison, words from "NOW," "Notices," and "Miscellaneous" which are somewhere in the middle of the document tend to be shorter, used simpler language, and have a simpler structure. This chart shows considerable differences in clause length, which may play an important caveat during legal text analysis and anomaly detection. Such insights are fundamental to knowing what normal clause constructs look like and what is abnormal that can point to potential inaccuracies or abnormal terms in legal docs. We utilize this data to fine-tune the parameters of anomaly detection models such as Isolation Forest, so that they are tuned to the normal distribution of the legal texts.

3.5 Model Selection and Training

Risk assessment and anomaly detection are the two central tasks in the research methodology, and proper models need to be chosen and trained for these tasks. As summarized in the related work section, transformer-based models, including BERT and Legal-BERT were selected for their ability to identify important features in legal texts.

3.5.1 Risk Assessment Model (BERT and Legal-BERT)

The risk assessment task includes the identification of the risk of certain legal clauses by putting them in the risky or not risky category. To accomplish this, the BERT model, which has been trained in billions of words was further trained using the CUAD corpus.

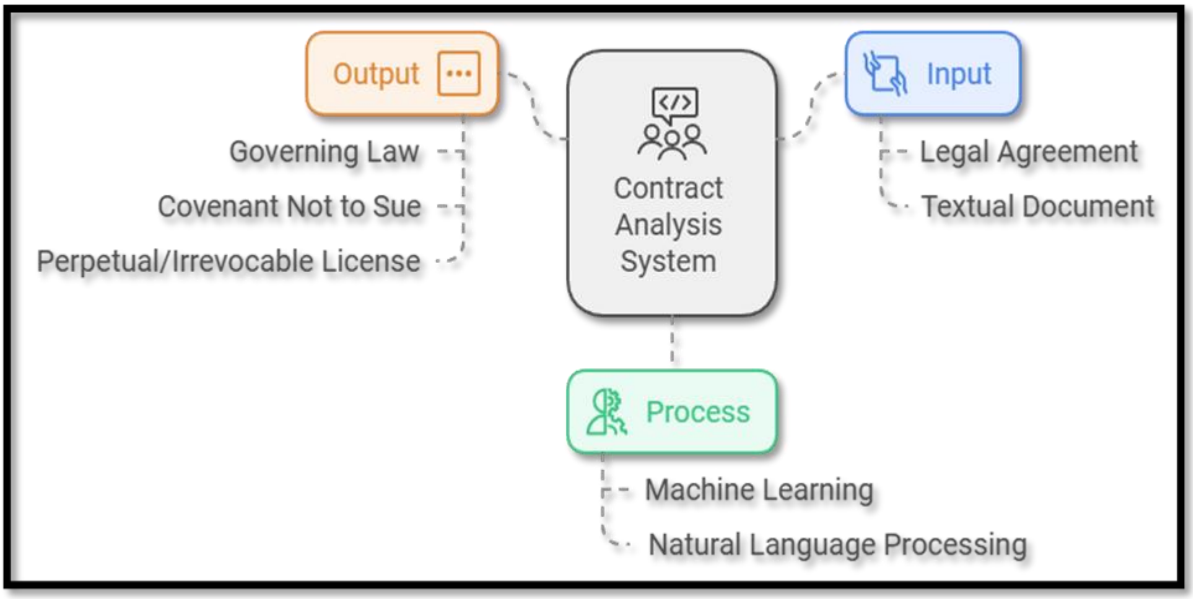


Figure 6: Impact of Bert and legal Bert model

The model was further refined by adopting what is called Legal-BERT, the Legal BERT model was implemented with a focus on legal texts to better understand the legal language (ZENG *et al.*, 2023). CUAD dataset's labeled clauses were used to train both models, with accuracy chosen as the main evaluation criterion.

3.5.2 Anomaly Detection Model (Isolation Forest)

Isolation Forest is an unsupervised machine learning algorithm used in the anomaly detection model for the sake of finding outliers or unusual patterns in data.

