EMPOWERING INDIAN LEGAL NLP: ADVERSARIAL SAMPLING AND MULTI LABEL CLASSIFICATION WITH LARGE LANGUAGE MODELS

A PROJECT REPORT

Submitted by

Shanmukha Naveen K (205002087)
Sharon Roshini S (205002088)

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY IN INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATIONTECHNOLOGY

Sri Sivasubramaniya Nadar College of Engineering

(An Autonomous Institution, Affiliated to Anna University)

MAY 2024

Sri Sivasubramaniya Nadar College of Engineering

(An Autonomous Institution, Affiliated to Anna University)

BONAFIDE CERTIFICATE

Certified that this Report titled "**EMPOWERING INDIAN LEGAL NLP: ADVERSARIAL SAMPLING AND MULTI LABEL CLASSIFICATION WITH LARGE LANGUAGE MODELS**" is the bonafide work of **Shanmukha Naveen K(205002087)** and **Sharon Roshini S (205002088)** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

Dr. A. SHAHINA

Professor and Head of Department

Department of Information Technology

Sri Sivasubramaniya Nadar College of Engineering
Kalavakkam – 603 110

Dr. S. KARTHIKA

Associate Professor

Department of Information Technology

Sri Sivasubramaniya Nadar College of Engineering

Kalavakkam – 603 110

Submitted for project viva-voce examination held on.....

EXTERNAL EXAMINER **INTERNAL EXAMINER**

ACKNOWLEDGEMENT

I thank **ALMIGHTY GOD** who gave me the wisdom to complete this Project. My sincere thanks to our beloved founder **Dr. SHIV NADAR, Chairman, HCL Technologies.** I also express my sincere thanks to **Ms. KALA VIJAYAKUMAR,** President, SSN Institution and our Principal **Dr. V.E. ANNAMALAI**, for all the help he has rendered during this course of study.

We are highly indebted to **Dr. A. SHAHINA, Head of the Department** for providing us with the opportunity and facilities to take up this project.

I am deeply obliged and indebted to the timeless help and guidance provided by **Dr. S. KARTHIKA, Associate Professor**, Department of Information Technology and also express my heartfelt thanks for making this project a great success.

I also thank all the faculty of the Department of Information Technology for their kind advice, support and encouragement and last but not the least I thank my parents and my friend for their moral support and valuable help.

ABSTRACT

Legal NLP, a field within natural language processing , involves techniques to understand and analyze legal texts that are filed under statutes. These Natural Language Processing techniques are widely applied for tasks such as identification, prediction, and summarization within legal documents. This paper focuses on multi-label classification using four models (Quantum BERT, Bidirectional Encoder Representations from Transformers, Bidirectional Long Short Term Memory, Gated Reccurent Units) with attention mechanisms trained on the Indian Legal Statute Identification dataset, comprising 100 sections, particularly working with women-related sections. The models are tested against adversarial attacks to assess their robustness. Despite the challenges posed by adversarial inputs, Quantum BERT, with its efficient circuit structure and computations, performed well with an accuracy of 0.85. Bidirectional Long Short Term Memory demonstrates exceptional accuracy of 0.90, followed by the other two models, showcasing their resilience in legal text analysis even under adversarial conditions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure 7.1 Heat Map

LIST OF ABBREVIATIONS

BERT - Bidirectional Encoder Representations from Transformers

BiLSM - Bidirectional Long Short-Term Memory

ILSI - Indian Legal Statute Identification

GRU Gated Reccurent Units

IPC - Indian Penal Code

CHAPTER 1
INTRODUCTION

In the specialised field of legal natural language processing (NLP), information is analysed, comprehended, and extracted from legal texts and documents using computational techniques. It includes an extensive array of methods intended to manage the intricacies present in legal language, including contracts, statutes, rules, court cases, and legal opinions. Legal professionals [2] have enormous hurdles in processing and analysing the vast amount of textual data that is generated on a daily basis in the legal area. In response to these difficulties, legal natural language processing (NLP) tools and techniques have been created, providing answers for legal research, document summarization, contract analysis, case law prediction, and legal document classification.

Legal NLP techniques are employed in legal document classification, allowing for the categorization of legal documents based on various criteria such as case types, legal topics, or document types. This classification enhances the organization and retrieval of legal information, making it more accessible and manageable for legal professionals. Understanding papers, which entails parsing legal documents to understand them, is one of Legal NLP's main goals .

Natural language processing (NLP) has a subfield called "legal NLP," which studies methods for deciphering and interpreting legal texts filed under statutes. NLP approaches are frequently used for tasks like summarising, predicting, and identifying information included in legal texts. In this paper,

we specifically deal on women-related areas of the Indian Legal Statute Identification (ILSI) dataset to train three models (BERT, BiLSTM, and GRU) using attention mechanisms for multi-label classification. To evaluate the robustness of the models, adversarial sampling is used as a test. Even in the face of adversarial inputs, BiLSTM outperforms BERT and GRU in terms of accuracy, showing that these models are resilient in the face of adversarial inputs when it comes to legal text analysis.

Legal Natural Language Processing (NLP) has become indispensable in various countries, revolutionizing legal practices from research to contract analysis and beyond. In the United States,[34] Legal NLP has significantly impacted legal research and analysis. Industry giants such as LexisNexis and Westlaw have integrated NLP techniques into their platforms, bolstering search capabilities and providing more efficient access to legal information [28]. The United Kingdom's legal industry has also embraced Legal NLP, particularly in contract analysis and due diligence. Firms like Luminance and Eigen Technologies have developed AI-powered platforms that leverage NLP to extract and analyze crucial information from contracts, streamlining the contract review process and enhancing accuracy [29]. In Australia, Legal NLP has gained traction, particularly in legal research and information retrieval. AustLII (Australian Legal Information Institute) has developed NLP-based search engines, allowing users to access and search through vast collections of Australian legal documents and case law with ease [30].

In Indian context , the utilization of Legal NLP in various facets of its legal system is prevalent recently. The National Judicial Data Grid (NJDG) employs NLP techniques to analyze case data from across the country,

offering insights into judicial trends and case outcomes. This application of NLP has contributed to more efficient case management and analysis [31].

1.1 LEGAL DECISION FRAMEWORK

In this domain , a new way of making decisions is emerging. It combines fancy language understanding with studying past cases to help lawyers work smarter and faster. This decision framework is depicted in Figure 1.

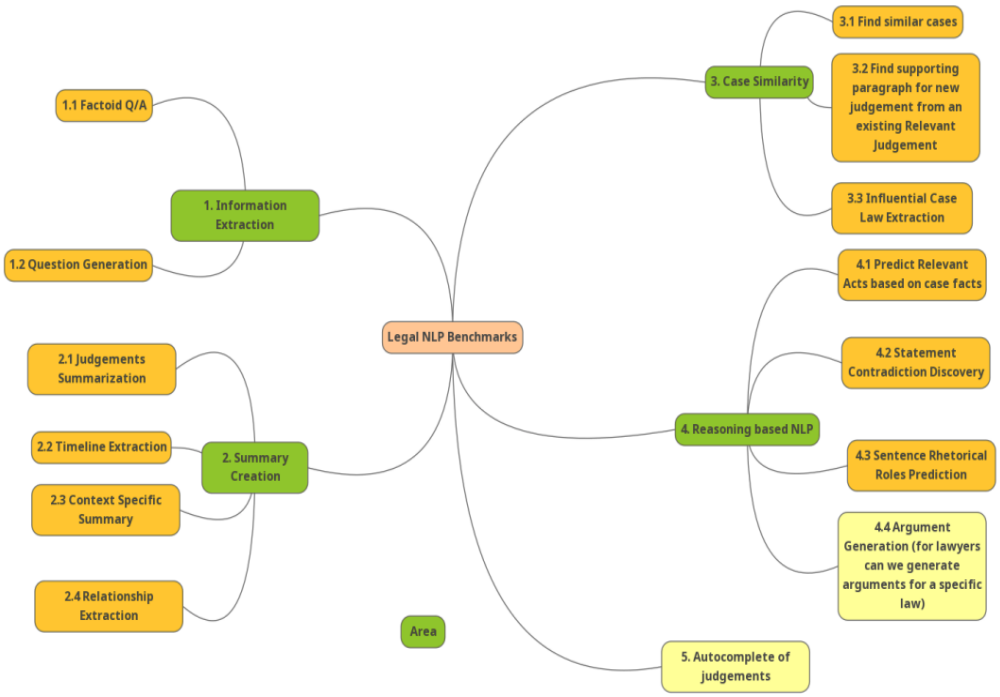

FIGURE 1 *Legal Decision Support Framework*

The process begins with information extraction, where various tasks are performed to gather relevant data. This includes question generation, which involves formulating queries based on the context of the legal case, and factoid question answering, which aims to extract specific pieces of

information from legal documents or databases. Moving on, the system proceeds to case similarity assessment. Here, it seeks to identify cases that are similar to the current one. This involves finding similar cases from historical records and extracting supporting paragraphs from these cases to inform the decision-making process for the new judgment.

Subsequently, the process involves influential case law extraction. This step entails analyzing past rulings to identify legal precedents that may be relevant to the current case. By leveraging historical legal data, the system can provide insights into how similar cases have been decided in the past, helping to inform the current judgment. Following influential case law extraction, the process moves into legal reasoning steps. In this section, the system predicts relevant laws or acts based on the facts of the case. Additionally, it performs tasks such as statement contradiction discovery, where inconsistencies or conflicts within the legal documents are identified, and argument generation, which involves generating persuasive arguments based on the available evidence and legal precedents.

Finally, the process concludes with the autocomplete of judgments. Here, the system assists in generating complete judgments based on the analysis conducted in the previous steps. By automating this aspect of the legal process, the system aims to streamline decision-making and improve efficiency in case management. Overall, the case management process described focuses on leveraging large language models (LLMs) and natural language processing (NLP) techniques to automate various tasks involved in legal information processing. From information extraction to legal reasoning

and judgment generation, these technologies play a crucial role in enhancing the efficiency and accuracy of legal decision-making processes.

1.1 OVERVIEW

Legal information is primarily in text form, making legal text processing an increasingly important area of research in natural language processing (NLP). This includes tasks such as crime classification, judgment prediction, and summarization. In countries like India with high population densities, there is a large number of pending legal cases, estimated at around 41 million [15]. This backlog is due to various factors, including the shortage of judges. A legal statute identification system , judgment prediction system and classification system could assist in several aspects, such as retrieving relevant articles or case histories and determining penalty terms. However, the accuracy of such systems is crucial, as even small errors could significantly impact judicial fairness. While many researchers [23] have focused on developing legal identification and judg- ment prediction systems using NLP models like LSTM, BERT, and legal-BERT trained on legal datasets, little attention has been given to ensuring the robustness of these models.

