A Meta Analysis of Attention Models on Legal Judgment Prediction System

G.Sukanya[1], J.Priyadarshini[2]
School of Computer Science and Engineering
Vellore Institute of Technology, Chennai Campus, Chennai, India

***Abstract*—Artificial Intelligence in legal research is transforming the legal area in manifold ways. Pendency of court cases is a long-lasting problem in the judiciary due to various reasons such as lack of judges, lack of technology in legal services and the legal loopholes. The judicial system has to be more competent and more reliable in providing justice on time. One of the major causes of pending cases is the lack of legal intelligence to assist the litigants. The study in this paper reviews the challenges faced by judgment prediction system due to lengthy case facts using deep learning model. The Legal Judgment prediction system can help lawyers, judges and civilians to predict the win or loss rate, punishment term and applicable law articles for new cases. Besides, the paper reviews current encoding and decoding architecture with attention mechanism of transformer model that can be used for Legal Judgment Prediction system. Natural Language Processing using deep learning is an exploring field and there is a need for research to evaluate the current state of the art at the intersection of good text processing and feature representation with a deep learning model. This paper aims to develop a systematic review of existing methods used in the legal judgment prediction system and about the Hierarchical Attention Neural network model in detail. This can also be used in other applications such as legal document classification, sentimental analysis, news classification, text translation, medical reports and so on.**

Keywords*—Legal judgment prediction; hierarchical attention neural network; text processing; transformer

I. INTRODUCTION

Legal Judgment Prediction (LJP) system helps in assisting litigants and attorneys to improve their work and time efficiency and reduce the risk of making mistakes with feasible judgment suggestions, which includes the prediction of charges, applicable law articles, and prison term based on the case facts [1]. Some of the LJP frameworks predict final judgment as a binary [2] and multilabel text classification [3] for cases in English of European Court Human Rights and Chinese Judgment Online dataset. One of the most important challenges in LJP is unlabelled data and that was tackled using the Long Short Term Memory framework [4] for Indian Supreme Court judgments with headnotes. At the initial stages, Machine Learning methods such as optimized Lasso Regression [5] [6] for Chinese cases and then deep learning models [7] [8] for automated judgment predictions were used. Providing fair and timely justice by the courts is not only the most important obligation of the country but is an important characteristic of democracy [9] [10] [11] [12] [13] [14]. India being the world's largest democracy is still in need of an intelligent judicial system. It also benefits civilians to know the possible judgment result before the trial by describing a case they are concerned about for win or loss rate. Surveys show that India has only twenty percent judges for every million citizens [15] [16].

Following are the benefits of using deep learning in LJP

- Accelerated decision process and outcomes.
- Instant verification of input data.
- Unbiased point of view or opinion.
- Ability to quickly showcase historical cases with similar patterns.
- Easier to spot corruption by identifying cases with high variance in human and AI decisions.

Case Facts of the real World has two main challenges. One is the difficulty faced in encoding lengthy documents and the other is lack of full external information. Existing models used for text classification and prediction like RNN and LSTM work sequentially and takes a long duration in training large corpus for the UK and Chinese court cases. Most of the LJP framework consider judgment prediction as text classification task while some works consider it as a Legal Reading Comprehension [3]. Any text classification has two main phases. The first phase is the representation of the document and the latter is to use a good classifier model. Both are very much important to give good prediction accuracy for new case facts or queries. Earlier deep learning models used BOW (Bag Of Words) and word embeddings, like Word2Vec, GloVe and Fast Text for encoding.

Legal information such as legal cases, contracts, bills are often represented in textual form. Processing of legal text is a grooming area in the current era. In NLP, the legal text is being utilized in various applications like "legal topic classification, court opinion generation and analysis, legal information extraction, legal interpretation and entity recognition" [17] [18] [19] [20] [21] [22]. Models that predict the legal outcomes are emerging for the past few years and these models aid the legal practitioners and citizens with a reduction in the cost of legal issues paves way for faster justice. The legal judgment prediction model can be utilized by lawyers and judges to predict the win or loss chance of a case [23] [24]. Human rights organizations and legal research scholars can adapt them to examine whether fair judicial decisions are given or are do they correlate with biases.

