Empowering Indian Legal NLP: Adversarial Sampling and Multi-label Classification with Large Language Models

Dr.S.Karthika
Information Technology
Sri Sivasubramaniya Nadar
College of Engineering
skarthika@ssn.edu.in

Shanmukha Naveen
Information Technology
Sri Sivasubramaniya Nadar
College of Engineering
shanmukhanaveen2010809@ssn.edu.in

Sharon Roshini
Information Technology
Sri Sivasubramaniya Nadar
College of Engineering
sharonroshini2010942@ssn.edu.in

***Abstract*—Legal NLP, a field within natural language processing (NLP), involves techniques to understand and analyze legal texts that are filed under statutes. NLP techniques are widely applied for tasks such as identification, prediction, and summarization within legal documents. This paper focuses on multi-label classification using three models (BERT, BiLSTM, GRU) with attention mechanisms trained on the subset of Indian Legal Statute Identification (ILSI) dataset, comprising 100 sections, particularly working with women-related sections. These models are tested against adversarial sampling to assess their robustness. Despite the challenges posed by adversarial inputs, BiLSTM demonstrates exceptional performance with test accuracy of 0.92, followed by BERT and GRU, showcasing their resilience in legal text analysis even under adversarial conditions.**

***Index Terms*—LSI, ILSI, IPC, BiLSTM ,GRU**

I. Introduction

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 judgment 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. Essentially, 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 classification 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.

The remainder of this paper is structured as follows. Section 2 provides the various supporting works for the developed system. Section 3 illustrates the dataset and text preprocessing flow and Section 4 explains about the novel architecture used to assess the performance. In Section 5 the proposed methodology is explained and in the Section 6 experimental results and performance analysis of the models are discussed and in the section 7 the research work is concluded with the future work.

II. LITERATURE SURVEY

A. LeSICiN model for statute identification

In their recent study, ”A Heterogeneous Graph-based Approach for Automatic Legal Statute Identification from Indian Legal Documents,” Paul et al. (2021) present an innovative method that integrates textual analysis with legal citation networks for Legal Statute Identification (LSI) [13]. 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.

B. CJPE model for judgment prediction

Another survey, introduces the ILDC dataset, titled ”ILDC for CJPE: Indian Legal Documents Corpus for Court Judg- ment Prediction and Explanation,” consisting of 35,000 Indian Supreme Court cases annotated with original decisions. [10]Their study centers on Court Judgment Prediction and Explanation (CJPE), where the authors explore baseline models and propose a hierarchical occlusion-based model for enhancing explainability.

C. Adversarial training on transformer models

To test the robustness of the models, this study addresses the need for robust legal judgment prediction systems, highlighting their vulnerability to adversarial attacks. [17]. 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.

D. Other related works

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. [16] 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. [20]. [7] 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 [1] 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. [12] 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 [3].

[5] 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 .

Adversarial assaults have been used extensively in research to analyse NLP models; [8] however, each attack is implemented in a separate code repository. Creating NLP attacks and applying them to enhance model performance is still difficult. This work [9]presents TextAttack, a Python framework for data augmentation, adversarial training, and adversarial attacks in natural language processing. Because of TextAttack’s modular architecture, [23]researchers may quickly assemble attacks by combining both new and pre-existing components.

[2] explains the working of BiDirectional LSTM on sentiment analysis using text preprocessing which gives better understanding of the models working. Also [24]and [22] reviews attention mechanism working on text data from which prediction and summarization of customised dataset are capa- ble of achieving better accuracy. This paper [4] 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. [14] and [18]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. [19] and [21] 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.

III. DATASET

A. Description

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.

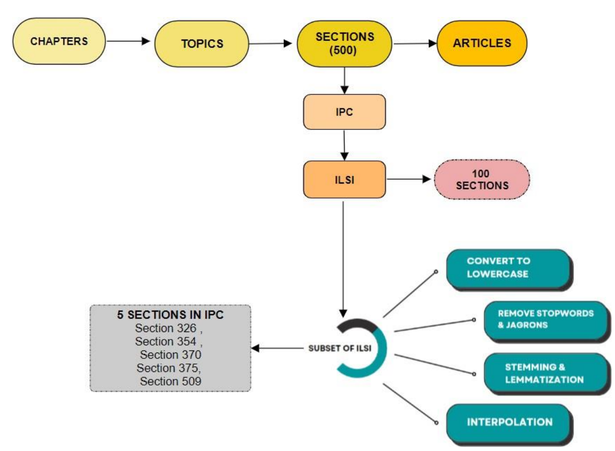

Fig. 1. PREPROCESSING

From the figure 1, it is inferred that 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. The 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 [13].

