Evaluating Transformer Models for Legal Judgement Prediction: A Comparative Study

2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 979-8-3503-7024-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCCNT61001.2024.10725043

Trisha Ghosh
Computer Science and Engineering
Delhi Technological University
Delhi, India
trishaghosh2950@gmail.com

Shailender Kumar
Computer Science and Engineering
Delhi Technological University
Delhi, India
shailenderkumar.verma@gmail.com

***Abstract*— Current developments in natural language processing (NLP) show that there is great potential in the area of law. This study investigates the performance of six transformer-based models BERT, XLNet, RoBERTa, DeBERTa, ELECTRA, and BigBird in predicting legal decisions using ILDC (Indian Legal Documents Corpus)-single dataset. Given the complexity of Indian judiciary, our study examines how recent advances in transformer technology can enhance classification accuracy. Our preliminary findings indicate that the best performance from the newer models (DeBERTa, ELECTRA, BigBird) reaches 80% accuracy, while the old models peaked at 76% accuracy, which is almost a 4% increment. This paper provides a detailed comparative analysis of each model, focusing on their respective merits and potential for automating legal judgements. In addition, we address the shortcomings of our work and make recommendations for further research.**

Keywords— Natural Language Processing, Transformer models, Court Judgement Prediction, Legal AI applications, BERT, XLNet, ROBERTa, DeBERTa, ELECTRA, BigBird, Indian Law.

I. INTRODUCTION

Lately, Natural Language Processing (NLP) has significantly evolved, with a notable shift towards Transformer-based models [1]. These models, particularly effective in specialized domains like legal documents, leverage the method of attention to interpret each lawful token's significance in relation to the document as a whole. India's legal system, inherited from the British common law, faces additional challenges due to the country's vast population and diverse legal framework. The judicial system often suffers due to the unstructured nature of legal texts. Court judgment prediction, a critical application of NLP in the legal field, involves analyzing legal documents to predict court decisions (ruled in favour of petitioner, ruled against petitioner). The advent of BERT [2] and other Transformer models has introduced tools which have the capability to comprehend the subtleties and complexity of legal vocabulary.

This paper explores the comparative effectiveness of six Transformer models—BERT [2], XLNet [3], and RoBERTa [4] as used in previous works with newer models- DeBERTa [5], ELECTRA [6], and BigBird [7] on the ILDC-single dataset for court judgment prediction [8]. This study is structures into five sections. Section 2 is mentions related work. Section 3 provides a detailed methodology of our study. Section 4 outlines the study's outcomes. Lastly, Section 5 is the conclusion, mentioning limitations and future work.

II. RELATED WORK

With the advent of Transformer-based models [1], Legal NLP has thrived in automating complex text analysis, specifically in the Indian jurisdiction. This section focuses on the relevant literature, reviewing the application of NLP in the legal sector with respect to the Indian context as well as previous works done with the dataset used in our work.

A. NLP in Indian Legal Domain

Quite a few notable works have been done in Legal NLP when it comes to the Indian context. These include tasks of catchphrase identification in legal texts [9] using unsupervised methods, and enhancing text similarity analysis using topic modeling and neural networks [10] and Hier-SPCNet, which merges precedent citations with legal statutes(written laws) [11]. Other tasks include comparison of various summarization algorithms [12], rhetorical role labeling using deep learning models such as BiLSTM and BiLSTM-CRF [13], which was later extended to check for its effectiveness across different jurisdictions [14], introduction of LeSICIN which integrates textual and citation data into a heterogeneous graph for legal statute identification [15].

B. Court Judgement Prediction on ILDC Dataset

The Indian Legal Documents Corpus Dataset (ILDC) was introduced for the function of Court Judgement Prediction and Explanation(CJPE) [8], aiming to use AI to predict and explain court judgments. This work compares performances of various models like Classical Models(LR,SVM,RF), Sequence Models(BiGRU+Attn), transformer models (BERT [2], RoBERTa [4], XLNet [3]), and hierarchical transformer models. XLNet+BiGRU showed highest accuracy on the ILDC-multi dataset, which is a superior set of the dataset (ILDC-single) used in our work. As an extension of this work, embeddings like InLegalBERT and InCaseLawBERT [16, 17], which have been created by retraining LegalBERT [18] and CaseLawBERT [19] respectively on Indian legal data for achieving better performance. In this work, using InLegalBERT in the encoder module showed the highest macro-F1 score, which justifies the utility of re-training legal language models on jurisdiction-specific legal data.