This review aims to discuss one of the most effective deep learning models (HAN) applicable to judgment predictions. The other objectives to be covered in this article are:

- Classify and summarize recent works based on empirical methods of LJP and conceptual literature of text classification.
- Encoding and Decoding architecture using a transformer.
- Showcase the important features and challenges of existing LJP methods.
- Future Scope with various applications.

The following section of this article is organized as follows. Section II briefs up the related work. In Section III, a review of the HAN and Transformer model is done. Major Findings and suggestions are revealed in Section IV. Finally, Section V contains the conclusion.

II. Background

Machine learning has begun to revolutionize various industries already and it would aid India's legal system in producing legal judgments with higher accuracy and precision. If law firms/advocates in India use this AI application for risk assessment for an out-of-court settlement/Alternative Dispute Resolution (ADR) mechanism, the number of cases would reduce and faster justice could be provided [25][26]. In India, legal search engines use Artificial Intelligence for legal analytics and visualization through its Case Map and Taxonomy and it also provides an analysis of judge's disposition through data analyzed by AI. Areas in legal that already use ML are client due diligence and contract management.

The problem of judgment prediction has taken a great attraction in the legal research area. This section describes Empirical Literature on Legal judgment Prediction methods, Conceptual Literature on Text Classification Methods and transformer model in detail.

A. Empirical Literature on Legal Judgment Prediction Methods

Most of the legal judgment prediction research works were classified using Binary Classification. Zhong *et al*. [1] have introduced the TOPJUDGE, a topological multi-task learning framework that shows the dependencies among subtasks of case facts as a Directed Acyclic Graph. The challenge faced in TOPJUDGE is that it failed to exhibit interaction between subtasks. Besides, a Convolutional Neural Network-based encoder was utilized to generate the fact descriptions. The outcomes of the TOPJUDGE model are law articles, charges and terms of penalty. TOPJUDGE exhibited a higher consistency over the existing models.

A new HAN model with Google's Bidirectional Encoder Representations from Transformers (BERT) was developed by Chalkidis *et al*. [2] for the prediction of cases from the European Court of Human Rights. A wide variety of neural models like BiGRU-Att, HAN, LWAN, BERT and Hierarchical BERT were evaluated on the proposed dataset. The research work had surpassed the drawbacks of the existing models and does (1) binary violation classification (2) multi-label classification (3) case importance prediction using sentence scores. Long *et al*. [3] formulated the Legal Reading Comprehension framework for the judgment predictions which works based on answering the questions of Comprehension rather than using the text classification method. This model was developed to handle multiple and complex textual inputs. Also, they have conceptualized the AutoJudge framework to infuse law articles for judgment prediction. The results of AutoJudge were better than the base model in terms of consistency and reliability. The problem of unavailability of labeled data for the prediction task is tackled by assigning classes/scores to sentences in the training set, based on their match with reference summary produced by humans using LSTM by Anand et al. [4].

In 2020, Guo *et al*. [5] introduced the TenLa by blending the concepts of both the tensor decomposition and an optimized Lasso regression model. Based on the similarities between legal cases the judgment charges were predicted. The major process undergone in the proposed works was: (a) ModTen to legal cases as three-dimensional tensors, (b) ConTen to decompose tensors obtained by ModTen (c) OLass, which was trained with the Core tensors got by ConTen. The results of the TenLa had exhibited higher accuracy than the traditional models.

Guo *et al*. [6] have preferred TenRR that amalgamates the tensor decomposition and ridge regression for judgment prediction of legal cases, and the proposed model had enclosed three major contributions. In the initial contribution, RTenr was developed as a tensor representation method to express the legal cases as three-dimensional tensors. In the second contribution, the ITend was introduced to decompose the original tensors representing legal cases into core tensors. In the contribution, the ORidge was built to construct an optimized Lasso judgment prediction model for legal cases. The results of the proposed work had exhibited higher accuracy than traditional methods for judgment prediction.