This research is driven by the huge amount of legal data and the need to classify it. Traditional legal research takes a lot of time and effort, slowing down legal corpus classification. Legal NLP (Natural Language Processing) offers a new way to speed up this work by automatically finding important information in legal documents. [13] This can help law enforcement agencies find charges or crimes faster, compare cases more accurately, and pick out key phrases that sum up legal situations. Essen- tially, Legal NLP could

change how law is practiced in India, making it faster and easier to understand complex legal issues. Hence, this paper focuses on performing multi label classifi- cation using a subset of ILSI dataset , particularly focusing on women related sections. An adversarial algorithm is introduced on the test dataset so as to check the performance of the model even though pertubations are introduced in dataset. This is performed using highly resilient models BERT , BiLSTM and GRU . These models overperforms the pre exisiting models performance and its accuracy.

1.2 OBJECTIVE

- To utilize NLP models to tackle the drawbacks in the classification of the Indian legal system.
- To explore the working of Quantum BERT and train this model to perform classification task.
- To ensure NLP models' (BERT , BiLSTM , GRU) accuracy to prevent errors in legal information processing, vital for maintaining judicial fairness.
- To test models with adversarial sampling to evaluate their resilience and ability to perform accurately despite potential data manipulation.

1.3 MOTIVATION

The Indian legal system grapples with a massive backlog of cases, hinders timely resolution. Traditional legal research[3], relying heavily on manual efforts, further exacerbates this issue. This necessitates innovative solutions to navigate this complex legal landscape.Legal NLP (Natural Language Processing) offers a promising approach. By applying NLP techniques to legal text

analysis, we can automate crucial tasks like identifying relevant statutes, predicting case outcomes, and generating summaries. These capabilities can significantly reduce research time and effort for legal professionals, ultimately leading to faster case resolution.

However, the accuracy of these NLP models is paramount. Even minor errors in legal information processing can have severe consequences, impacting judicial fairness. This underscores the importance of ensuring adversarial robustness. Adversarial attacks are deliberate attempts to manipulate data, aiming to deceive machine learning models. By testing our models against such attacks, we evaluate their resilience and ability to perform accurately even with corrupted data. This research focuses on multi-label classification using highly robust models like QuantumBERT , BERT, BiLSTM, and GRU. We train these models on a subset of the ILSI dataset, focusing specifically on women-related legal sections. By introducing adversarial attacks on the test data, we assess the models' performance under challenging conditions.

In essence, this research strives to develop robust Legal NLP models, mainly focusing on the developing QuantumBERT, capable of accurate and reliable legal text analysis, even in the face of potential data manipulation. This will contribute to streamlining legal processes, ensuring better access to justice, and promoting fairness within the Indian legal system.

1.4 ORGANISATION OF THE REPORT

Chapter 1 Deals with the Introduction of the project

Chapter 2 Deals with the literature survey carried out and the results of the investigation.

Chapter 3 Deals with the dataset distribution.

Chapter 4 Deals with design of the system.

Chapter 5 Deals with the results and discussion of the pre-existing trained models.

Chapter 6 Deals with the implementation of QuantumBERT.

Chapter 7 Deals with the performance analysis of the robust model(BiLSTM).

Chapter 7 Deals with the conclusion and future directions

CHAPTER 2
LITERATURE SURVEY

2.1 RELATED WORKS

In their recent study, "A Heterogeneous Graph-based Approach for Automatic Legal Statute Identification from Indian Legal Documents," presents an innovative method that integrates textual analysis with legal citation networks for Legal Statute Identification (LSI)[1]. Their LeSICiN model, developed and trained on a carefully curated dataset, demonstrates superior performance compared to existing approaches by leveraging both graphical structures and textual features. This novel approach highlights the potential of incorporating heterogeneous graph-based techniques to enhance the accuracy of LSI tasks, offering valuable insights for researchers and practitioners in the field of legal text analysis.

Another survey, introduces the ILDC dataset, titled "ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation," consisting of 35,000 Indian Supreme Court cases annotated with original decisions. Their study[2] centers on Court Judgment Prediction and Explanation (CJPE), where the authors explore baseline models and propose a hierarchical occlusion-based model for enhancing explainability. Through their experiments, they shed light on the challenges inherent in aligning algorithmic predictions with human interpretations, offering valuable insights into the complexities of legal text analysis.To test the

robustness of the models, this study addresses the need for robust legal judgment prediction systems, highlighting their vulnerability to adversarial attacks[3]. By proposing a novel approach and conducting experiments on multiple legal datasets, significant improvements were achieved in handling such attacks, marking a noteworthy advancement in this field.

The relevance of pre-trained language models in the legal domain is underscored in recent studies such as "Pre-trained Language Models for the Legal Domain: A Case Study on Indian Law" by Paul et al. (2022), which highlights their potential to advance legal NLP. These models have proven instrumental in addressing complex tasks, as demonstrated in "Large Scale Legal Text Classification Using Transformer Models" by Shaheen et al. (2020), where the focus lies on tackling the challenging problem of large multi-label text classification using datasets like JRC-Acquis and EURLEX57K.[4]Additionally, "jurBERT: A Romanian BERT Model for Legal Judgment Prediction" by Masala et al. (2021) introduces a specialized Romanian BERT model pre-trained on a large legal corpus, illustrating the efficiency of ransformer models in NLP , particularly in the legal context.[5] presents a novel approach using LSTM models to evaluate the rationality of judicial decisions by measuring judgment deviation, based on analysis of Chinese judicial texts, aiding in efficient case handling and upholding judicial justice.

This paper [6] introduces CNN-BiGRU, a hybrid model combining CNN and BiGRU for legal judgment prediction, achieving high accuracy and efficiency, validated on the CAIL 2018 dataset, showcasing its effectiveness

in handling the growing volume of legal cases with improved prediction accuracy. This survey [7] systematically reviews recent advancements in adversarial training for enhancing the robustness of deep learning models against adversarial examples, introduces a novel taxonomy, addresses generalization issues, and identifies remaining challenges and potential future directions in the field. In this paper BERT, a bidirectional language representation model is pre-trained on unlabeled text, capable of achieving state-of-the-art performance on various natural language processing tasks with minimal task-specific modifications performing NLP tasks [10].

[8] introduces TextFooler, a robust baseline for generating adversarial text that effectively attacks state-of-the-art models in text classification and textual entailment tasks, outperforming previous attacks in success rate and perturbation rate while preserving semantic content, grammaticality, and computational efficiency .

This paper[9] reviews the landscape of Quantum Natural Language Processing (QNLP), categorizing approaches by theoretical or hardware implementation, task type, and evaluation resource, highlighting advantages and potential replacements for deep learning-based methods.[11] and [12] investigates the linguistic phenomena accounted for by language models in Conversational Question Answering tasks, identifying areas of improvement for finetuned RoBERTa, BERT, and DistilBERT models through error analysis and multitask learning, resulting in enhanced performance across various question classes.[13] and [14] explaining about transformer models explores text classification using BERT and DistilBERT models on English

and Brazilian Portuguese datasets, revealing DistilBERT's faster training time and smaller size while maintaining high language comprehension accuracy compared to BERT.

2.2 RESEARCH GAP

While significant advancements have been made in Legal NLP for tasks like legal statute identification, a critical gap remains in ensuring the robustness of these models against adversarial attacks. Existing research, such as the work by Paul et al. (2021) on legal statute identification using a heterogeneous graph-based approach, focuses on improving model performance but doesn't necessarily address adversarial robustness [1]. Similarly, studies like the one on the ILDC for CJPE dataset (Indian Legal Documents Corpus for Court Judgment Prediction and Explanation) concentrate on specific tasks within legal NLP, but the focus on adversarial robustness is often missing.

This research gap highlights the vulnerability of existing NLP models in legal applications. Malicious actors could potentially manipulate legal data, leading to inaccurate model outputs and potentially impacting judicial decisions. Our research addresses this gap by employing highly robust models (QUANTUMBERT , BERT, BiLSTM, GRU) specifically chosen for their resilience against adversarial attacks. We train these models on a curated subset of the ILSI dataset, concentrating on women-related legal sections. This focus ensures the model's applicability to a specific and significant legal domain.

Furthermore, we introduce adversarial sampling on the test data, simulating real-world scenarios where data might be corrupted. By evaluating the models' performance under such challenging conditions, we aim to surpass the accuracy of existing methods that haven't been trained with adversarial robustness in mind. In essence, our research bridges the gap by developing Legal NLP models that are not only accurate but also resilient against potential data manipulation. This focus on robustness is crucial for ensuring the reliability and fairness of NLP applications within the legal system.

CHAPTER 3

DATASET

A dataset commonly used in Indian context containing criminal case documents and statutes from the Indian judiciary is used here for the multi label classification task. This Section describes the dataset.

3.1 DATA COLLECTION AND PREPARATION

In Indian Law, most criminal offences are described in the Indian Penal Code (IPC), which is an Act. The IPC Act has a hierarchical structure – the Act is divided into coarse-grained categories called Chapters, which are further subdivided into fine-grained categories called Topics. Each Topic groups together a set of Sections that are based on the same crime. Sections are statutory legal articles that are usually cited from case documents. The text of a Section describes the nature and circumstances of the crime, and litigation procedures involved. From fig 3.1, it is demonstrated that rhe IPC contains more than 500 Sections, but a large majority of them are seldom cited. Hence, we chose to focus on the 100 most frequently cited Sections of IPC as the set of labels in our dataset. We consider the Facts from only those case documents that cite at least one of these top 100 Sections, and we end up with 66, 090 such Facts. The dataset is available at [2].