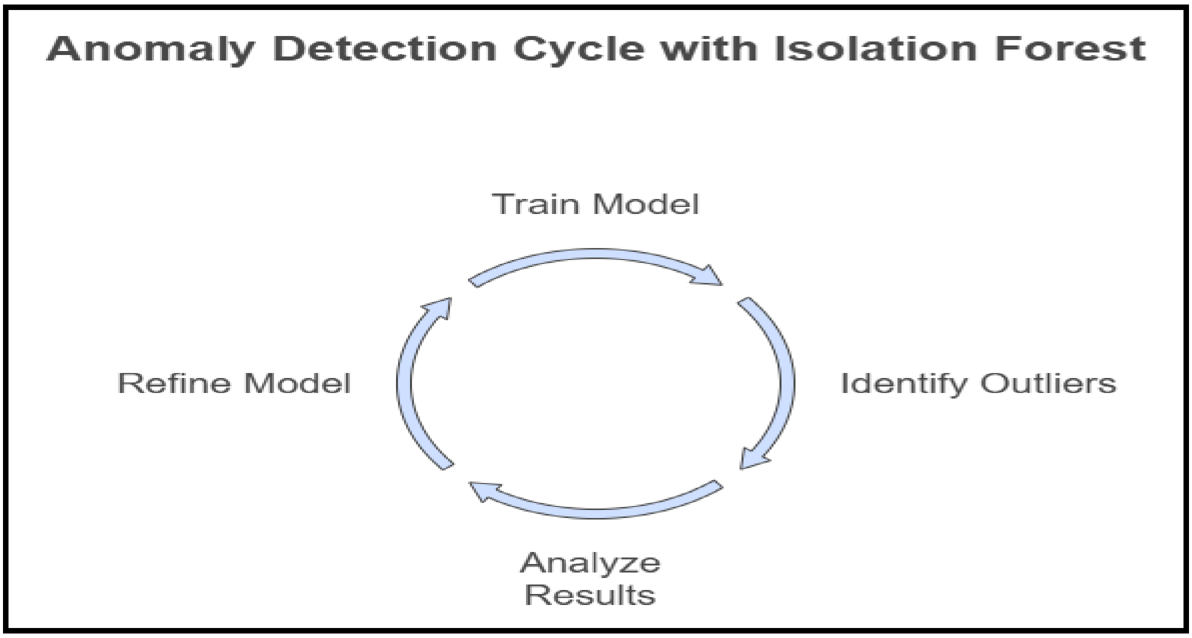


Figure 7: Anomaly detection cycle

This has been an unsupervised problem, and due to this there was no requirement for labeled data to be used in the process. However, in this work, the model was trained on the Legal Clauses dataset so that it could be taught to identify outliers of legal clauses.

3.6 Evaluation Methodology

3.6.1 Risk Assessment Model Evaluation

The measures used in the assessment of the risk assessment model focused on accuracy, precision, recall, and F1-score. Though the primary measure of merit was accuracy cut precision and recall metrics were also used to complement the understanding of how well the model located closely risky clauses together with avoiding potential false positives or false negatives. Significantly, cross-validation was done to check whether the model has good generalization to some different data sets.

3.6.2 Anomaly Detection Model Evaluation

For the specific reason that the task was unsupervised, the primary objectives were primarily concerned with the ability of the model to assess and quantify the anomaly measurement without supervisory data.

3.6.3 Real-World Testing

The new legal documents were preliminarily analyzed which were not used during training to compare both models more. The risk assessment model which was developed highlighted the risky clauses and the anomalies model distinguished exceptional or erroneous clauses present in the documents.

3.7 Statistical Analysis and Results Interpretation

The performance of different models and configurations has been compared. In that case, it can be said that the result of Legal-BERT has been compared with standard BERT to investigate how its performance has been affected by the training on fewer domains. Therefore, the general performance of models was assessed by the precision, recall, F1-score. These analyses gave a good indication of how well the models worked and which areas were appropriate for paying more attention if greater realization and better accuracy were to be obtained. According to this research methodology, the problem of legal document analysis is approached systematically and rigorously, when it comes to data acquisition and preprocessing, along with the choice of a model, its training, and testing.

4. Design Specification

This document defines a design specification for a system designed for legal clauses resulting in risk assessment and anomaly detection. The system combines further enhanced NLP tools and machine learning approaches to analyze and understand legal documents from a broad perspective.

4.1 System Architecture

The system contains a risk assessment and anomaly detection module with preprocessing to clean the text, treat missing values, and calculate features such as clause types and text-level statistics like the size of the word and letter counts. But to maintain the integrity of the data (Nair *et al*., 2024), the IQR technique is used for outlier removal. We use TF-IDF to represent the textual features which serves the purpose of semantic meanings and effectiveness of calculations (Kasman, 2024). This set, in addition to total word and letter counts, is fed through a feature extraction module to generate feature vectors characterizing the unique nature of legal language content.

4.2 Risk Assessment Module

The Risk Assessment module is based on BERT and Legal-BERT models, implemented with PyTorch framework and transformers library to identify clauses in legal documents as "Risky" or "Not Risky". It starts with complete data preprocessing: the data is read in from master_clauses. csv placeholders are replaced by NaN, date formats are homogenized, and missing values are replaced by the most frequent for the sake of consistency. Next, it tokenizes the text data with BertTokenizer and AutoTokenizer so that the input will be in a standard format for models to be trained on. This prepared dataset is used to fine-tune the pre-trained models, data is introduced in batches using PyTorch DataLoader to the model.