Chen *et al*. [7] analyzed the case description and predicted the judgment employing the deep learning model. The outcome of the deep learning model was in the form of three aspects: penalty, accusation and legal provisions. They predicted the latter aspects based on the FastText and TextCNN method. The resultant of the proposed judicial decision-making model was more accurate and persuasive. Yang *et al*. [8] proposed a Multi-Perspective Bi-Feedback Network. It had a word-level attention mechanism based on the topology structure among subtasks. Also a multi-perspective forward prediction and backward verification framework were designed to make use of the dependencies among multiple subtasks effectively. The word collocations features of fact descriptions were integrated into the proposed work to distinguish cases with similar descriptions but different penalties. The resultant of the proposed work had achieved significant improvements in terms of prediction accuracy. Shang Li et al. [12] discussed a Multichannel Attentive Neural Network(MANN) framework which predicts applicable charges, punishment terms and articles for Chinese court cases based on case facts for single defendant person using two-tier hierarchical architecture. K.

Zhu et.al [22] proposed Sequential Generation Network using a nested hierarchical attention mechanism for multi-charge prediction with single case defendants.

Kongfan Zhu et al [34] proposed Transformer-Hierarchical-Attention-Multi-Extra (THME) Network to extract the semantics of external information of the fact for prediction of Legal judgment based on multiple classes.

Jerrold sho et al [35] proposed a comparative study on NLP methods against statistical models on 6227 novel Singapore Supreme Court Judgments for the topic model, word embedding, and language model.

Hui Wang et al [36] have designed the LJP framework based on FastText and TextCNN for multilabel text classification for accusation prediction.

B. Limitations

We see that most of the existing legal judgment prediction system is applicable for a single defendant person and the judgments are predicted only based on case facts. In the real world, along with case facts, other external information such as evidence and emotions play a vital role in judgment which is a drawback found, which indirectly affects the prediction accuracy of the judgment. Also the problem of unavailability of structured legal data prone to an imbalanced dataset, gives a biased prediction. The lengthy case fact also is found to be a major challenge. Practically, Casefact with legal opinion comes around 60 to 100 pages. So a great amount of time is spent on extracting important points from them to make it into around 200 words per document either manually or by using a text summarization tool for a basic RNN model to work on it. In Table I, the features and challenges of existing Judgment prediction works are enlisted**.**

TABLE I. FEATURES AND CHALLENGES OF EXISTING PREDICTION WORKS

Author [Citations]	**Methodology**	**Data Sets Used**	**Features**	**Challenges**
Zhong *et al*. [1]	TOPJUDGE	**CJO**, **PKU**, and **CAIL**. CJO has criminal cases published by the Chinese government from China Judgement Online. PKU contains criminal cases published by Peking University Law Online CAIL(Chinese AI and Law Challenge)	✓ integrates multiple subtasks and make judgment predictions through topological learning framework ✓ judgment predictions through topological framework ✓ neural encoder for fact representation and subtasks with DAG dependencies.	✓ Limited to work on single defendants and charges. ✓ Need to explore how to infuse temporal factor into LJP
Chalkidis *et al*. [2]	HAN model with BERT	English legal judgment prediction dataset	✓ binary violation classification of articles ✓ multi-label classification of charges ✓ case importance prediction using sentence scores	✓ few-shot learning is not taken into account ✓ need to break the problem of charge prediction into different subtasks
Long *et al*. [3]	AutoJudge	Chinese Referee Document Network	✓ captures the complex semantic interactions among facts, pleas, and laws based on legal reading comprehension framework ✓ Improved F1 score, accuracy and precision	❖ don't have access to groundtruth law articles ❖ increases the computational complexity ❖ reduces the accuracy and stability
Anand *et al*. [4]	neural network	Indian Supreme Court Judgments (1947 to 1993)	✓ tackles the problem of unavailability of labeled data ✓ Uses Feed Forward Neural Network and Long Short Term Memory for case text summarization	❖ Higher cost ❖ Higher computational complexity ❖ Need sentence simplification approaches for complex and long sentences
Guo *et al*. [5]	TenLa	3,000,000 legal cases in the past five years from multiple provinces and cities in China	✓ higher accuracy ✓ removes redundant, meaningless, and inaccurate information	❖ Need to prevent over fitting ❖ Complex
Guo *et al*. [6]	TenRR	Chinese Referee Document Network	✓ greatly reduce the dimension of original tensors ✓ Removal of the meaningless and inaccurate information in original tensors	❖ Need to improve the accuracy of predictions
Baogui Chen *et al*. [7]	FastText and TextCNN	CAIL 2018 data set	✓ more accurate and persuasive decision-making ✓ Better in accuracy, and recall rate	❖ Need to improve the accuracy of model prediction
Yang *et al*. [8]	Multi-Perspective based BiFeedback Network (MPBFN) and a Word Collocation Attention (WCA) mechanism	Chinese AI and Law challenge (CAIL2018)	✓ improves the overall performance ✓ improve the performance of multitasking	❖ Need to reduce the misjudgment of penalty prediction