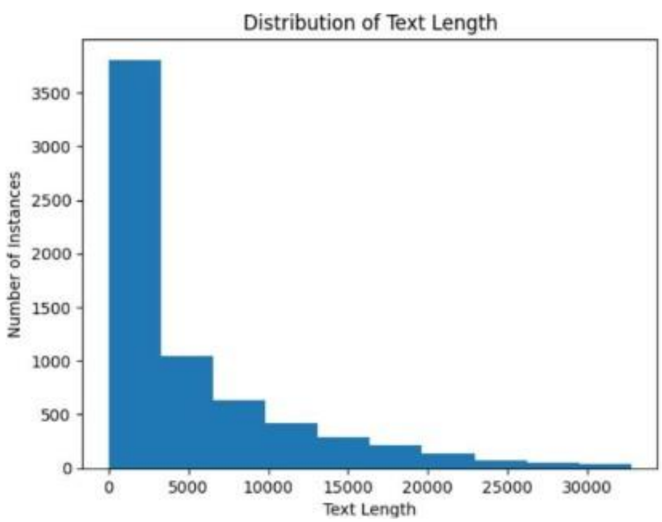

Fig. 2. Word Cloud

Figure 2 depicts the plot of legal text cases in the 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 cases filed. A peak at a specific text length indicates that a large portion of data consists of lengthy text cases.

Figure 3 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 1. The dataset has been preprocessed and modified to facilitate the training of a robust model, aiming to accurately perform classification task. The following subsection gives the format of preprocessing which is previously depicted in Figure 1.

B. Data preprocessing

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

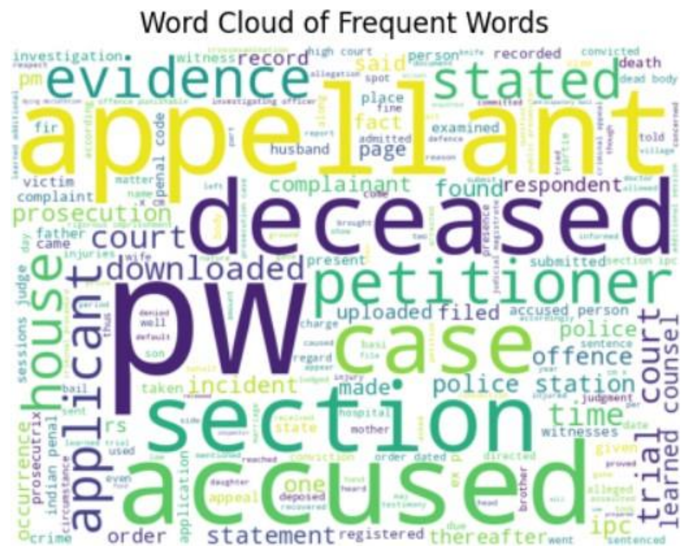

Fig. 3. Word Cloud

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.

IV. SYSTEM ARCHITECTURE

In this work, Figure 4, a novel architecture plays a crucial role, particularly in the training phase following preprocessing. Feature extraction is conducted and optimized to enhance the performance of three models: BERT, BiLSTM, and GRU. Subsequently, an adversarial algorithm is implemented on the test data to evaluate the classification performance and assess the models' robustness in withstanding adversarial attacks. This approach ensures thorough testing and validation of the models' capabilities, especially in challenging real-world scenarios where adversarial inputs may occur.

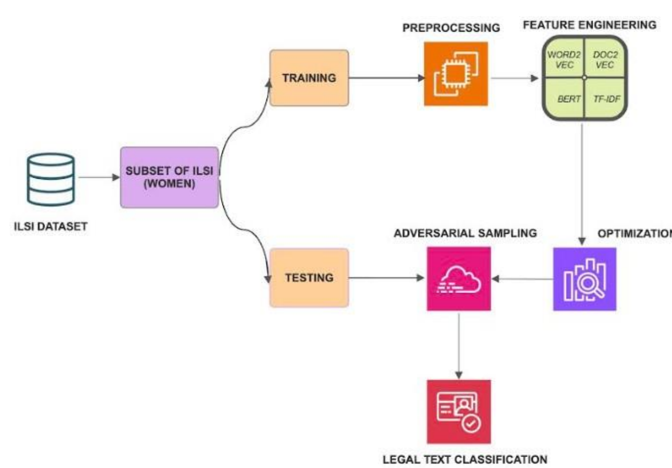

Fig. 4. System Architecture of the proposed model

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 classsification tasks. Adversarial training will be incorporated to fortify the model against potential pertubations. The major steps involve data preprocessing, including tokenization and removal of stopwords and special characters. Model implementation will explore various architectures such as BiLSTM, BERT, and GRU Attention mechanisms. Evaluation metrics will primarily focus on accuracy and F1 macro score to assess model performance.