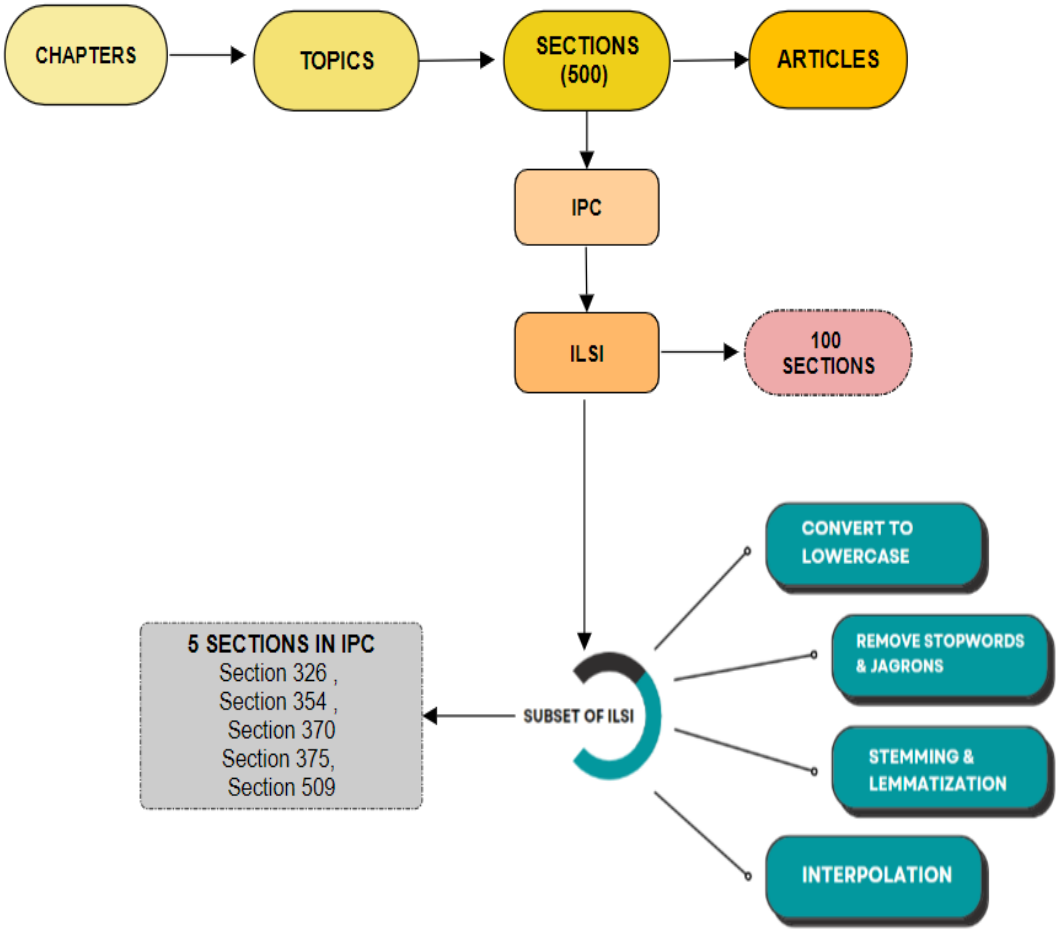

FIG 3.1 *TEXT PREPROCESSING FLOWCHART*

The dataset comprises approximately 43,000 rows, each representing a unique case filed and categorized under crime-related offenses. Each row consists of three columns: 'id', serving as a unique identifier for each case; 'text', containing the descriptions or factual details pertaining to the case; and 'labels', containing the sections within the Indian Penal Code (IPC) to which the respective cases belong. The dataset has been preprocessed and modified to facilitate the training of a robust model, aiming to accurately classify cases based on their corresponding IPC sections.

3.2 DATA PREPROCESSING

It is observed that only 5 Sections from the 100 sections of ILSI dataset is chosen to train the models. The original dataset, initially in JSON format, is set for preprocessing to enhance its ability for model training. Firstly, the text data was converted to lowercase to ensure subsequent processing steps. Subsequently, tokenization was applied to segment the text into individual words or tokens for analysis. Hyperlinks, often present in textual data, were removed to eliminate irrelevant information. Punctuation marks were stripped from the text to focus solely on the textual content. Furthermore, words and digits containing digits were excluded from the dataset to refine the dataset's quality and improve transformer models performance.

ID	TEXT	LABELS
1000008	['(a),Section 5r/w 27 of the Arms Act	['Section 395','Section 120', 'Section 5']
1000190	['05.09.13 Item No.44 Court No.17	['Section 438','Section 34', 'Section 498A']
1000196	['JUDGMENT R.KChowdry,J,'For of	['Section 120B','Section 161', 'Section 467', 'Section 109']

TABLE 1 *DATASET BEFORE PREPROCESSING*

ID	TEXT	SECTION 326	SECTION 354	SECTION 375	SECTION 376	SECTION 509
1000008	section rw of the arms act	1	0	0	0	0
1000190	item no court no	1	0	0	0	0
1000196	Judgment rk choudry for of	0	0	0	1	0

TABLE 2 *DATASET AFTER PREPROCESSING*

The differences can be seen by comparing both the tables 1 and 2 before and after preprocessing.

3.3 DATA ANALYSIS

Exploratory Data Analysis (EDA) is a crucial step in the data science workflow. It involves investigating, summarizing, and visualizing a dataset to understand its characteristics, identify patterns, and uncover potential relationships between variables.

This plot gives the outline of the various features and comparisons between them. It includes the text column and its respective id with sections under which the text filed .

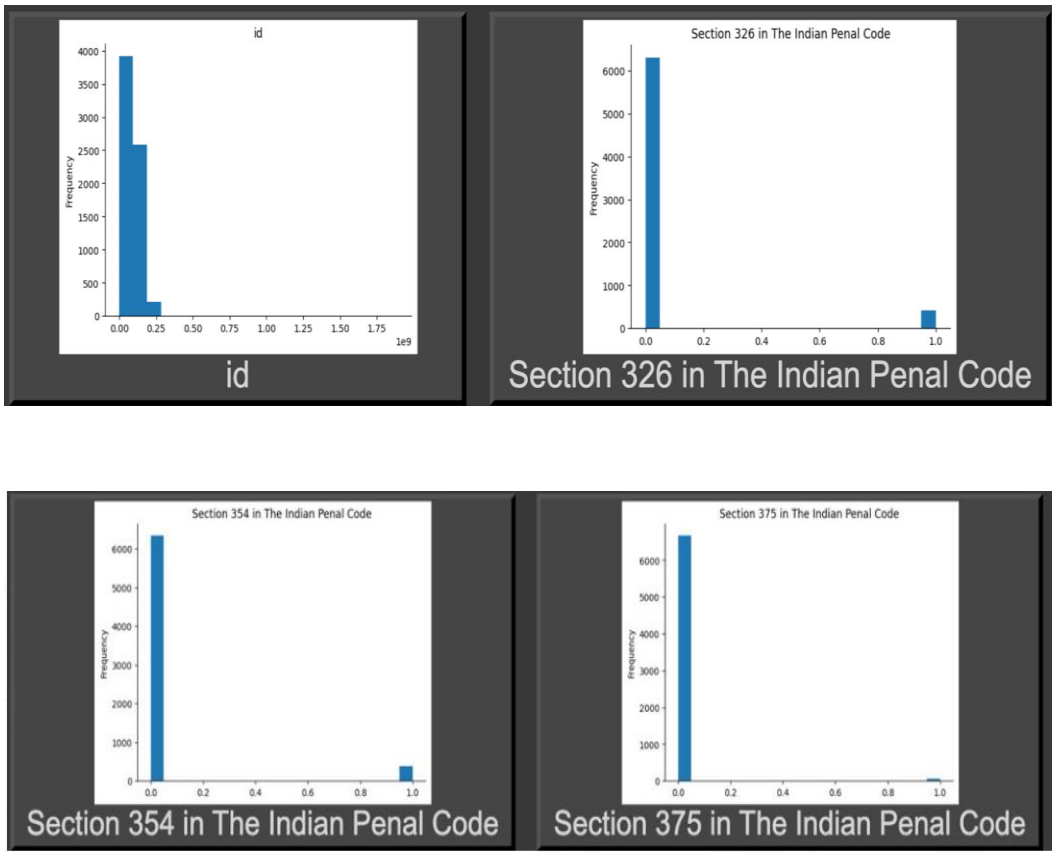

FIG 3.2 *ANALYSIS OF VARIOUS FEATURES*

This plot gives the outline of the various features and comparisons between them. It includes the text column and its respective id with sections under which the text is filed.

This plot in fig 3.2 visualizes how the number of instances in your dataset is distributed across different text lengths. The x-axis typically represents the text length (number of words or characters), while the y-axis represents the number of instances with that text length.

By analyzing the shape of the distribution, you can gain valuable insights into the structure of the data. For example, a peak at a specific text length indicates that a large portion of the data consists of texts with that length. Conversely, a flat distribution suggests that the text lengths are more evenly spread across the range.

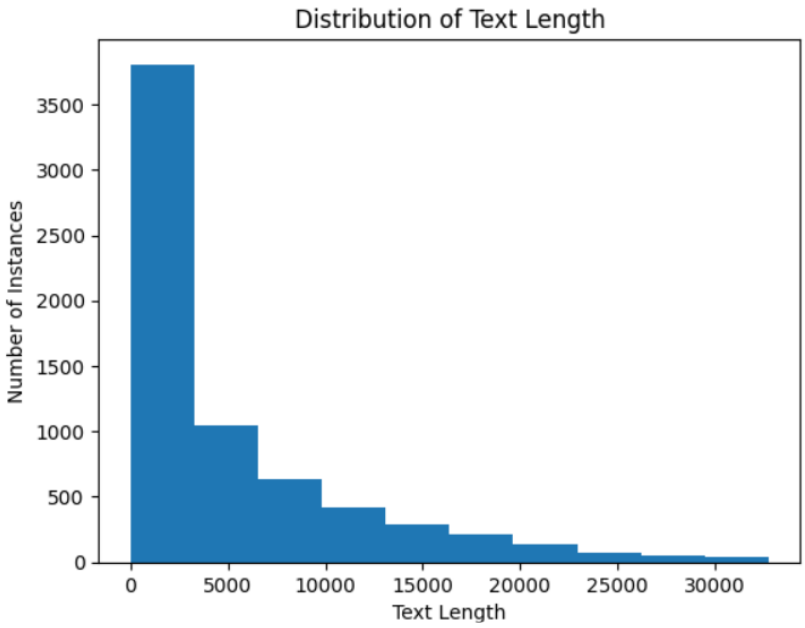

FIG 3.3 *DISTRIBUTION OF TEXT LENGTH*

This plot in fig 3.3, shows the distribution of one of the target variables in your dataset, which appears to be related to sections of the Indian

Penal Code (IPC). The x-axis typically represents the different categories or labels within the target variable (IPC sections), and the y-axis represents the number of instances (data points) belonging to each category.