C. Conceptual Literature on Text Classification Methods

Before applying text classification methods text preprocessing has to be done. Preprocessing of raw text data gives good results on classification

1) Based on text preprocessing: Jin Wang et al [37] in 2019 implemented a regional CNN with LSTM model which comprises of two parts: to predict the Valency Arousal ratings of texts on Stanford Sentiment Treebank 1 dataset. The local information within the sentences is observed using regional CNN and long-distance dependencies are extracted by using LSTM across sentences that can be considered in the prediction process. Hao Fei et al [38] in 2020 finds multiple emotions using text as a multilabel classification problem using variational autoencoder and capsule module, to extract rich features in a sentence. Latent Topic attention-based routing algorithm is used in capsule module for pertaining the task. Zhang et al [39] in 2019 proposed a coordinated CNN-LSTM attention model to capture meaningful emotional dependent information, where filters of different widths are used in between word representation and pooling unit to get semantic information. Hao Peng et al [40] in 2019 uses a graphical capsule neural network model to capture rich information through a routing mechanism. This type of neural network model is found to be better than LSTM-RNN while considering long-term dependency. Zhongqing Wang et al [41] proposed the Hierarchical Attention Model in 2020 using Linguistic Attention based on argument representation, dependency representation and sentiment representation to extract meaningful words.

Word segmentation and tokenization are the initial steps in Natural Language Processing. The raw content of case fact description has to be preprocessed according to the application we choose. Mingjie Ling et al [46] use ELMo(Embeddings from Language Models) word embedding Language Model to overcome polysemy phenomena in word representation. The work of researchers Matthew E Peters et al [47] has proved that ELMO word embeddings are good to avoid polysemy phenomena than Skip-gram models and other word embeddings, like word2vec and GloVe which were widely used earlier. The main advantage of ELMo is that they have different word vectors under different contexts for the same word.

D. Transformer Model

Transformers which are at their budding stage in the application can be replaced for other methods in encoding and decoding text representation. They are pre-trained word embedding models used for text summarization and translation. Original transformer models have a large number of parameters and are compute-intensive. Also, they can be used for fixed-length documents only. Some of the transformer models are G-BERT, BioBERT, M-Bert, Trans-Bert, Clinical BERT, etc.

Jacob Devlin et al [42] in 2019 proposed deep bi-directional transformers by smoothing the existing pre-trained BERT model by adding one additional output layer, which is simple and powerful compared to the existing RNN models. Zhenzhong Lan et al [43] in 2020 proposed A Lite BERT which is better than BERT in terms of less memory consumption of the model by using two parametric reduction techniques. Chi Sun et al [44] in 2020 worked on different types of fine-tuning like single task and multitask tuning of parameters of the BERT for text classification. The tuned hyperparameters were then applied on eight different datasets and analyses were done. Zihang Dai et al [45] in 2020 proposed an attention model using XLnet which could learn 80% more dependency than RNN and better than existing vanilla transformers. The drawback is that the transformer model is highly compute-intensive and that it has a little struggle in handling negative sentences [47].