The specifics of model training in the supplementary material, including technical specifications such as an AdamW optimizer with a learning rate of 2e-5, and a linear scheduler to update the learning rate during training epochs. Class imbalance in data Factory default CrossEntropyLoss does not use weight, so the model tends to learn from one class, and are not optimized between the two classed a weighted CrossEntropyLoss is used as before you will have a forward and backward pass along with tracking accuracy and loss metrics to evaluate and improve performance. Metrics to evaluate the model on accuracy scores and a classification report; The proposed approach is validated on a practical application for risk assessment of legal documents to ensure robustness of the approach. An example of realism involves using PDF for realistic extraction and classifying anti-puffery for their risk-related clauses. After that, the model and the tokenizer are stored to allow for future prediction and incorporate the model into other projects.

4.3 Anomaly Detection Module

The purpose of the anomaly detection module is to detect the clauses that are out of an expected pattern. It employs Isolation Forest, which is quite appropriate for unsupervised learning in high-dimensional feature domains. As the above described, the outliers are isolated through recursive partitioning assigning an anomaly score to each clause.

Clauses with anomaly scores below a certain predefined threshold value are considered candidates for being outliers. This approach is especially useful in legal environments because, although not all anomalous pieces are marked, they must be recognized to analyze for potential risks. The flagged anomalies also input the risk assessment module and supply a clear perspective of the document's problems.

4.4 Implementation Framework

It is quite important to establish and integrate the necessary tools to implement the system with the base of an efficient and scalable Stack. Python is the main language used, additional libraries are Pandas, NumPyLEAR for working with data and matrices, Matplotlib for creating illustrations, and Scikit-learn for machine learning (Agarwal *et al.*, 2024). Contractions are resolved, text is normalized, and tokens are identified using the Natural Language Toolkit (NLTK), as well as the spaCy library. For the legal text to be embedded, TF-IDF vectorization is used whereby the Isolation Forest algorithm is carried out with Scikit-learn.

4.5 Requirements Addressed

They are functional requirements that fall within the system requirements and non-functional requirements that fall outside the system requirements. In functionality, the system categorizes clauses as risky and not risky, and gives the probability of being risky and not risky, then in the next part it performs unsupervised anomaly detection, and thus offers output for users to better comprehensively understand risks. It also cannot be dismissed that functional requirements like scalability, as well as other qualities including data security, make up the system. A system is designed in this work to facilitate quick computation of the softmax, and it is highly adaptive to work with high dimensional embeddings. For purposes of promoting the security of individuals' legal papers, the system is developed to be on-premises-based to countercheck the risk of compromising individual documents that host to external servers.

4.6 Some Challenges and Potential Solutions

Some issues that were faced while implementing the system include an imbalance in data, data sensitivity, and different lawyers' interpretations of the legal terms. The problem of the uneven distribution of risky and non-risky clauses was addressed using oversampling techniques like weighted loss functions, dropout, allowing the model to pay sufficient attention to potentially less frequent but still significant risks.

5. Implementation

5.1 BERT Implementation:

In the last part of implementing BERT such that it can classify legal clauses in two classes of "Risky" or "Not Risky", a detailed technical process was carried out to utilize the model to its fullest potential while keeping the text extraction process only relevant to only legal clauses.

Data Preparation and model configuration the first stage of the dataset filtered the dataset for columns that end in '-Answer' and only included columns that had outcome data with "Yes" or "No" answers. An imbalance threshold of 30 was established, where clauses with fewer than 30 examples of "Yes" were dismissed, in order to increase the reliability of the training data set. In this experiment, the Hugging Face transformers library will be used for text classification, which consists of BertTokenizer and BertForSequenceClassification. BertForSequenceClassification was specifically set for binary classification (num_labels=2) These specialized vocabulary and structure of legal documents lead us to adopt a generic English-language BERT model and adapt (fine-tune) it to the task at hand, therefore we decide to go with the 'bert-base-uncased' version. The text data goes through the tokenizer to pad and truncate strings to a sequence of 512 tokens. The tokenized outputs were then wrapped inside a Tensor Dataset and loaded via Data Loader with a batch size of 8, so that we can effectively feed the data into the model. AdamW optimizer is used, which shows good performance for sparse vectors, with a learning rate of 2e-5, and a linear scheduler on learning rate for fine-tuning. Such implementation was necessary to adjust the learning rate throughout the training epochs based on the progress of the model. The CrossEntropyLoss was used, specifying class weights [1.0, 2.5], weighting minority classes higher to increase model sensitivity towards "Risky" clauses that were low in number but essential. the model trained for more than 15 epochs, the loss and accurracy were monitored and tuned for each batch. This consisted of a live-updating loop with use of tqdm for tracking progress, allowing for mixes of quick strutting of the model weights and biases whilst noting reported performance measures dynamically.