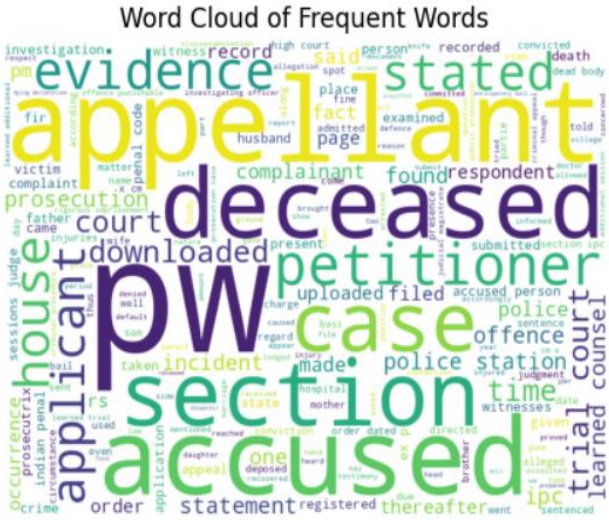

FIG 3.4 *WORD CLOUD OF FREQUENT WORDS*

Figure 3.4 depicts the cloud of frequently used words , which is used to calculate importance score in the adversarial algorithm. These highlighted words are repeated in most of the women related cases filed under the 5 sections mentioned in Figure 3.1. The dataset has been preprocessed and modified to facilitate the training of a robust model, aiming to accurately perform classification task.

CHAPTER 4

SYSTEM DESIGN

4.1 PROPOSED METHODOLOGY

This section introduces a novel approach for training BERT, BiLSTM, and GRU models. We propose implementing an adversarial training strategy on the test data to evaluate their robustness against such attacks. Figure 4.1 illustrates the system architecture for this approach.

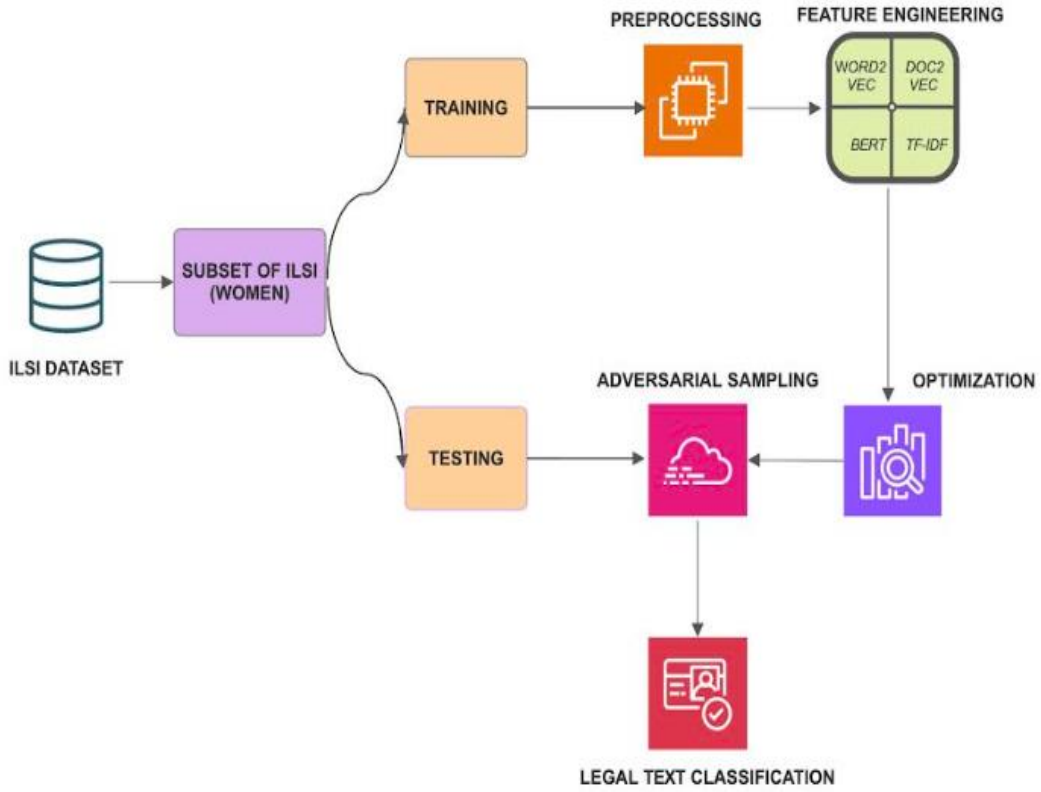

Figure 4.1 *SYSTEM ARCHITECTURE OF THE PROPOSED MODEL*

This section delves into the core of the research, introducing a novel training approach that goes beyond standard preprocessing steps. The process begins with feature extraction, a crucial step in NLP tasks where relevant information is identified and optimized from the legal text data. This optimization aims to enhance the performance of the three chosen models: QUANTUMBERT, BERT, BiLSTM, and GRU. Each model has its strengths and weaknesses in handling different types of text data, and this optimization helps them better understand the specific characteristics of legal text for the classification task.

However, the innovation lies in the subsequent step. Instead of a traditional training and evaluation process, this research employs an adversarial algorithm on the test data. This essentially simulates real-world challenges where malicious actors might attempt to manipulate legal data. By introducing these adversarial attacks, the research team can assess the models' robustness – their ability to maintain accurate classification performance even when faced with corrupted inputs.

This approach offers a significant advantage over traditional methods. By subjecting the models to adversarial attacks during evaluation, the research ensures thorough testing and validation of their capabilities. This not only provides a more realistic picture of how the models would perform in real-world applications but also helps identify potential vulnerabilities that could be exploited by malicious actors.

4.2 BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a cutting-edge language model developed by Google that has revolutionized natural language processing (NLP) tasks. BERT utilizes a transformer architecture, a neural network architecture specifically designed for processing sequential data such as text. Unlike previous models that processed text in a left-to-right or right-to-left manner, BERT is bidirectional, meaning it considers the context from both directions when understanding a word's meaning within a sentence. This bidirectional understanding allows BERT to capture the nuances and complexities of language more effectively, can be observed in FIG 4.2, leading to remarkable performance improvements in various NLP tasks such as text classification, sentiment analysis, question answering, and more. BERT's architecture referred from [16], with its attention mechanisms and multi-layered structure, enables it to capture long-range dependencies in text, contextualize words based on their surrounding context, and generate high-quality representations of language, making it a powerful tool for identification and classification tasks in NLP.

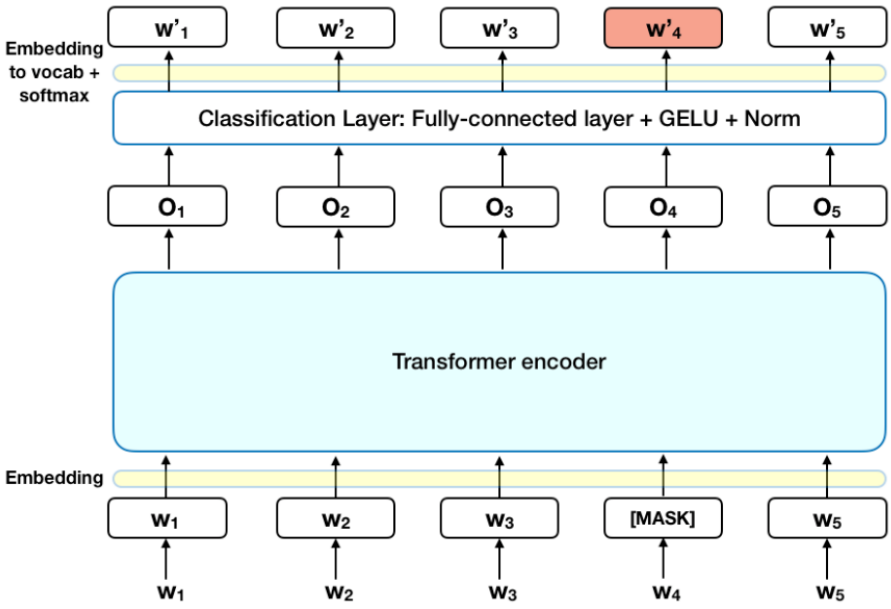

FIG 4.2 *BERT ARCHITECTURE*

Here , with our dataset , its performance with adversarial samples is distinctive approach to processing and encoding information. Its bidirectional architecture allows it to capture contextual information effectively, enabling a deeper understanding of the relationships between words and phrases within the text. It is pre-trained on a massive corpus of text data using self-supervised learning tasks, which helps it learn quickly. Fine-tuning BERT on our dataset further enhanced the performance by adapting its parameters. Moreover, BERT's attention mechanisms enable it to focus on relevant parts of the input sequence, facilitating accurate classification of multi-label data. Furthermore, BERT's multi-layered architecture allows it to learn hierarchical representations of text, capturing both low-level features and high-level semantic information.

4.3 BiDirectional LSTM

Pre-trained Bidirectional Long Short-Term Memory (BiLSTM) networks are recurrent neural networks (RNNs) that have shown significant effectiveness in various sequential data processing tasks, including natural language processing. BiLSTMs enhance traditional LSTM networks fig 4.3 , by processing input sequences in both forward and backward directions, allowing them to capture contextual information from both past and future time steps. This bidirectional processing enables BiLSTMs [17] to better understand the semantics and dependencies within sequences, making them particularly useful for tasks such as sequence labeling, sentiment analysis, and named entity recognition. By considering the entire input sequence bidirectionally, BiLSTMs can effectively capture long-range dependencies and contextual information, enabling more accurate identification and classification of patterns within sequential data.