III. HAN and Transformer Model

Recurrent Neural Networks in deep learning models were widely used in Natural language processing tasks. Though it can capture contextual information over long distances compared to CNN, it suffers from the Vanishing Gradient and Exploding Gradient problem. While passing the information down between hidden layers during backpropagation, larger derivatives increase exponentially and then explode eventually creating Exploding Gradient problem. Similarly for smaller derivatives, the gradient decreases and vanishes eventually creating the Vanishing Gradient problem. To solve this semantic bias CNN with the max-pooling stage is adopted to get the most important information from text. Again approaches using the basic CNN model cannot represent the text semantically due to fixed window size. Other solutions for vanishing and exploding gradient problems are reducing the number of layers, by limiting the gradient size and by considering random initialization of weights [48] between the hidden layers. The gradient problems of RNNs were overcome by long short-term neural networks (LSTMs). It captures the contextual information of longer context in the documents than basic RNN. But, LSTM works unidirectional and sequentially which takes longer time consumption. This limitation in unidirectional LSTM was overriden by using bidirectional LSTMs, where we can read the context from both directions. Nowadays attention mechanisms were infused in the existing framework of RNN which is the hierarchical attention models. In this survey importance of the Hierarchical Attention Neural Network and transformer model has been studied and analyzed.

A. HAN

In NLP, the advancement in machine learning is making the decision-making capability a more relevant one. Bots in the market are already used to ‘smart search’ existing judgments and rulings to help in the preparation of a new case. Most of the existing judgment prediction models reviewed by researchers are based on traditional machine learning algorithms [28].

Nowadays, attention mechanism has been used in deep learning across a wide variety of contexts ranging from image captioning, image generation, and language modeling and translation. Hence, in the judgment prediction model, the attention mechanism is used to extract important words from the lengthy document, by assigning more weights to them.

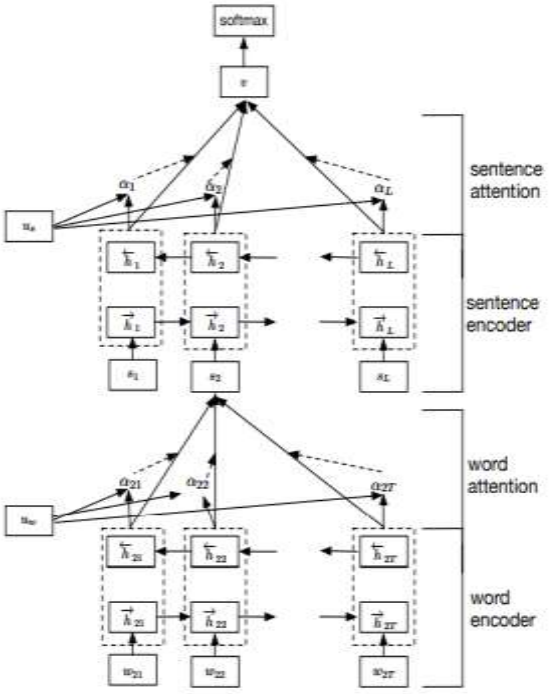

Fig. 1. General Architecture of Hierarchical Attention Network (HAN).

A hierarchical attention model is used in modeling the hierarchical relationships of words and sentences in a document for document classification. The Hierarchical Attention Network (HAN) [10] [22] [26] [27] [29] [30] [31] [32] [33] is a deep-neural-network that is utilized for Document Classification. A HAN attempts to classify a document based on the knowledge it can infer about the document from its composite parts, in other words, the sentences and words that make up the document. The 'hierarchical' in HAN comes from the design that this knowledge is built hierarchically, starting from using the words in a sentence and followed by using the sentences in a document [29].

The overall architecture of the Hierarchical Attention Network (HAN) is shown in Fig. 1. It consists of several parts: "a word sequence encoder, a word-level attention layer, a sentence encoder and a sentence-level attention layer". The word vectors are encoded using the word sequence encoder and the word-level attention layer is utilized for aggregating the information of the informative words that do not contribute equally. Then, the sentence vectors are encoded in the word sequence encoder and the sentence level attention mechanism is utilized to reward attention (weights) to the sentences that are clues to correctly classify a document.