III. METHODOLOGY

A. Experimental Setup

Table I tabulates the basic requirements for our implementation.

TABLE I. MODEL REQUIREMENTS

Requirements	Details
Hardware	NVIDIA A100 Tensor Core GPU
Software	Google Colab Pro
Programming Language	Python (version 3.x)
Framework	Pytorch
Pre-trained Models	bert-base-uncased [21], xlnet-base-cased [22], roberta-base [23], deberta-large [24], electra-large-discriminator [25], bigbird-roberta-base [26]
Hardware	NVIDIA A100 Tensor Core GPU

B. Dataset

The ILDC dataset [8] mentioned before was created by taking annotated cases from the Indian Supreme Court for the problem of Court Judgement Prediction and Explanation (CJPE). This dataset is not publicly available but can be obtained through a request made to the authors, by filling out a Google form. The dataset is divided into 3 parts:

- ILDC-multi: ILDC-multi includes 35K cases from the Supreme Court of India that feature those cases for which multiple petitions were raised from the same petitioner. This dataset presents a significant challenge as it requires the prediction of multiple, potentially differing outcomes from the same case.
- ILDC-single: It is a subset of the ILDC-multi dataset and contains only those cases where a single petition was filed.
- ILDC-expert: It comprises of 56 documents annotated with expert explanations for the decisions, and is used to evaluate how well judgment prediction algorithms can not only predict outcomes but also provide explanations that align with expert reasoning.

TABLE II. TRAIN-VALIDATION-TEST SPLIT

Dataset	Percentage	Count
Train	80	6073
Validation	10	760
Test	10	760

For our work, we have used the ILDC-single dataset. It has 2 classes: (1: petition accepted, 0: petition rejected). The dataset was further broken down into train, test and validation datasets for effective working of our models. Table II illustrates this division.

C. Detail of Model Used

Hugging Face [20] is a company that provides tools and technologies in the domain of natural language processing (NLP). It is renowned for its 'Transformers' library, which has become a standard in the AI community for building cutting-edge models for a wide range of NLP tasks. Details of the models used are given in Table III.

TABLE III. MODEL DETAILS

Model, Ref.	layers	attn heads	hidden units	Params	Features
bert-base-uncased, [21]	12	12	768	110M	focuses on tasks where case sensitivity is not crucial, initially trained in Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) [2]
xlnet-base-cased, [22]	12	12	768	110M	employs permutation-based training to capture bidirectional contexts, integrates techniques from Transformer-XL, suitable for tasks requiring case sensitivity [3]
roberta-base, [23]	12	12	768	125M	modifies BERT's training process by taking out the Next Sentence Prediction task, using bigger mini-batches for training, and using more data [4]
deberta-large, [24]	24	16	1024	304M	presents a brand-new disentangled attention mechanism that distinguishes between position-based and content-based attention and employs an enhanced mask decoder during pre-training [5]
electra-large-discriminator, [25]	24	16	1024	335M	trains a generator and a discriminator concurrently to distinguish between "real" and "fake" tokens, allows for more efficient learning as the model benefits from all input tokens, not just the masked ones [6]

bigbird-roberta-base, [26]	12	12	12	125M	merges RoBERTa's architecture with Google Research's BigBird capabilities, uses a sparse attention mechanism that extends beyond the typical 512-token limit of standard transformers, combines local, random, and global attention [7]

D. Hyperparameters

From the case descriptions which are present in the 'text' column of the dataset, only the last 512 tokens were used for creating the embeddings and then training the model as it has been mentioned in the previous work [8] that the decisions of the court appear towards the end of the case descriptions and hence using the last 512 tokens perform better. Details of the hyper-parameters used during training of each model, keeping in mind the limitations in resource and the best performance achieved after checking validation accuracies are given in Table IV.