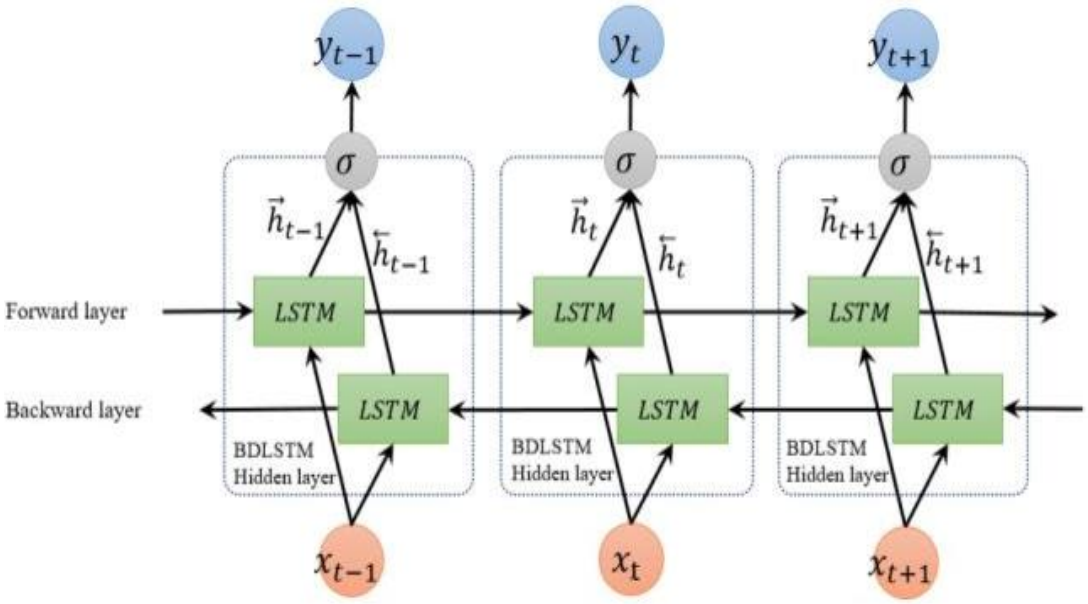

FIG 4.3 *BiDIRECTIONAL LSTM ARCHITECTURE*

This capability makes BiLSTMs a valuable tool in NLP tasks where understanding the context of words and phrases is crucial for accurate analysis and classification.

4.3.1 GATED RECURRENT UNITS(GRU)

The Gated Recurrent Unit (GRU) model is a type of recurrent neural network (RNN) architecture designed to address some of the limitations of traditional RNNs, such as the vanishing gradient problem. GRUs achieve this by introducing gating mechanisms that regulate the flow of information within the network. These gates, including the update gate and the reset gate, enable GRUs to selectively retain or discard information from previous time steps, allowing for more effective long-range dependencies

modeling. This architecture aids in identification and classification tasks by enabling GRUs to capture and remember relevant contextual information from the input sequences. The ability to selectively update memory and control the flow of information helps GRUs in learning meaningful representations of sequential data, making them particularly useful for tasks such as sentiment analysis, sequence labeling, and speech recognition. Additionally, GRUs have a simpler architecture compared to other recurrent units like LSTMs, making them computationally more efficient while still providing competitive performance in various identification and classification tasks. This is achieved from this highly resilient architecture observed in Fig 4.4[18].

This dataset containing IDs, text, and binary labels for multiple sections, applying GRU with attention mechanism entails incorporating attention weights that dynamically adjust the contribution of each hidden state of the GRU. Initially, the GRU processes the input text sequence, generating a sequence of hidden states.

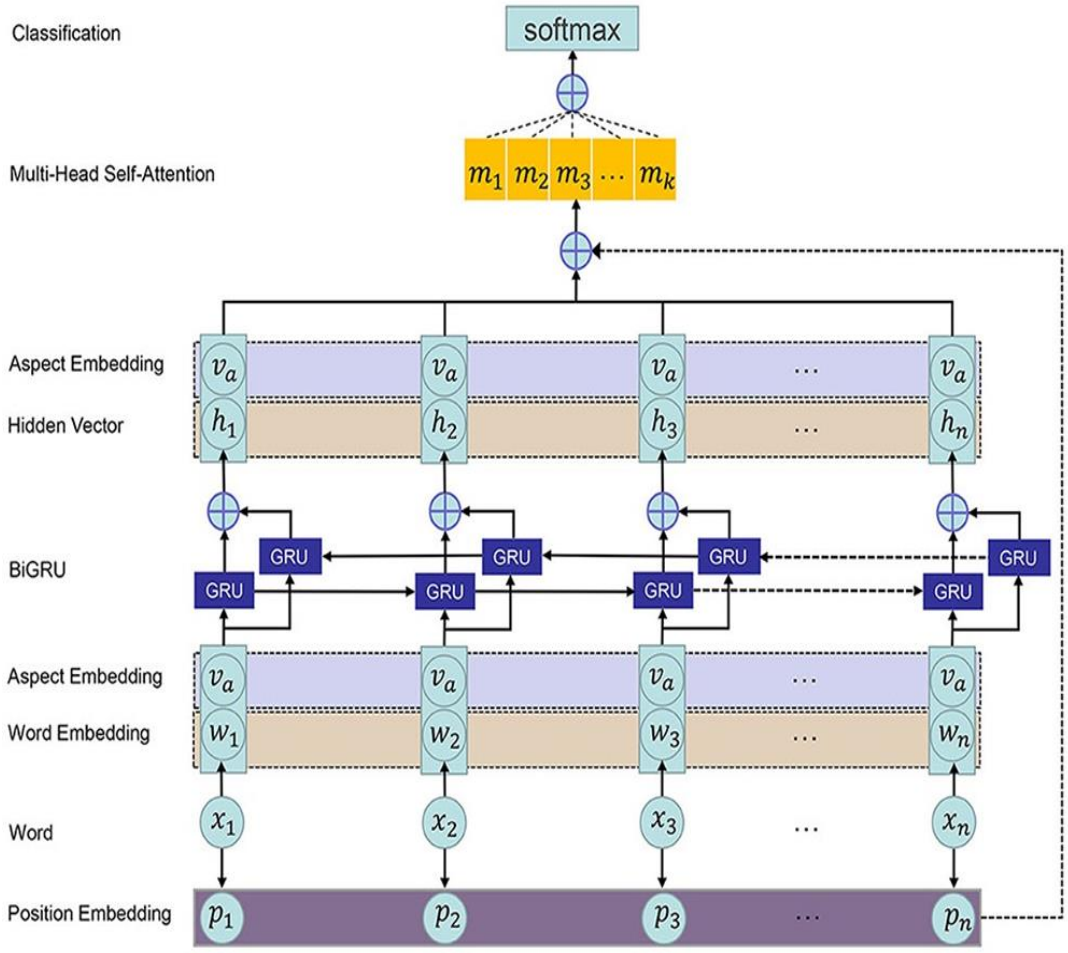

FIG 4.4 *GRU ARCHITECTURE*

Then, the attention mechanism computes attention weights for each hidden state based on its relevance to the prediction task. These attention weights are computed through a learned function that considers the current hidden state of the GRU and possibly other contextual information. Finally, the weighted sum of the hidden states, weighted by their corresponding attention weights, is used to make predictions for the labels of the sections. This mechanism enables the model to focus on the most relevant parts of the text sequence for

accurate identification of the binary labels, enhancing performance by effectively leveraging the contextual information within the data.

CHAPTER 5

EXPERIMENTAL RESULTS

5.1 TRAINING MODEL

To initially assess the performance of basic models on the dataset, the focus is narrowed down to sections related to women among the 100 sections available. Specifically, attention is given to six sections within the Indian Penal Code pertaining to women: Section 326 , Section 354 , Section 370,Section 375, Section 376 , Section 509. Out of these, only five sections are labeled in our dataset.

The performance of the BERT, LSTM, and GRU models are evaluated using these labeled sections:

MODEL	**PRECISION**	**RECALL**	**F1 SCORE**	**ACC.**
BERT on ILSI train set (epochs=10)				
Doc2Vec + LR (0.001)	0.81	0.81	0.80	0.80
Word2Vec + LR (0.001)	0.79	0.78	0.79	0.79
Bi-directional LSTM on ILSI train set (epochs=10)				
Sen2Vec+BiLSTM +att.	0.95	0.97	0.96	0.92
Doc2Vec+BilSTM tatt.	0.88	0.89	0.91	0.92
GRU on ILSI train set (epochs=10)				
TF-IDF + GRU	0.55	0.55	0.54	0.56

TABLE 3 *Training results of the proposed models*

From the run results in table 3, it is obvious that the Bidirectional LSTM model outperformed both BERT and GRU in terms of accuracy. the combinations tried 3 different types of word embeddings: Doc2Vec , WOrds2Vec and TF-IDF vectorizer with different learning rate combinatons from 0.01 to 0.0001. It is obvious that the Bidirectional LSTM model outperformed both BERT and GRU with a training accuracy of 0.90 since it performed well with attention mechanism.

This suggests that while GRU may have been more conservative in its predictions with accuracy of 0.56 , LSTM and BERT were more effective at capturing the underlying patterns in the data. The transformer architecture in BERT had its better performance by processing sequential data and capturing long-range dependencies.

The hyperparameters, including learning rate, epochs, and batch size, play crucial roles in determining the accuracy of neural network models such as BERT, LSTM, and GRU. A higher learning rate, as seen in the GRU model with a rate of 0.001, can lead to faster convergence but may risk overshooting the optimal solution, potentially affecting accuracy adversely. Conversely, a lower learning rate, exemplified by the LSTM model's rate of 0.0001, facilitates finer adjustments during training, contributing to higher accuracy. The number of epochs, with both LSTM and GRU models trained for 10 epochs, affects model performance by allowing the network to see the data multiple times, potentially enhancing accuracy, albeit risking overfitting.

Additionally, batch size influences training dynamics, with smaller sizes, like the LSTM's batch size of 32, increasing the accuracy.

The proposed methodology involves developing a Natural Language Processing (NLP) model and training it on the Indian Legal Sentences and Indications (ILSI) dataset to perform multi label classification tasks. Adversarial training will be incorporated to fortify the model against potential attacks. The major steps involve data preprocessing, including tokenization and removal of stopwords and special characters. Model implementation will explore various architectures such as LSTM, BERT, and Attention mechanisms. Evaluation metrics will primarily focus on accuracy and F1 macro score to assess model performance.

5.2 ADVERSARIAL SAMPLING

During training, the model is exposed to both the original and adversarial examples, forcing it to learn robust features that are resilient to potential attacks. By repeatedly training on a mixture of original and adversarial data, the model gradually improves its ability to correctly classify both regular and adversarial inputs.Evaluation of the adversarial training process typically involves testing the trained model on a separate dataset containing adversarial examples to assess its robustness. Metrics such as accuracy and F1 score are commonly used to measure the model's performance under adversarial conditions.

Algorithm 1- Adversarial Sampling

Require:

- Legal judgement prediction model M(·)
- Legal sample sentence X = (w1, w2, ..., wn)
- Perturbation Generator P(X, i) which replaces wi with a perturbed word using counter-fitted word-embedding

Ensure: Adversarial legal sample Xadv

1. Calculate importance score I(wi) of each word wi using equation 1.
2. Take top-k words and rank them in decreasing order according to I(wi) and store them in set R = (r1, r2, ..., rk)
3. Set X' = X
4. **for** i = 1, ..., |R|, in R **do**
5. Generate adversarial sample Xp by perturbing sentence X' using P(X', i)
6. **if** sim(Xp, X) > threshold **then**
7. **if** M(Xp) ≠ y **then**
8. Set X' = Xp
9. **end if**
10. **end if**
11. **end for**
12. **return** X' as Xadv

Even Though the performance of these models was poor initially, it increased gradually when data augmentation was executed. This task is specifically critical, since a slight variation in the input may affect judgment fairness. Hence during deployment, if the input sequence is disturbed intentionally , classification may change drastically. It is the main reason for adversarial training.Here is the sample of adversarial samples generated in Fig 5.2.