B. Transformer Model Architecture

Transformers have taken a vivid shape in NLP since 2019. Researchers use transformers nowadays in NLP while using RNN, LSTM, GRU, etc. Integration of transformer models with language models is at its budding stage in all NLP tasks. The Attention mechanism makes transformers have extremely long-term memory. A transformer model can remember all previously generated tokens that have been generated. Also, they have infinite reference windows, thereby overcoming the short reference window of RNN. It is an encoder-decoder architecture. The inputs given into the encoder are represented as a continuous vector.

The main benefit of using transformer-based models are:

1) The input tokens are not processed sequentially one by one as in RNN, rather the full sequence is taken as one input at a single shot.

2) Also labeled data is not necessary. Just giving a large amount of unlabeled data is enough to train a transformer-based model.

Bidirectional Encoder Representation Transformer (BERT) is a multiheaded attention-based encoder-decoder used as a pre-trained model for the word to a vector representation. It can be applied for Legal Judgment prediction which is based on lengthy case facts in an efficient way [2]. Fig. 2 shows the steps used in BERT architecture.

1) The input is fed into a word embedding layer where ever word is mapped into a vector and a lookup table is formed

2) Positional information is sent into embeddings since the encoder of the transformer doesn't have it. Sine and Cosine functions are used for positional encoding, at every even and odd index.

3) Encoder layer has the information for the entire sequence. It contains two subgroups. Multiheaded attention and then a Fully Connected Network. Multiheaded attention model uses a self-attention mechanism, i.e. it relates each word in input to other words. Query, key and Value factors are used to create a self-attention mechanism.

4) A dot product between Query and the key value is done to produce a score matrix. This gives a clue about how much importance should be given to each word in a sentence. Greater the score, the more important those words are. Thus queries are mapped to keys.

5) On the scaled score, a Softmax is applied which gives attention weights between 0 and 1.After doing this, greater scores are made still higher and vice-versa.

6) Multiply the above Softmax output with the value vector, which gives the output vector.

7) Decoder layer is also similar to the encoder layer and it has two multiheaded attention layers and a feedforward layer. It is appended with the linear layer that acts as a classifier and a Softmax is applied to get the probabilities of words. Masking is done to make all the negative values represented as zero.

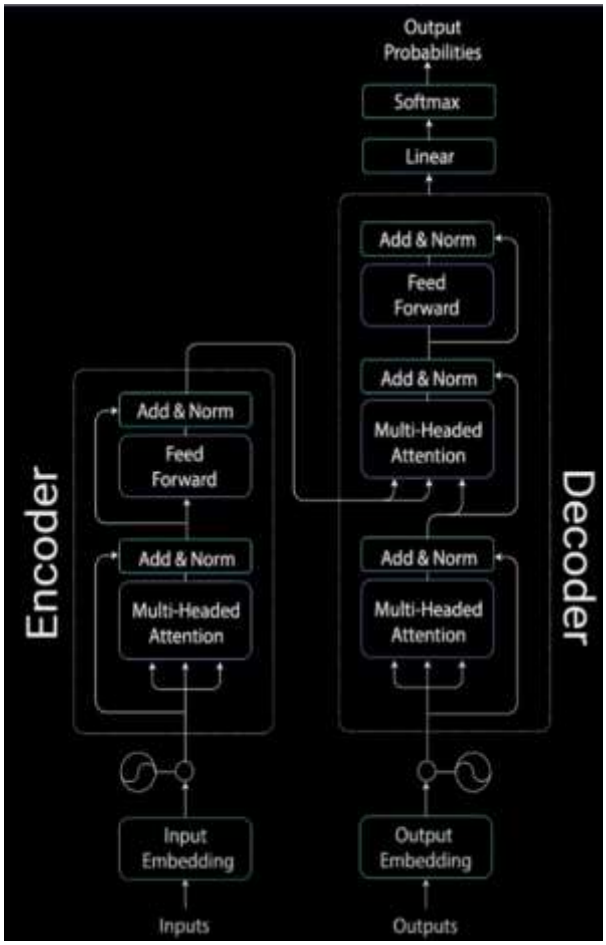

Fig. 2. Transformer Model.