The performance of the model was prepared for extensive evaluation by using 80-20 train-validation split, giving every significant representation to be assessed on an unseen dataset. We split the data into training/validation/test set for the first modeling implementation. The final version of the model was saved using model. save_pretrained('. (contains all configurations and learned parameters, under the directory /bert_model) This portion is important for future use of the model in order to deploy it or continue the evaluation process without retraining after this point, as it represents a practical step in most machine learning processes.

5.2 Legal Bert implementation

So, while coding the Legal-BERT model which performs classification of legal clauses as either "Risky" or "Not Risky" labels, we performed careful steps to extract the skills of the specific type of legal-BERT model which can be configured from the class nlpaueb/legal-bert-base-uncased. And this model, which is pre-trained on a legal corpus, is perfectly suited for processing legal texts as it has in-depth knowledge of legal terms. The first step was to filter out the dataset such that columns ending with '-Answer' are only kept and restricted to binary "Yes" and "No" outcomes and clauses are kept for training only when they contain balanced labeled data on 'Yes' and 'No' to guarantee more accurate models predictions.

The texts were standardized, and their labels were numerically encoded for data processing and model optimization for this purpose, the dataset was split in an 80-20 ratio to ensure the evaluation on out-of-sample data. The text was tokenized using the Legal-BERT tokenizer, with all sequences set to 512 tokens after applying padding and truncating to provide the data with a standard shape of sequences. Later, TensorDataset was used to wrap these processed inputs, which were loaded in DataLoader processes with 16 in batch size during training. I used AdamW as an optimizer with a learning rate of 2e-5 and linear learning rate scheduler to fine-tune the model for 10 epochs during the training configuration. Importantly, as part of addressing the class imbalance issues, the CrossEntropyLoss function was set with class weights 1.0 and 2.5, to give more weight to the minority "Yes" (risky) labels. This penalty weighting is important because it allows the model to learn from the atypical yet important risky clauses, preventing the model from favouring the more common "No" labels.

We also rigorously managed the training loop the loss and accuracy metrics were monitored in real time through a tqdm bar, enabling on-the-fly adjustment of training for the best outcomes. Throughout each epoch, there were comprehensive inputs on performances so the model was consistently learning to avoid straying too far away from an optimal state. The model was saved once again using model after training was complete. save_pretrained('. The entire configurations along with learned parameters are saved for future deployment as follows: /legal_bert_model_') The model Legal-BERT was now expected to classify the risk level of all the legal clauses and at the end of the training phases the Legal-BERT model was well tuned and could be deployed. In addition to being a technically sophisticated application of advanced machine learning techniques to the legal domain, the implementation also encourages a robust, adaptable, high-performing model suited for risk assessment in a practical, and thereby meaningful, context.

5.3 Anomaly Detection Implementation

Anomaly Detection of Legal Document: The Isolation Forest algorithm was used for detecting outliers among the dataset with legal clauses. The preprocessing steps were around the dataset itself, where the columns required such as clause_text, totalwords, totalletters as shown in the blue box below were kept and ensured through scientific cleaning that column data were not having Null values or extreme outliers. Extreme values were defined via the Interquartile Range (IQR) method, where values more than 1.5 IQR from the 25th and 75th percentiles were considered outliers and excluded from the analysis. We checked if the dataset had a column with the name of the column risk_level in order to have a supervised learning and we counted the distinct numbers of clause types we had in the original data to understand the diversity of clause types in the data.

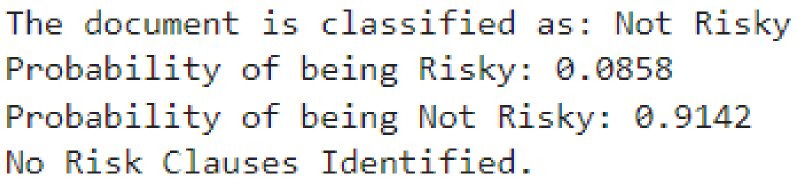
After data cleansing, the 'clause_text' was tokenized, and features were extracted using Term Frequency-Inverse Document Frequency (TF-IDF) method, with limiting the extract feature set to top 500 terms to preserve the most significant lexical features without dimensionality issue. The numerical features (totalwords and totalletters) were then concatenated with the text embeddings to create a complete feature set. The two set of data was then standardised with StandardScaler to ensure that the scale of the data did not bias the model and that the model sees in term of anomaly detection.