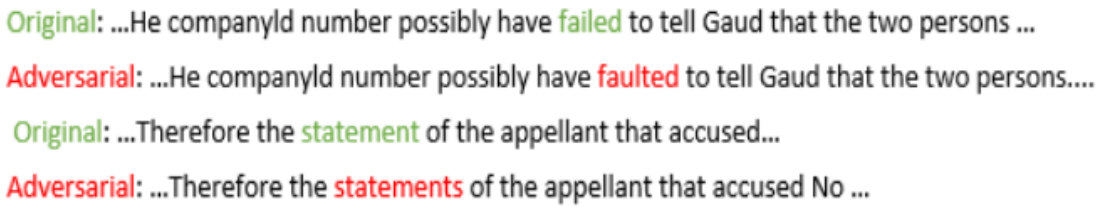

FIG 5.2 *ADVERSARIAL SAMPLE*

To generate adversarial examples, a method depicted in Fig 5.3 shows that word importance scoring is employed. First, the importance score of each word in the text is calculated using a model trained on the dataset. This importance score reflects the impact of removing a word from the original text on the model's prediction. A greedy search algorithm is then used to iteratively remove words with the highest importance scores, altering the original text to create adversarial samples.

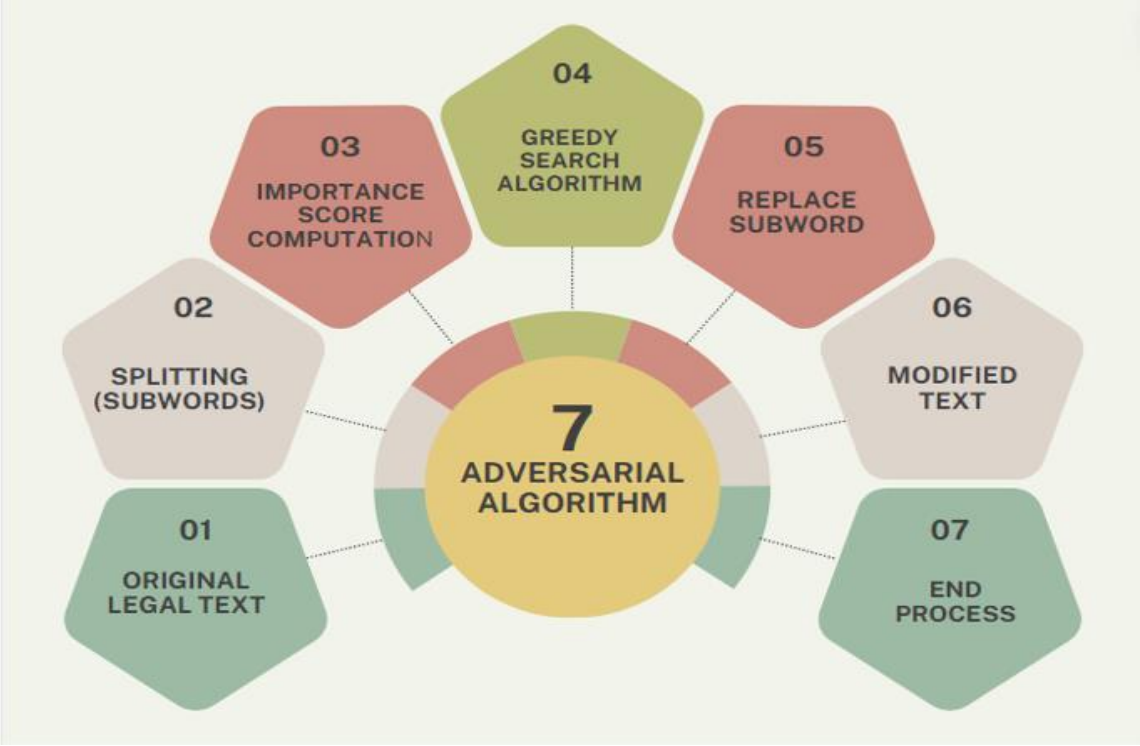

FIG 5.3 *STEPS IN ADVERSARIAL SAMPLING*

Several evaluation metrics are utilized to assess the effectiveness of the adversarial samples. These include the word importance score, which indicates the contribution of each word to the model's prediction. Additionally, textual differences between the original text and its corresponding adversarial sample are measured using standard text similarity metrics, such as cosine similarity or edit distance. This provides insights into the extent of modifications introduced by the adversarial generation process.

5.3 OUTPUT OF THE CLASSIFIER

Basic transformer models were used on this adversarial dataset, but the results were disappointing. This happened because the dataset was purposely altered to trick the models. Even small changes in the text can confuse the model and make it predict the wrong thing. Even Though , transformer models are powerful in capturing complex sequential patterns, they failed to generalize well to adversarial inputs due to their susceptibility to slight modifications in the input data. To improve the models' performance, alternative ways were implemented to make them more resistant to these adversarial attacks.

The performance of the BERT , Bidirectional LSTM and GRU model on this adversarial text was evaluated using these labeled sections. The following results was obtained after fine tuning hyperparameters while working with test dataset exposed to adversarial attack,

5.3.1 RUN RESULTS

These metrics are essential for evaluating classification models because they give us insights into different aspects of their performance: precision focuses on the accuracy of positive predictions, recall focuses on finding all positive instances, and the F1 score balances both precision and recall.

1. Precision:

Precision tells us what proportion of positive identifications was actually correct. In other words, it's the accuracy of the model when it predicts something as positive.

$$Precision = \frac{TP}{TP + TN}$$

Equation 5.1

2. Recall :

Recall tells us what proportion of actual positives was identified correctly. It's the ability of the model to find all the positive instances.

$$Recall = \frac{TP}{TP + FN}$$

Equation 5.2

3. F1 Score:

F1 Score is a balance between precision and recall. It's the harmonic mean of precision and recall. It gives us a single number that represents both precision and recall, helping us to compare models more easily.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Equation 5.3

4. Accuracy :

Accuracy is the ratio of the number of correctly classified instances to the total number of instances. It is given by the formula:

$$Accuracy = \frac{TP + TN}{TS}$$

Equation 5.4

From the run results in table 4 , it is evident that BiLSTM achieved remarkable performance with a precision, recall, and F1 score all exceeding 0.9, showcasing its effectiveness in accurately identifying women-related cases. This notable performance can be attributed to the carefully tuned hyperparameters, including a dropout probability of 0.5, a batch size of 32, and 10 epochs of training, which allowed the model to learn meaningful representations of the sequential data and effectively capture long-range dependencies.

MODEL	BERT		BILSTM		GRU	
HYPER PARAMETERS	**RECALL**	**F1 SCORE**	**RECALL**	**F1 SCORE**	**RECALL**	**F1 SCORE**
Epochs = 5 LR = 0.001 Batch size = 32	0.74	0.74	0.92	0.92	0.56	0.56
Epochs = 5 LR = 0.001 Batch size = 32	0.78	0.79	0.92	0.92	0.53	0.51
Epochs = 5 LR = 0.001 Batch size = 32	0.79	0.79	0.92	0.92	0.56	0.56

TABLE 4 *Run results: BERT, BiLSTM, GRU*

Similarly, the GRU model with attention mechanism also demonstrated robust performance, leveraging its ability to selectively update memory and control the flow of information within the network. Despite the simpler architecture compared to other recurrent units, GRU achieved competitive results, emphasizing its efficiency and effectiveness in identifying women-related cases. Further experimentation with a combination of transformer models could potentially enhance the robustness and performance of the classification task, offering promising avenues for future research in legal text analysis.

5.3.2 INFERENCES

After testing, it's clear that BERT, BiLSTM, and GRU performed well, with BiLSTM leading with an F1 score of 0.92. However, it's obvious that these models are ubiquitous in the field of natural language processing (NLP) and are extensively utilized for various tasks such as identification, summarization, and classification. Despite their effectiveness, their widespread usage often leads to a saturation point , prompting the exploration of alternative approaches. Hence, to break away from the conventional norms and introduce novel methodologies, we delve into the implementation of a relatively underexplored model known as Quantum BERT.The following section provides an overview of QUANTUMBERT , and its architecture classification approaches, and the outcomes of our research.

CHAPTER 6

QUANTUM BERT

Quantum machine learning is a research area that explores the interplay of ideas from quantum computing and machine learning . For example, we might want to find out whether quantum computers can speed up the time it takes to train or evaluate a machine learning model. On the other hand, we can leverage techniques from machine learning to help us uncover quantum error-correcting codes, estimate the properties of quantum systems, or develop new quantum algorithms.

5.1 OVERVIEW

Quantum BERT, an evolution of the renowned BERT (Bidirectional Encoder Representations from Transformers), marks a significant stride towards harnessing quantum computing's potential in natural language processing (NLP). Leveraging quantum computing's principles, Quantum BERT aims to tackle NLP tasks with unprecedented efficiency and scalability. At its core, Quantum BERT integrates the power of quantum computing, which operates on the principles of quantum mechanics, to manipulate and analyze data in quantum bits or qubits, providing a quantum advantage over classical computing methods. By encoding and processing information in qubits, Quantum BERT holds the promise of revolutionizing NLP by exponentially

enhancing computation speed and handling complex linguistic nuances with greater precision.

Furthermore, Quantum BERT stands poised to address the limitations of classical BERT models, particularly in handling vast datasets and performing intricate language understanding tasks. Its quantum-enhanced architecture enables Quantum BERT to explore a vastly larger solution space in parallel, thereby potentially unlocking novel insights and improving performance across various NLP applications. Additionally, Quantum BERT holds the potential to unlock new frontiers in machine learning and artificial intelligence by offering solutions to previously intractable problems, paving the way for groundbreaking advancements in language understanding, sentiment analysis, and information retrieval. As research and development in quantum computing continue to progress, Quantum BERT emerges as a promising avenue for driving innovation and reshaping the landscape of NLP.

6.2.1 PENNYLANE

PennyLane is a cross-platform Python library for programming quantum computers. Its differentiable programming paradigm enables the execution and training of quantum programs on various backends. PennyLane connects quantum computing with powerful machine learning frameworks like NumPy’s autograd, JAX, PyTorch, and TensorFlow, making them quantum-aware.