The success of this architecture is highlighted by its excellent performance in both training and testing, particularly demonstrated through adversarial sampling of the dataset as outlined in the experimental and performance analyses.

V. PROPOSED WORK

To initially assess the performance of basic models on the dataset, the focus is narrowed down to sections relatedto 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.

A. Adversarial Sampling

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** 1 RESULTS AFTER TRAINING*

From the run results of training in table 1 , 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. Despite BERT showing slightly lower accuracy, it experienced a lower test loss compared to the other models. This suggests that while GRU may have been more conservative in its predictions, 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 BiLSTM model's rate of 0.0001, facilitates finer adjustments during training, contributing to higher accuracy. The number of epochs, with both BiLSTM and GRU models trained for 10 epochs, affects model performance by allowing the networkto 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 16, increasing the accuracy.

B. Adversarial Sampling

Fig. 5. Steps in adversarial sampling

Figure 5 explains a technique used in machine learning to enhance the robustness of models against adversarial sampling [11]. In the context of natural language processing (NLP), adversarial sampling involves generating confused or fake examples, known as adversarial examples, from original data and incorporating them into the training process. The goal is to expose the model to these perturbations during testing so that it learns to analyse it and make more accurate predictions on unseen or adversarially crafted inputs.The process typically begins by creating adversarial examples from the original dataset. This can involve various methods such as adding small, carefully crafted perturbations to the input data or modi- fying certain features to mislead the model's predictions whilestill retaining semantic meaning. These adversarial examples are then combined with the original data to form an augmented dataset.

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

During deployment, if the input sequence is disturbed intentionally , classification may change drastically. It is the main reason for adversarial training. Figure 6 shows the sample of adversarial samples generated:

Original: ...He companyld number possibly have failed to tell Gaud that the two persons ...
Adversarial: ...He companyld number possibly have faulted to tell Gaud that the two persons....
Original: ...Therefore the statement of the appellant that accused...
Adversarial: ...Therefore the statements of the appellant that accused No ...

Fig. 6. Adversarial sample

C. Implementation

To generate adversarial examples, a method based on 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. 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.

Hence, this algorithm was implemented to generate the test dataset new adversarial dataset, and the changes that occurred have been noted down. Now, the working plan is to train this dataset on a robust model capable of withstanding adversarial attacks.

VI. EXPERIMENTS AND RESULTS

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 results in misclassification. Since , transformer models are powerful in capturing complex sequential patterns, it works to generalize well to adversarial inputs inspite of their susceptibility to slight modifications in the input data. To improve the models' performance, hyperparameters tuning was executed with various combinations as mentioned in table 3.

The performance of the BERT , Bidirectional LSTM and GRU model on this adversarial text was evaluated using these

Algorithm 1 Adversarial Sampling

Require: Legal judgement prediction model $M(\vartheta)$, legal sample sentence $X = (w_1, w_2, ..w_n)$, Perturbation Generator $P(X, i)$ which replaces w_i with a perturbed word using counter-fitted-word-embedding

Ensure: Adversarial legal sample X_{adv}

Calculate importance score $I(w_i)$ of each word w_i using equation 1.
Take top-k words and rank them in decreasing order according to $I(w_i)$ and store them in set $R = (r_1, r_2, ..r_k)$
$X' \leftarrow X$
for $i = r_1, ..r_n$ in R **do**
    $X_p \leftarrow$perturb the sentence X' using $P(X', i)$
    if $M(X_p) = y$ **then**
        if sim(X_p, X) > threshold **then** {Check similarity of X and X'}
            $X' \leftarrow X_p$
        end if
    end if
end for
return X' as X_{adv} =0

labeled sections. The following results was obtained after fine tuning hyperparameters while working with test dataset exposed to adversarial sampling:

A. Adversarial Sampling

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 2 *TEST RUN RESULTS*

The heatmap depicts a confusion matrix for BiLSTM multilabel classification task. The columns represent the actual labels (IPC sections), while the rows represent the predicted labels of the model. It is inferred that the prominent dark diagonal suggests good overall performance, as most instances are classified correctly according to their actual IPC sections while lighter colored squares away from the diagonal represent misclassifications.