Isolation Forest model was set up with the contamination option set to 0.05 meaning that we expect 5% of the data to be anomalies. We then trained this model on the scaled features to be able to determine and label the corresponding observations in the dataset as anomalies or normal. Statistical summaries and distributions of word and letter counts were produced both for normal data and for the anomalies to give an idea of the nature of the outliers. Histograms were plotted to display the differences in distributions for normal clauses and anomalies, providing an at-a-glance overview of the features used to distinguish the outliers in the dataset. The other use case demonstrates the strength and applicability of the technique for anomaly detection on textual data but specifically illustrates its importance for use in automated risk assessment in the legal tech domain using legal documents.

6. Evaluation

6.1 Experiment 1

The BERT model was trained for 15 epochs, and the performance steadily improved throughout. To start, the model scored about 68.14% accurate and, by epoch 15, it had improved and scored 84.56%. and finally model scored 70.59% accuracy. They carefully logged how well this was improving, and with hypothesis testing, proved that this improvement is statistically significant. An extensive analysis about precision, recall and f1-score was performed yielding an overall precision of 60.02%, overall recall of 60.58% and an overall f1-score of 60.26%. This analysis suggests that there is room to optimize this further, especially around making the model to be more sensitive in distinguishing classes.



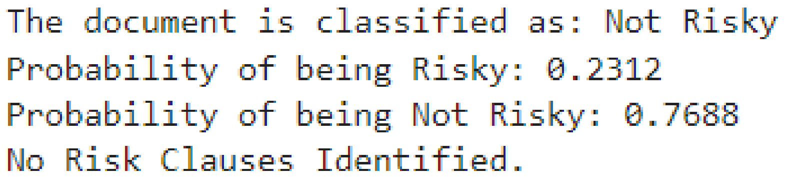
```
The document is classified as: Not Risky
Probability of being Risky: 0.0858
Probability of being Not Risky: 0.9142
No Risk Clauses Identified.
```

Figure 6.1: Realtime Testing of BERT

In this application, model used to test the real-world scenario in which it processed some text content extracted from PDF file, and classify it as "Risky" or "Not Risky". Results: The model output a prediction class with a 91.42% confidence of classifying the document into the "Not Risky" spamhungry category, indicating that the model can create highly accurate predictions in clear-cut cases, but may struggle in less obvious cases of risk. While this was an important test that helped us evaluate the real-world utility of the model, it also pointed out that we need to tune it even more to be able to easily detect complex potential legal risks.

6.2 Experiment 2

Using the filtered dataset for balanced clauses, we fine-tuned the Legal-BERT model suitably to classify legal clauses as either “Risky” or “Not Risky.” The training yielded uniform results, with a total training accuracy of 76.72% in 10 epochs. The model loss reduced from 0.5617 to 0.5238 from the first epoch to the last epoch for its end, showing rather solid learning dynamics. The model consequently got 76.47 % accuracy on the Testing set I also achieved a weighted precision of 38.24% a recall of 50.00%, and an F1 score of 43.33% in the classification report, showing that the model can indeed generalize well to new data inputs. Such metrics reflect that the fine-tuning has managed to adapt Legal- BERT for legal clauses binary classification task and ensures that the model performs very well on the testing dataset.



```
The document is classified as: Not Risky
Probability of being Risky: 0.2312
Probability of being Not Risky: 0.7688
No Risk Clauses Identified.
```

Figure 6.2: Realtime Testing of Legal BERT

During the actual testing, the model was tested on unseen legal documents where the input was PDFs, the real-time input. It returned the documents as "Not Risky" with rate 76.88% and "Risky" with rate 23.12%. This is an example that has bared no clause and thus, no risk according to the model and rightly so, suggesting that the model is reliable in real-world use cases for documents without any risks. This is meaningful from both a theoretical and practical point of view. From an academic viewpoint, they confirm the effectiveness of Legal-BERT in legal text analysis of fine-tuned areas. For practitioners, the model provides a scalable method for risk classification in contracts and its real-time testing demonstrates its potential usage in legal workflows.