IV. Major Findings

The explanation given in section III conveys the working procedure of HAN without transformer models and Transformer models with attention mechanism in a detailed way.

Hierarchical attention models are used mainly to automatically read and extract information from case facts efficiently. They differ from classic sequence to sequence model in two main ways. Most of the hierarchical model has two-tier architecture one for word attention vector and the other for sentence attention vector. The Encoder passes more data to the decoder instead of passing the last hidden state of the encoding stage, which turns to be advantageous than RNN. It consumes more time to train word embedding for a particular application if we use an attention model without transformers. Traditional RNN needs labeled data for the word which is not necessary if we use the pre-trained transformer model.

Attention mechanism with transformer model also has its pits and falls. The main benefits are it can cope up with any application using its good contextualized inbuilt word embedding which supports large words. A pre-trained transfer model in word to vector representation proves to be a time-saving one. On the other hand, original transformer(BERT) models are very big which are prone to intense computation. Due to the large number of parameters present in the transformer model, it is advised to fine-tune the architecture according to the needs of the application. BERT has its own limitations [26]. Firstly, it fails to capture longer-term dependency beyond the predefined context length. The maximum length of the sequence for BERT is 512 tokens which have to be taken into consideration. For shorter sequence padding has to be done and for longer sequence, the sentence has to be trimmed. Secondly, it struggles in handling negative sentences. Thirdly, it is unable to generalize to positions beyond those undergone for training.

There are different types of transformer models available. Though BERT has been renowned as the most efficient one on many NLP tasks, now it's overrun by XLNet from Google. XLNet uses the permutation language modeling concept in a sentence. CamemBERT is used mainly for legal tasks enduring with Part Of Speech tagging and Named Entity Recognition with less number of parameter compared to basic BERT, which in turn takes less time for computation. Pappagari et al. [26] proposed fine tuned BERT models such as Transformer over BERT(ToBERT) and Recurrence over BERT(RoBERT) methods for classification of long documents which performed better than pre-trained BERT. Therefore we suggest using HAN with fine-tuned transformer model for future endeavors in judgment predictions.

V. Conclusions

The purpose of this review was to identify an effective deep learning model used for judgment predictions. Based on the analysis conveyed integration of Hierarchical Attention Neural network models with fine-tuned transformer concept will give an efficient improvement based on quality and time in judgment prediction. Also, the improvement of multilabel classification for complex case facts with multiple defendants and charges still needs further investigation. A future exploration into the following legal areas such as summarization of legal judgment, legal data curation, and legal document simplification could be very much useful for the legal society.

References

[1] Haoxi Zhong, Zhipeng Guo∗ , Cunchao Tu, Chaojun Xiao, Zhiyuan Liuy , Maosong Sun,"Legal Judgment Prediction via Topological Learning",Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,pp.3540–3549,2018.

[2] Ilias Chalkidis, Ilias Chalkidi and Nikolaos Aletras, "Neural Legal Judgment Prediction in English",arXiv:1906.02059v1 [cs.CL],Jun 2019.

[3] Shangbang Long, Cunchao Tu, Zhiyuan Liu, Maosong Sun,"Automatic Judgment Prediction via Legal Reading Comprehension",arXiv:1809.06537v1 [cs.AI],Sep 2018.

[4] Deepa Anand, Rupali Wagh,"Effective Deep Learning Approaches for Summarization of Legal Texts",Journal of King Saud University - Computer and Information Sciences,2019.

[5] Xiaoding Guo,Hongli Zhang, Lin Ye,Shang Li,"TenLa: an approach based on controllable tensor decomposition and optimized lasso regression for judgement prediction of legal cases",Applied Intelligence,2020.

[6] X. Guo, H. Zhang, L. Ye, S. Li and G. Zhang, "TenRR: An Approach Based on Innovative Tensor Decomposition and Optimized Ridge Regression for Judgment Prediction of Legal Cases," in IEEE Access, vol. 8, pp. 167914-167929, 2020, doi: 10.1109/ACCESS.2020.2999522.