From the run results in table 2, 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 hyperpa-

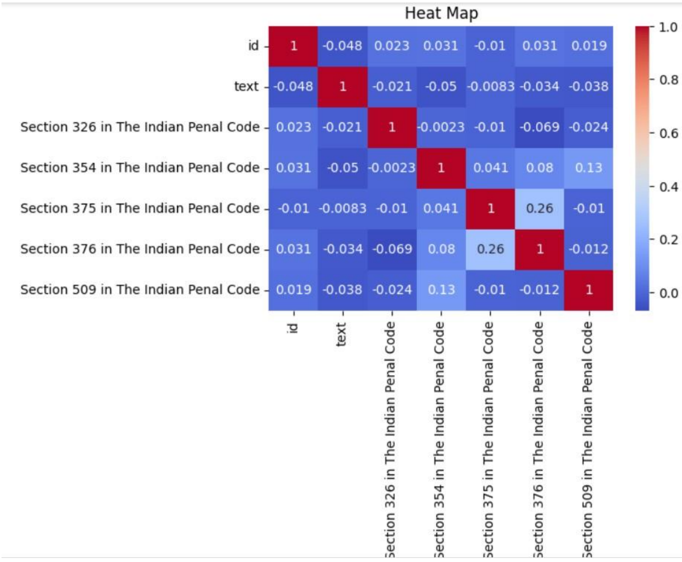

Fig. 7. Heat map of the model

rameters, 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.

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, with batch size of 32 emphasizing its efficiency over 0.56 in classifying 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.

Evaluation metrics: To assess the performance of the models, these metrics include precision, recall, and F1 score, where precision measures the accuracy of positive predictions, recall assesses the model's ability to correctly identify all relevant instances, and F1 score provides a harmonic mean of precision and recall. Hence , F1 Macro score plays a vital role in classification tasks, equation is given by the formula ,

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The classification report of the best performing model , BiLSTM is depicted as follows:

	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 3 *CLASSIFICATION REPORT*

With the best F1 score and accuracy , it is very obvious that BiLSTM has classified the legal text cases more accurately in

With the best F1 score and accuracy , it is very obvious that BiLSTM has classified the legal text cases more accurately inits respective sections in the Indian Penal Code. It is also crossvalidated to check the classification.

VII. CONCLUSION AND FUTURE WORKS

In this work , it is demonstrated that BiLSTM performed well with best accuracy of 0.9 which can withstand any adversarial pertubations whereas pre-existing models depicted less performance against adversarial attacks [9]. Classification task is successfully implemented with the better performance of the transformer models using attention mechanism. In future, many other tasks including judgment prediction, identification and summarization has to be tested with adversarial attacks so as to check the robustness of the model and their performance. Furthermore, there has been less research done on Quantum Bert in machine learning. [5] QuantumBERT, which contains built-in circuits, gates, and variational layers, wouldbe far more helpful in the classification problem than the conventional transformer models. Therefore, implementing an adversarial algorithm on this model , in the future, has better chances of achieving classification accuracy even further and produce better outcomes.

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] Chaohui Chai and Dongru Ruan. A sentiment classification algorithm of bi-lstm model fused with weighted word vectors. In *2021 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*, pages 249–253. IEEE, 2021.

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

[4] Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification. *arXiv preprint arXiv:2004.01970*, 2020.

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

[6] 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.

[7] 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.

[8] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 115–124, 2017.

[9] Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*, 2020.

[10] 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.

[11] John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin,and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*, 2020.

[12] Jinseok Nam, Jungi Kim, Eneldo Loza Menc´ıa, Iryna Gurevych, and Johannes Fürnkranz. Large-scale multi-label text classification—revisiting neural networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14*, pages 437–452. Springer, 2014.

[13] Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. Lesicin: a heterogeneous 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.

[14] Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. Pretrained 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.

[15] Felipe Maia Polo, Gabriel Caiaffa Floriano Mendonça, Kauê Capellato J Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Batista Ferreira, Leticia Maria Paz de Lima, Antônio Carlos do Amaral Maia, and Renato Vicente. Legalnlp–natural language processing methods for the brazilian legal language. *arXiv preprint arXiv:2110.15709*, 2021.

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

[17] 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.

[18] Ieva Staliūnaitė and Ignacio Iacobacci. Compositional and lexical semantics in roberta, bert and distilbert: A case study on coqa. *arXiv preprint arXiv:2009.08257*, 2020.

[19] 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.

[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.