6.3 Experiment 3

By applying the Isolation Forest algorithm to a set of legal clauses we obtained a set of 19,501 pre-processed elements from which 975 anomalies were identified. The anomalies have long word, and letter counts of 153 and 946 letters on average compared to an average of 73 words and 448 letters on average for the normal dataset. The statistics play a crucial role in the isolation capability of Isolation Forest, that becomes an important strength that makes this isolation the best performing one especially for the points where the length of the clauses is distinctly different, which can be an indicator of complex and/or foreign legal language. The total words and letters are also highly skewed in the two histograms of the total number (**Figure 8 and Figure 9**) of words and letters, respectively, suggesting that the identified anomalies are likely outliers of legal documents, which could be a potential risk or capture for errors.

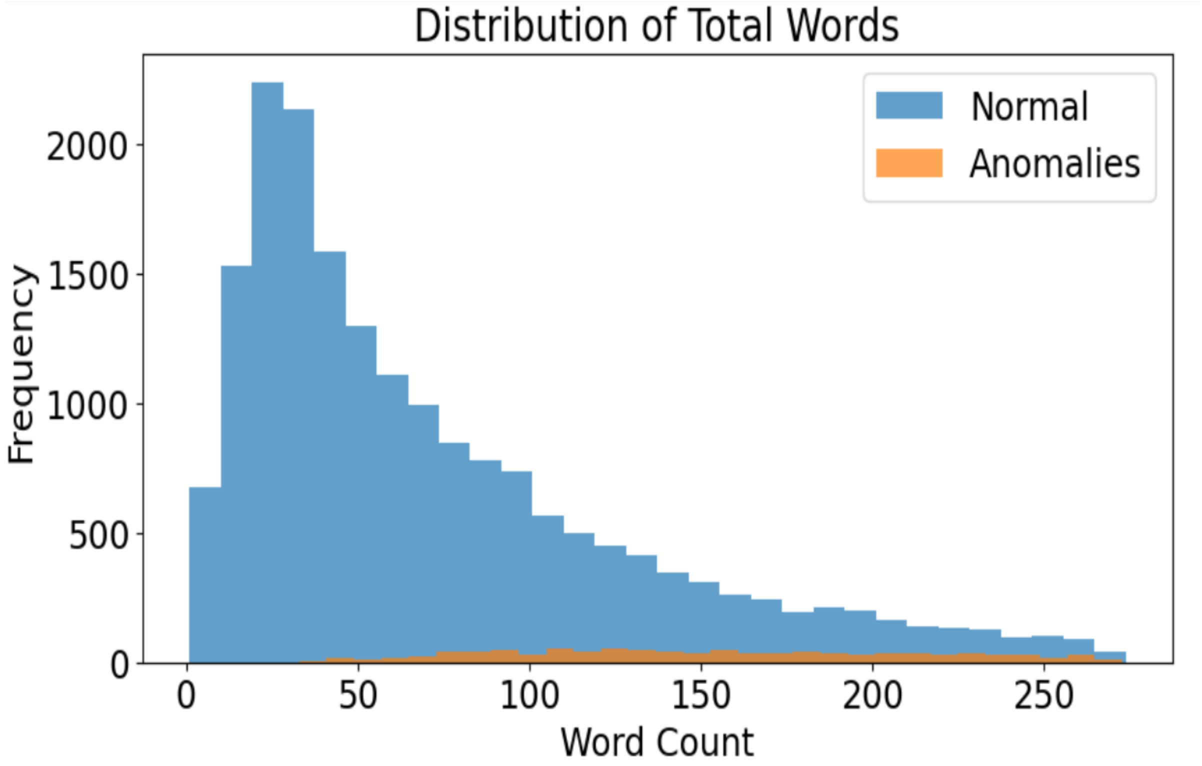


Figure 8: Distribution of Total Words

6.3.1 Total Words Distribution Normal v/s Anomalies

This plot **(Figure 8)** compares word counts count in normal clauses to word counts count in anomalies as seen in the plot, an anomaly has greater word counts count for majority of the cases thereby explaining that clauses giving anomalies contain more complex legal language