TABLE IV. HYPER-PARAMETERS USED DURING TRAINING

Model	Learning rate	Number of epochs	Batch-size
BERT	5e-6	3	16
XLNet			16
RoBERTa			16
DeBERTa			8
ELECTRA			16
Big Bird			8

IV. RESULTS AND ANALYSIS

The previous work [8] had used the Transformer models BERT [2], XLNet [3] and RoBERTa [4] out of which XLNet + BiGRU combination showed the best result on ILDC-single dataset(accuracy of 76% was mentioned in the paper). In this work, we have used newer models, DeBERTa [5], ELECTRA [6] and Big Bird [7] and compared their performance with their predecessors. Our models were trained with the AdamW optimizer and binary cross-entropy was used to measure the loss. Model performance was recorded using metrics such as macro-F1 (mF1), accuracy, macro-precision (mP) and macro-recall (mR). These results are tabulated in Table V.

TABLE V. PERFORMANCE OF TRANSFORMER-BASED MODELS [1] FOR COURT JUDGEMENT PREDICTION TASK

Model	mP (%)	mR (%)	mF1 (%)	Accuracy (%)
BERT	66.78	66.74	66.76	67.94
XLNet	68.77	66.85	67.16	69.65
RoBERTa	73.04	69.17	69.62	72.61
DeBERTa	77	75.80	76.22	77.49
ELECTRA	77.56	77.16	77.33	78.08
Big Bird	**82.40**	**79.15**	**79.15**	**80.97**

As we can see from the above table, bigbird-roberta-base shows the highest accuracy of 80.97% which is approximately 4% greater than the performance noted in the original work. mF1 score of 79.15% was achieved by the model, which is approximately 3% better than the score shown in the original work(76.55%).

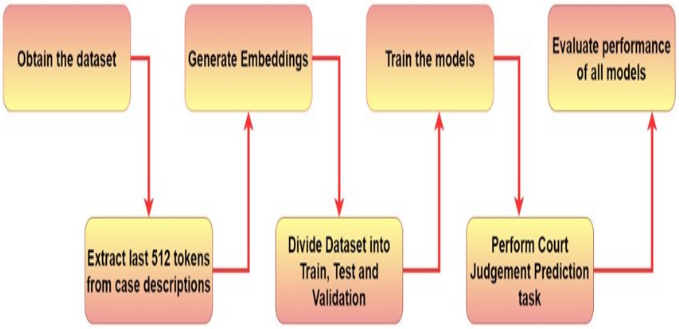
Obtain the dataset
Generate Embeddings
Train the models
Evaluate performance of all models
Extract last 512 tokens from case descriptions
Divide Dataset into Train, Test and Validation
Perform Court Judgement Prediction task

FIGURE 1 WORKFLOW FOR COURT JUDGEMENT PREDICTION TASK

It is to be noted that bigbird-roberta was able to give the best precision, recall, F-1 score and accuracy out of all the models used in this work (results in bold in Table V), though only its base version was used and it was run for a batch-size of 8, which is half of that used for the predecessors (Refer Table IV). This model also exhibited better results than its counterparts, i.e. models (DeBERTa [5] and ELECTRA [6]), whose 'large' category were used and hence had way more parameters.

DeBERTa, ELECTRA, and BigBird models have made significant improvements in performance thanks to some key architectural advancements. DeBERTa has a fancy attention mechanism that separates content and position-based attention. This helps it capture the intrinsic structure of legal texts with even more accuracy. ELECTRA has a unique training mechanism that involves a generator and a discriminator. Unlike other models that only learn from masked tokens, ELECTRA learns from all tokens, making its training more efficient and leading to better performance on different tasks. BigBird, on the other hand, retains the solid pretraining of Roberta while incorporating a sparse attention mechanism that can go beyond the 512-token limit of other standard transformers, making it specifically applicable to lengthy legal documents. These overall improvements lead to a better representation of the complex and voluminous legal texts in the ILDC. Finally, the reduced model size and batch size in bigbird-roberta could have produced an unintended regularizing effect, preventing overfitting and promoting better generalization. These possibilities should be explored further to thoroughly understand why such an outcome was achieved.