Its central job is to manage the execution of quantum computations, including the evaluation of circuits and the computation of their gradients. This information is forwarded to the classical framework, creating seamless quantum-classical pipelines for applications. PennyLane's design principle states that circuits can be run on various kinds of simulators or hardware devices without making any changes – the complex job of optimising communication with the devices, compiling circuits to suit the backend, and choosing the best gradient strategies is taken care of.

The library comes with default simulator devices, but is well-integrated with external software and hardware to run quantum circuits—such as IBM's Qiskit, or Google's Cirq, Rigetti's Forest, or Xanadu's Strawberry Fields.

6.3 ARCHITECURE

Quantum BERT architecture intertwines the foundational principles of BERT with the transformative capabilities of quantum computing. At its core, Quantum BERT harnesses the power of quantum mechanics to revolutionize natural language processing (NLP). While classical BERT relies on traditional neural network architectures, Quantum BERT diverges by incorporating quantum computing principles Fig 6.1, particularly quantum circuits and qubits, to process and analyze language data.

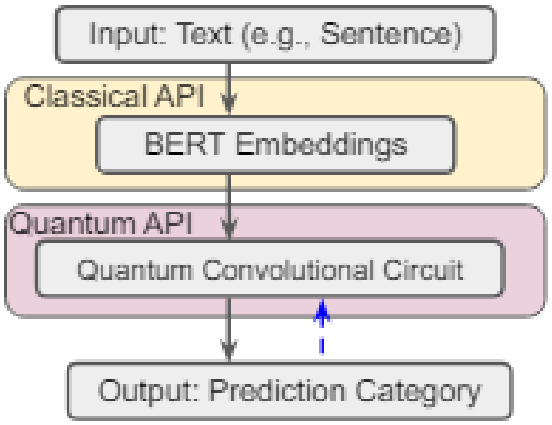

***FIG 6.1** OVERVIEW OF ARCHITECTURE*

The architecture of Quantum BERT involves several key components:

1. **Quantum Circuits**: Instead of traditional neural network layers, Quantum BERT utilizes quantum circuits to encode and manipulate language representations. These circuits leverage the principles of quantum superposition and entanglement to process information in parallel across a vast number of qubits noted as q1 and q2 in Figure 6.2 [27] , enabling Quantum BERT to explore a much larger solution space compared to classical architectures.

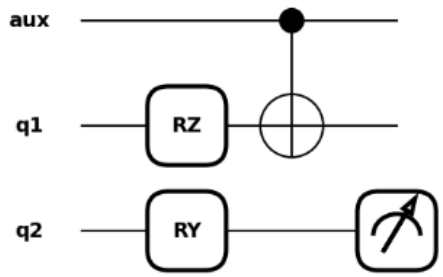

FIG 6.2 *QUANTUM CIRCUIT*

2. Qubits: Quantum BERT operates on qubits, the fundamental units of quantum information. Unlike classical bits, which can only exist in a state of 0 or 1, qubits can exist in a superposition of both states simultaneously, allowing for exponentially increased computational power and information storage capacity.

3. Quantum Gates: Quantum BERT employs quantum gates to perform operations on qubits within the quantum circuits. These gates manipulate the quantum states of qubits, enabling Quantum BERT to perform complex transformations on language representations in an efficient and scalable manner.

4. Hybrid Approach: Quantum BERT often adopts a hybrid approach, combining classical and quantum computing techniques to leverage the strengths of both paradigms. This hybrid architecture allows for the seamless integration of quantum enhancements into existing NLP frameworks, facilitating the transition from classical to quantum computing in practical applications.

From the figure 6.1 , when word embeddings are processed within quantum computational circuits, the advantage lies in the potential for exponential computational scalability and enhanced information processing. Quantum circuits exploit the principles of quantum superposition and entanglement, enabling simultaneous exploration of a vast solution space. This allows for the representation of complex linguistic features and relationships contextually rich word representations. Also, the inherent parallelism of quantum computation facilitates faster convergence and optimization, ultimately yielding more robust and accurate outputs for tasks such as natural language understanding and classification.

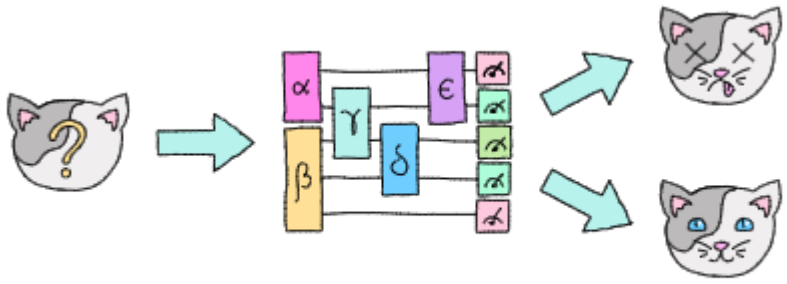

FIG 6.3 *QUANTUM BERT FOR CLASSIFICATION*

From figure 6.3 [27], It is shown that circuits, gates, deltas, and qubits together control the complex dance of information processing in quantum computing frameworks for quantum classification problems. Circuits are the blueprint that directs quantum operations, specifying the encoding, transformation, and final classification of data. Gates, the elemental building

blocks of quantum circuits, enact specific quantum operations on qubits, manipulating their quantum states to encode and process information[27]. Deltas, in this context, represent the incremental adjustments made to the parameters of the quantum gates, facilitating the iterative refinement of the classification model. Through the orchestrated interplay of circuits, gates, and deltas, quantum algorithms navigate through vast solution spaces to converge upon optimal classification boundaries, leveraging the inherent parallelism and quantum superposition to accelerate the classification process and uncover intricate patterns in the data.

This quantum classification paradigm lie qubits, the fundamental units of quantum information. Unlike classical bits, which exist in binary states of 0 or 1, qubits harness the phenomena of quantum superposition and entanglement, allowing them to exist in multiple states simultaneously. This unique property enables qubits to encode complex data representations and perform parallel computations, enhancing the capacity of quantum classification models to discern subtle patterns and relationships within the data. By harnessing the quantum properties of qubits and orchestrating their interactions through circuits and gates, quantum classification algorithms transcend the limitations of classical approaches, offering unparalleled computational power and potential for tackling high-dimensional, nonlinear classification tasks with unprecedented accuracy and efficiency.

6.4 MODEL IMPLEMENTATION AND RESULTS

The comparison between classical and quantum approaches to natural language modeling (NLM) and language processing (LP) reveals a fundamental shift in computational paradigms.It is shown in table 5. While classical methods rely on classical bits for input and processing, quantum counterparts harness the power of quantum bits (qubits) to encode and manipulate information. This transition from classical to quantum frameworks not only introduces a higher dimensionality and parallelism but also holds the promise of exponential computational scalability. With quantum LP operating on qubits, there's potential for enhanced efficiency and accuracy in classification .

APPROACH	**INPUT**	**MODEL**	**OUTPUT**
CLASSICAL	bits	NLM & LP	bits
QUANTUM	qubits	QuantumLP	qubits

TABLE 5 *An overview of different natural language processing (NLP) approaches: neural language model (NLM), logic programming (LP), and its quantum variants.*

After this interpretation of the terms in structural architecture of the models , the parameters set in here for training the model are mentioned in the table

below. Tuning these parameters causes changes in the structure of the model thus increasing the efficiency of the classification task .

QUANTUM PARAMETERS	**TRAINING**	**ACCURACY**	**TESTING**	**ACCURACY**
Step	2	0.865	2	0.851
Batch size	32		32	
Num epochs	10		10	
Rng_speed	42		42	
Qubits	4		4	
Depth	6		6	
delta	0.01		0.01	

TABLE 6 *RUN RESULTS OF QUANTUMBERT*

From the table 6 , the learning rate is set to 1e-2 to control the step size of parameter updates during optimization. Each training step processes a batch of 32 samples, determined by the batch size. Training occurs for 1 epoch, specifying the number of complete passes through the entire dataset. The random number generator is seeded with 42 to ensure reproducibility of results. The computation timer starts at the beginning of the computation. The quantum circuit consists of 4 qubits, while the depth of the circuit, i.e., the number of variational layers, is set to 6. Initial quantum weights have a spread of 0.01, denoted by q_delta, providing an initial range for weight initialization. Here , the quantum logic gate is controlled NOT gate [28] which computes the number of qubits. The CNOT gate operates on a quantum

register consisting of 2 qubits. The CNOT gate flips the second qubit (the target qubit) if and only if the first qubit (the control qubit) is |1⟩.

After training the model with this parameters setup, the model managed to achieve a training accuracy of 0.86. Along with this, adversarial sampling was introduced in test dataset to check the robustness of the quantum bert model. Subsequently , the model proved to be robust to withstand pertubations by yielding an accuracy of 0.85.

6.5 INFERENCES

Due to a number of important characteristics, the quantum BERT model outperformed classical methods with a test accuracy of 85%. First, quantum BERT may simultaneously explore a wider solution space due to the intrinsic parallelism of quantum computing, which facilitates more thorough feature extraction and representation learning. Furthermore, because of the deeper circuit (6 variational layers) and the enhanced expressiveness and capacity of quantum circuits, quantum BERT may more successfully identify complex linguistic nuances and patterns in the data. Furthermore, a varied range of initializations is guaranteed by the random quantum weights (0.01) initialization spread, which may improve training-phase convergence and optimisation. All things considered, these elements work together to enable quantum BERT to attain exceptional test accuracy, demonstrating the potential of classification task.

CHAPTER 7

PERFORMANCE ANALYSIS

HEAT MAP

A heatmap is a graphical representation of a matrix using color variations. It visualizes the magnitude of values in a two-dimensional data structure, where the intensity of a color typically reflects a higher value. Heatmaps are useful for exploring relationships between variables in a dataset.