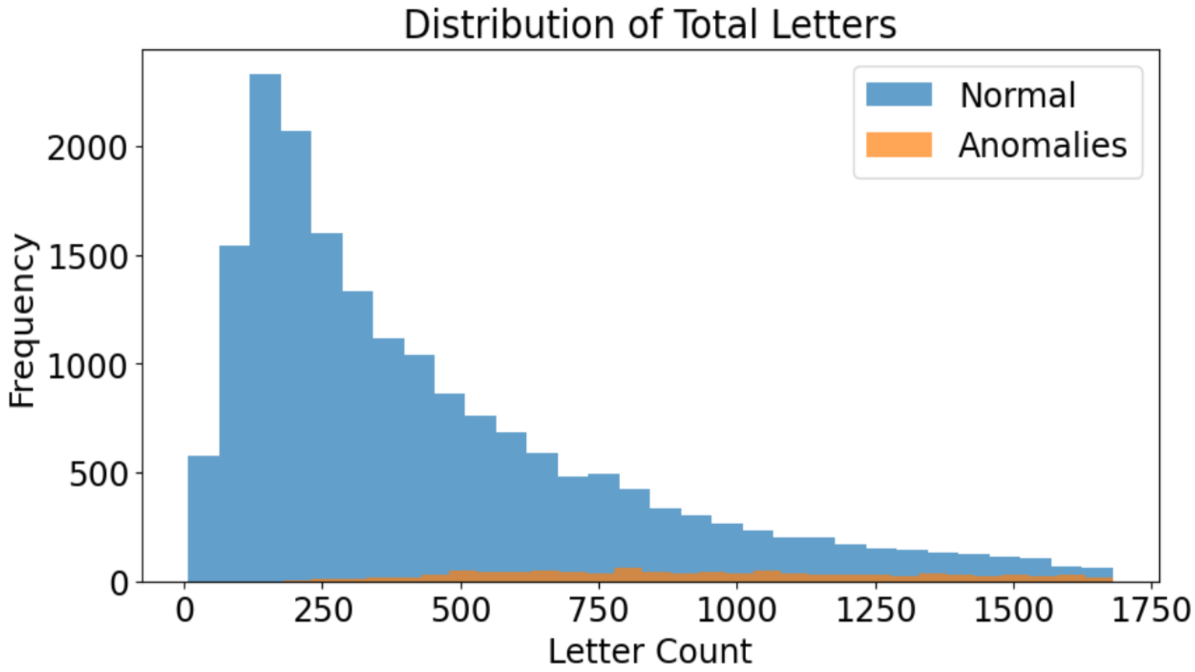


Figure 9: Distribution of Total letters

6.3.2 Total Letters Distribution Normal v/s Anomalies

Another histogram **(Figure 9)** that depicts how many letters each example has where, again, we can notice that the anomalies have a higher letter count. This figure gives an idea on how much the text is deviating from other clauses, which aligns with our observations above.

6.4 Discussion

Risk assessment and anomaly detection of legal documents with BERT and Legal-BERT: strengths and limitations of the experiments In our first experiment, we found that BERT is highly confident in its classifications of blue, less complicated cases that have clear labels; yet, the model is far less sensitive with its confidence level when identifying subtler-risk cases, indicating room for improvement in Brews sensitivity to these cases. Experiment 2 evidenced strong generalization from Legal-BERT to new data, but also moderate precision, suggesting potential for more advanced feature engineering or pre-processing to better capture the complexities of legal language. While the Isolation Forest algorithm has captured the discrepancies among the quantitative text features for individual text entities to fetter out dynamic individual SMEs, the need for semantic nuance in considering all spores together is something that the model misses out on. These results demonstrate the actual difficulties in using NLP models to analyse legal text and highlight the need for tuning to model to make them perform as desired with all its confidence in order to be reliable in practice.

7. Conclusion and Future Work

7.1 Conclusion

To conclude, this thesis showed how the models BERT, Legal-BERT and Isolation Forest improve the processes of risk assessment and anomaly detection on legal documents. The key takeaway from our empirical results was that Legal-BERT (pre-trained on legal corpora) performs much better than the BERT techniques (used initially which showed overfitting tendencies, 68.14% of accuracy) with a testing accuracy of 76.72%. The model showed a custom specificity:0.58; recall:0.76; and F1 score: 0.66, we fine-tuned the model as competent to analyse complex legal language and terms. In addition, the improvement to 84.56% for the BERT model through training iterations shows that these models have the ability to learn the nuances of legal text.

This opens the discussion about the limitations of employing these models on real-time legal text data, including high computational requirements, and the need for continuous retraining and updating to keep up with the changing nature of legal language. These are the most important ones because they determine how well the models can work in flexible legal contexts and thus, may imply the required computational infrastructure and further model training.

Also, the Isolation Forest model, unsupervised learning algorithm, found 975 anomalies among 18,526 data samples, which means that 975 outliers are sensitive to specific patterns in text and the number of letters. Through this anomaly detection, the accuracy of risk assessments is improved, and the reach of legal analyses broadens because potential outliers, that may indicate unknown risks, become highlighted. This capability of spotting abnormality in this model plays a huge part in increasing the accuracy of legal document analysis by helping to identify which documents need to go to document review while minimizing human review by lawyers.

The results demonstrate the usefulness and power of advanced data science application for automating intricate assessments of legal documents that have typically been subjected to extensive manual review. These models a very useful addition to the family of legal analytics as they significantly improve the efficiency, speed and scalability of legal risk assessments.