V. CONCLUSION

This study has systematically analyzed the performance of six transformer-based models—BERT-base [21], RoBERTa-base [23], XLNet-base [22], DeBERTa-large [24], ELECTRA-large [25], and BigBird-RoBERTa-base [26] on the task of legal judgment prediction using the ILDC-single [8] dataset. Our findings reveal that the newer models, DeBERTa [5], ELECTRA [6], and BigBird [7], offer some improvement in classification accuracy, demonstrating an approximate 4% enhancement over the earlier models. The new models are more precise and hold much promise in enhancing legal judgement prediction. With the best reported accuracy of 80.97%, BigBird-RoBERTa outperforms these other models because it can handle long sequence lengths, and its attention mechanism is very computationally efficient over sparse inputs. This increase, though seemingly modest, is quite significant for making predictions on legal judgments, as legal text is very complex and varies a lot. Legal texts are often characterized by intricate language, substantial citations, or nuanced arguments that complicate precise predictions. An improvement of 4% would thus mean a real advance in the capability of the model to learn and predict judicial results, likely leading to higher reliability and more consistent automated analysis of the law. This can have profound implications for legal practitioners in enabling them to predict case outcomes better and make legal processes highly cost-effective. Although the newer models used could show some improvement over the baselines, there are quite some limitations in our work. We only focused on the predicting court decisions part in this study, while the original work showed the explainability part as well. The original work has incorporated additional layers of attention and BiGRU to enhance the model performances, but we have only used the models in their default form. In addition to these, due to limitations in resources, we could only run these models on a subset (ILDC-single) dataset. In future, we would like to address all these issues and would also like to use the newer models used in this work with specialized embeddings (i.e., embeddings re-trained on data of Indian jurisdiction) like InLegalBERT [16,17] to see how domain-specific embeddings affect the performances of these models on the models used in this work. We would also like to incorporate evolving language models like GPT and Llama into this task.

REFERENCES

[1] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017)

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

[3] Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. "Xlnet: Generalized autoregressive pretraining for language understanding." Advances in neural information processing systems 32 (2019).

[4] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[5] He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. "Deberta: Decoding-enhanced bert with disentangled attention." arXiv preprint arXiv:2006.03654 (2020).

[6] Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. "Electra: Pre-training text encoders as discriminators rather than generators." arXiv preprint arXiv:2003.10555 (2020).

[7] Zaheer, Manzil, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham et al. "Big bird: Transformers for longer sequences." Advances in neural information processing systems 33 (2020): 17283-17297.

[8] Malik, Vijit, 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)

[9] Mandal, Arpan, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. "Automatic catchphrase identification from legal court case documents." In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 2187-2190. 2017

[10] Mandal, Arpan, Raktim Chaki, Sarbajit Saha, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. "Measuring similarity among legal court case documents." In Proceedings of the 10th annual ACM India compute conference, pp. 1-9. 2017

[11] Bhattacharya, Paheli, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. "Hier-spcnet: a legal statute hierarchy-based heterogeneous network for computing legal case document similarity." In Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, pp. 1657-1660. 2020

[12] Bhattacharya, Paheli, Kaustubh Hiware, Subham Rajgaria, Nilay Pochhi, Kripabandhu Ghosh, and Saptarshi Ghosh. "A comparative study of summarization algorithms applied to legal case judgments." In Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41, pp. 413-428. Springer International Publishing, 2019

[13] Ghosh, Saptarshi, and Adam Wyner. "Identification of rhetorical roles of sentences in indian legal judgments." Legal knowledge and information systems: JURIX (2019): 3

[14] Bhattacharya, Paheli, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. "DeepRhole: deep learning for rhetorical role labeling of sentences in legal case documents." Artificial Intelligence and Law (2023): 1-38

[15] Paul, Shounak, 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, vol. 36, no. 10, pp. 11139-11146. 2022

[16] Paul, Shounak, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. "Pre-training transformers on indian legal text." arXiv preprint arXiv:2209.06049 (2022).

[17] Paul, Shounak, 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, pp. 187-196. 2023

[18] Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. "LEGAL-BERT: The muppets straight out of law school." arXiv preprint arXiv:2010.02559 (2020)

[19] Zheng, Lucia, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. "When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings." In Proceedings of the eighteenth international conference on artificial intelligence and law, pp. 159-168. 2021

[20] https://huggingface.co/

[21] https://huggingface.co/google-bert/bert-base-uncased

[22] https://huggingface.co/xlnet/xlnet-base-cased

[23] https://huggingface.co/FacebookAI/roberta-base

[24] https://huggingface.co/microsoft/deberta-large

[25] https://huggingface.co/google/electra-large-discriminator

[26] https://huggingface.co/google/bigbird-roberta-base