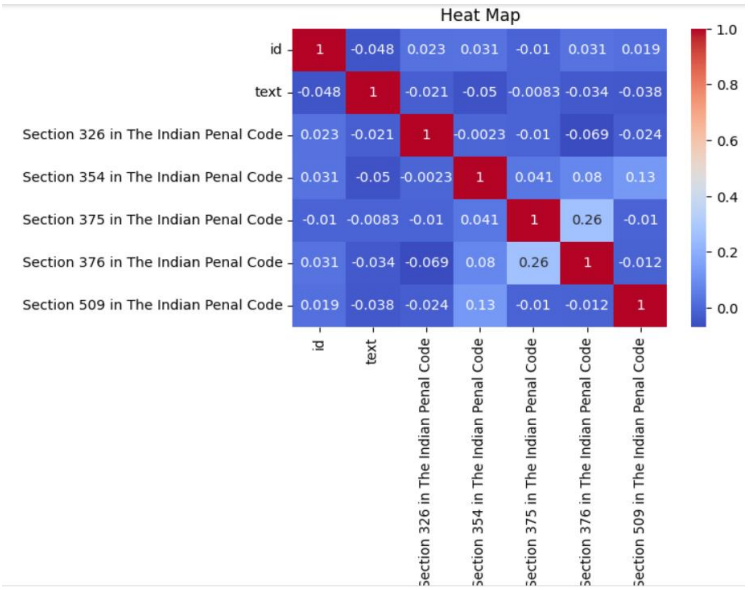

FIG 7.1 *HEAT MAP*

Here in our result case in figure 7.1 , The heatmap depicts a confusion matrix for BiLSTM multi-label classification task. The columns represent the actual

labels (IPC sections), while the rows represent the predicted labels by your model. Each cell shows the number of instances (data points) where a specific prediction was made. The color intensity in each cell corresponds to the number of instances.

Darker colors: Indicate a higher number of instances. Ideally, these darker squares should be along the diagonal, representing correct classifications for each IPC section.

Lighter colors: Indicate a lower number of instances. Ideally, these should be present off-diagonally, signifying occasional misclassifications. Specific Observations (depending on the image).

Dominant Diagonal: A prominent dark diagonal suggests good overall performance, as most instances are classified correctly according to their actual IPC sections.

Off-Diagonal Elements: Lighter colored squares away from the diagonal represent misclassifications. These squares are analysed to identify specific IPC sections where the model struggles. Here, a dark square at the intersection of "Predicted Section 354" and "Actual Section 376" indicates that some instances belonging to Section 376 were misclassified as Section 354.

CLASSIFICATION REPORT

Based on the classification report, it's evident that BiLSTM achieved the highest F1 score and accuracy among the models, indicating its superior

performance in accurately classifying legal text cases within the sections of the Indian Penal Code. Furthermore, the cross-validation process reinforces the robustness of the classification results, affirming the reliability of BiLSTM's performance in this task.

	Precision	**Recall**	**F1-score**	**Support**
0	0.95	0.97	0.96	2094
1	0.15	0.08	0.10	124
Accuracy			0.97	2218
Macro average	0.55	0.53	0.53	2218
Weighted average	0.90	0.92	0.01	2218

TABLE 7 *Classification report*

Table 7 indicates , that the macro-average F1 score (0.55) indicates varying performance across classes, while the higher weighted average F1 score (0.90) reflects overall better performance, accounting for class distribution.

CHAPTER 8

CONCLUSION AND FUTURE WORKS

In this work, the performance of the Quantum BERT model, achieving an accuracy of 0.85, inspite of introducing adversarial sampling during testing phase , is undeniably remarkable. This exceptional performance can be attributed to the increased capacity and expressiveness of quantum circuits, allowing for more nuanced understanding of textual data. Unlike traditional models, the Quantum BERT architecture offers a unique advantage, as tuning parameters can significantly alter the model's structure, increasing circuit's efficiency leading to improved performance. Given its impressive results and its robustness against adversarial sampling, the implementation of Quantum BERT for various other tasks like judgment prediction and summarization is highly beneficial.

Also , it is demonstrated that BiLSTM performed well with a best accuracy of 0.9 which can withstand any adversarial attack whereas pre-existing models depicted poor performance against adversarial attacks. Classification task is successfully implemented with the better performance of the transformer models using attention mechanisms.

This research also explores the successful implementation of classification tasks using Transformer models, which leverage attention mechanisms. These models potentially offer strong performance. Looking ahead, the authors propose testing various tasks, including judgment prediction, identification, and summarization, with adversarial sampling. This would be crucial for

evaluating the robustness of different models and their ability to maintain performance when faced with deliberately crafted inputs designed to deceive them.

REFERENCES

[1] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. arXiv preprint arXiv:2102.01356,2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert:Pre-training of deep bidirectional transformers for language un- derstanding. arXiv preprint arXiv:1810.04805, 2018.

[3] Siddhant Garg and Goutham Ramakrishnan.Bae:Bert-based adversarial

examples for text classification. arXiv preprint arXiv:2004.01970, 2020.

[4] Raffaele Guarasci, Giuseppe De Pietro, and Massimo Esposito. Quantum natural language processing:Challenges and opportunities. Applied sciences,12(11):5651, 2022.

[5] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is Bert really robust?a strong baseline for natural language attack on text classification and entailment.In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 8018–8025, 2020.

[6] Shang Li, Hongli Zhang, Lin Ye, Xiaoding Guo, and Binxing Fang. Evaluating the rationality of judicial decision with lstm-based case modeling. In 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), pages 392–397, 2018.

[7] Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripa Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. arXiv preprint arXiv:2105.13562, 2021.

[8] Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. Lesicin: a hetero- geneous graph-based approach for automatic legal statute identification from indian legal documents. In Proceedings of the AAAI conference on artificial intelligence, volume 36, pages 11139–11146, 2022.

[9] Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. Pre- trained language models for the legal domain: a case study on Indian law. In Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law,pages 187–196, 2023.

[10] Felipe Maia Polo, Gabriel Caiaffa Floriano Mendonc a, Kauˆe Capellato J Parreira,Lucka Gianvechio,Peterson Cordeiro,Jonathan Batista Fer- reira,Leticia Maria Paz de Lima, Antˆonio Carlos do

Amaral Maia,and Renato Vicente.LegalNlp–natural language processing methods for the brazilian legal language. arXiv preprint arXiv:2110.15709, 2021.

[11] Zein Shaheen, Gerhard Wohlgenannt, and Erwin Filtz. Large scale legal text classification using transformer models. arXiv preprint arXiv:2010.12871, 2020.

[12] Rafael Silva Barbon and Ademar Takeo Akabane. Towards transfer learning techniques—bert, distilbert, bertimbau, and distilbertimbau for automatic text classification from different languages: a case study. Sensors, 22(21):8184, 2022.

[13] Ieva Stali unait e and Ignacio Iacobacci. Compositional and lexical semantics inroberta,bert and distilbert:A case study on coqa. arXiv preprint arXiv:2009.08257, 2020.

[14] Chenlu Wang and Xiaoning Jin. Study on prediction of legal judgments based on the cnn-bigru model. In Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence, pages 63–68, 2020.

[15] Sabine Wehnert, Viju Sudhi, Shipra Dureja, Libin Kutty, Saijal Shahania, and Ernesto W. De Luca. Legal norm retrieval with variations of the bert model combined with tf-idf vectorization. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, ICAIL '21, page 285–294, New York, NY, USA, 2021. Association for Computing Machinery.

[16] Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Yu Tsao, and Pin-Yu Chen.When bert meets quantum temporal convolution learning for text classification in heterogeneous Computing In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),pages 8602–8606. IEEE, 2022

[17] https://www.gabormelli.com/RKB/Bidirectional_LSTM_%28BiLSTM%29_

[18] https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for

[19] Chenlu Wang and Xiaoning Jin. Study on prediction of legal judgments basedon the cnn-bigru model. In Proceedings of the 2020

6th International Conference on Computing and Artificial Intelligence, pages 63–68, 2020.

[20] Sabine Wehnert, Viju Sudhi, Shipra Dureja, Libin Kutty, Saijal Shahania,and Ernesto W. De Luca. Legal norm retrieval with variations of the bert model combined with tf-idf vectorization.In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law ICAIL '21, page 285–294, New York, NY, USA, 2021. Association for Computing Machinery.

[21] Chao-Han Huck Yang, Jun Qi, Samuel Yen-Chi Chen, Yu Tsao, and Pin-Yu Chen. When bert meets quantum temporal convolution learning for text classification in heterogeneous computing. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8602–8606. IEEE, 2022.

[22] Shudong Yang, Xueying Yu, and Ying Zhou. Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In 2020 International workshop on electronic communication and artificial intelligence (IWECAI), pages 98–101. IEEE, 2020.

[23] Jin Yong Yoo and Yanjun Qi.Towards improving adversarial training of nlp models. arXiv preprint arXiv:2109.00544, 2021.

[24] Muhammad Zulqarnain, Rozaida Ghazali, Muhammad Ghulam Ghouse,and Muhammad Faheem Mushtaq. Efficient processing of gru based on word embedding for text classification. JOIV: International Journal on Informatics Visualization, 3(4):377–383, 2019.

[25] Rafael Silva Barbon and Ademar Takeo Akabane. Towards transfer learning techniques—bert, distilbert, bertimbau, and distilbertimbau for automatic text classification from different languages: a case study. Sensors , 22(21):8184, 2022.

[26] Ieva Stali ̄unait e and Ignacio Iacobacci. Compositional and lexical semantics in roberta, bert and distilbert: A case study on coqa. arXiv preprint arXiv:2009.08257, 2020.

[27] https://docs.pennylane.ai/en/stable/introduction/circuits.html

[28] https://en.m.wikipedia.org/wiki/Controlled_NOT_gate

[29] LexisNexis. (n.d.). LexisNexis Legal Research & Analytics. Retrieved from https://www.lexisnexis.com/

[30] Luminance. (n.d.). Legal Contract Analysis & Review Software - Luminance. Retrieved from https://www.luminance.com/

[31] National Judicial Data Grid. (n.d.). About NJDG. Retrieved from https://njdg.ecourts.gov.in/njdg_public/

[32] AustLII. (n.d.). Australasian Legal Information Institute. Retrieved from https://www.austlii.edu.au/

[33] Lippi, M., & Torroni, P. (2016). Argumentation Mining: State of the Art and Emerging Trends. ACM Transactions on Internet Technology (TOIT), 16(2), 1-26.

[34] Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective. PeerJ Computer Science, 2, e93.

[35] Martin, A. D., & Katz, D. M. (2019). Machine Learning and Law: A Critical Overview. Annual Review of Law and Social Science, 15, 297-316.

[36] Winkels, R., Hepp, T., & Breuning, M. (2019). The Digital Lawyer: Legal Technology and Legal Design. Artificial Intelligence and Law, 27(3), 261-290.van Opijnen, M., van Noort, G., & van den Berg, M. (2020). Machine Learning in Law Firms: Understanding the Role of LegalTech and its Impact on the Legal Profession. Journal of Open Innovation: Technology, Market, and Complexity, 6(2), 43.