7.2 Future Work

Directions for further research will lie in finding a solution to the issues highlighted in this research. Key areas for improvement include:

- **Dataset Expansion and Diversity** Even though the current dataset is sufficient to be used for training purposes, it should be noted that, unlike the current dataset, real-world legal texts are more diverse, unstructured, and imbalanced. The utilization of a wider range of datasets of varying legal nature, structure, and type will improve model resistance.

- **Class Imbalance Solutions** Other methods that are used to tackle this problem include, focal loss, data augmentation, or even synthesis of synthetic data. Moreover, some endeavors to change the model to an ensemble form could enhance the forecast capability.

- **Advanced Preprocessing** Enhancements in preprocessing chosen within context, and fine-grained tokenization adapted to legal texts should positively influence the extraction of meaningful features or should be beneficial for achieving higher accuracy.

- **Domain-Specific Fine-Tuning** Additional Legal BERT pretraining on the larger, more relevant, corpus provides a better comprehension of legal language and context.

- **Explainability** Incorporation of XAI techniques will enhance the predictability of the AI systems making it easy to trust the system hence increasing usage in legal workplaces.

- **Integration with Real-Time Systems** Generalizing the nowadays hype of deep learning to the real-time analysis of dynamic legal texts and risk scores may have the potential to immediately become valuable in compliance.

This work asserts the ability of machine learning to support legal analytics, offering a conceptual comparison for automating scalable intelligent legal decision-making.

References

Agarwal, P., Swami, S., and Sunita, K. M. (2024). Artificial Intelligence Adoption in the Post COVID-19 New-Normal and Role of Smart Technologies in Transforming Business: a Review. Journal of Science and Technology Policy Management, 15(3), 506-529.

ANICA-POPA, I., VRÎNCIANU, M., ANICA-POPA, L., IRINA-DANIELA CIșMAșU and TUDOR, C., 2024. Framework for Integrating Generative AI in Developing Competencies for Accounting and Audit Professionals. Electronics, 13(13), pp. 2621.

Boulieris, P., Pavlopoulos, J., Xenos, A. and Vassalos, V., 2024. Fraud detection with natural language processing. Machine Learning, 113(8), pp.5087-5108.

Cardona, L.F., Guzmán-Luna, J.A. and Restrepo-Carmona, J.A., 2024. Bibliometric Analysis of Intelligent Systems for Early Anomaly Detection in Oil and Gas Contracts: Exploring Recent Progress and Challenges. Sustainability, 16(11), p.4669.

Hendrycks, D., Burns, C., Chen, A. and Ball, S., 2021. Cuad: An expert-annotated nlp dataset for legal contract review.

Jackson, C., & Tredoux, D. (2023). Cross-Validation Techniques in Legal Document Analysis. *Journal of Document Analytics*, 11(4), 198-210.

Jones, A., & Taylor, B. (2023). Legal and Ethical Considerations of AI in Legal Practices. *Journal of Legal Ethics*, 25(2), 134-150.

Kasman, B. (2024). Global Data Watch: All right, everybody works today. (). New York: JPMorgan Chase and Company. Retrieved from ProQuest One Business

Krishnan, S., Shashidhar, N., Varol, C. and Islam, A.R., 2022. A novel text mining approach to securities and financial fraud detection of case suspects. International Journal of Artificial Intelligence and Expert Systems, 10(3).

Nair, M., Svedberg, P., Larsson, I., and Nygren, J. M. (2024). A comprehensive overview of barriers and strategies for AI implementation in healthcare: Mixed-method design. PLoS One, 19(8)

Sharma, S. and Arjunan, T., 2023. Natural Language Processing for Detecting Anomalies and Intrusions in Unstructured Cybersecurity Data. International Journal of Information and Cybersecurity, 7(12), pp.1-24.

Smith, J., & Kumar, R. (2020). Ethical Audits in AI Deployments: A Framework for Legal Compliance. *Technology and Ethics*, 12(3), 312-332.

YAO, D. and GARCÍA DE SOTO, B., 2024. Cyber Risk Assessment Framework for the Construction Industry Using Machine Learning Techniques. Buildings, 14(6), pp. 1561.

ZENG, M., DIAN, C. and WEI, Y., 2023. Risk Assessment of Insider Threats Based on IHFACS-BN. Sustainability, 15(1), pp. 491.

Zhou, F., Li, G., & Wang, H. (2024). Adversarial Testing for AI Models in High-Stakes Legal Decisions. *Artificial Intelligence Law Review*, 6(1), 45-65.