[7] Baogui Chen,Yu Li,Shu Zhang, Hao Lian,"A Deep Learning Method for Judicial Decision Support",IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C),2019.

[8] Wenmian Yang, Weijia Jia, Xiaojie Zhou and Yutao Luo,"Legal Judgment Prediction via Multi-Perspective Bi-Feedback Network",Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19),2019.

[9] D. Huang and W. Lin, "A Model for Legal Judgment Prediction Based on Multi-model Fusion," 2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE), Xiamen, China, pp. 892-895, 2019.

[10] S. Li, B. Liu, L. Ye, H. Zhang and B. Fang, "Element-Aware Legal Judgment Prediction for Criminal Cases with Confusing Charges," 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, , pp. 660-667, 2019.

[11] C. Wang and X. Jin, "Study on the Multi-Task Model for Legal Judgment Prediction," 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, pp. 309-313, 2020.

[12] S. Li, H. Zhang, L. Ye, X. Guo and B. Fang, "MANN: A Multichannel Attentive Neural Network for Legal Judgment Prediction," in IEEE Access, vol. 7, pp. 151144-151155, 2019. doi: 10.1109/ACCESS.2019.2945771.

[13] L. Chen, N. Xu and Y. Wang, "Legal Judgment Prediction with Label Dependencies," 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Calgary, AB, Canada,

[14] L. Yuan et al., "Automatic Legal Judgment Prediction via Large Amounts of Criminal Cases," 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, pp. 2087-2091.pp. 361-365, 2020.

[15] R. Sil and A. Roy, "A Novel Approach on Argument based Legal Prediction Model using Machine Learning," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, pp. 487-490, 2020.

[16] X. Yang, G. Shi, J. Lou, S. Wang and Z. Guo, "Interpretable Charge Prediction with Multi-Perspective Jointly Learning Model," 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, pp. 1850-1855, 2019.

[17] V. G. Pillai and L. R. Chandran, "Verdict Prediction for Indian Courts Using Bag of Words and Convolutional Neural Network," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, pp. 676-683, 2020.

[18] C. Jin, G. Zhang, M. Wu, S. Zhou and T. Fu, "Textual content prediction via fuzzy attention neural network model without predefined knowledge," in China Communications, vol. 17, no. 6, pp. 211-222, June 2020.

[19] T. Goto, K. Sano and S. Tojo, "Modeling Predictability of Agent in Legal Cases," 2016 IEEE International Conference on Agents (ICA), Matsue, pp. 13-18, 2016.

[20] Y. Yin, F. Zulkernine and S. Dahan, "Determining Worker Type from Legal Text Data using Machine Learning," 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Calgary, AB, Canada, pp. 444-450, 2020.

[21] K. Kowsrihawat, P. Vateekul and P. Boonkwan, "Predicting Judicial Decisions of Criminal Cases from Thai Supreme Court Using Bi-directional GRU with Attention Mechanism," 2018 5th Asian Conference on Defense Technology (ACDT), Hanoi, pp. 50-55, 2018.

[22] K. Zhu, B. Ma, T. Huang, Z. Li, H. Ma and Y. Li, "Sequence Generation Network Based on Hierarchical Attention for Multi-Charge Prediction," in IEEE Access, vol. 8, pp. 109315-109324, 2020.

[23] B. Chen, Y. Li, S. Zhang, H. Lian and T. He, "A Deep Learning Method for Judicial Decision Support," 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C), Sofia, Bulgaria, , pp. 145-149, 2019.

[24] J. Guo, B. Wu and P. Zhou, "BLHNN: A Novel Charge Prediction Model Based on Bi-Attention LSTM-CNN Hybrid Neural Network," 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC), Hong Kong, Hong Kong , pp. 246-252, 2020.

[25] Chang Yin,Cuiqing Jiang,Zhao Wang,"Evaluating the credit risk of SMEs using legal judgments",Decision Support Systems,2020.

[26] Raghavendra Pappagari, Piotr ˙Zelasko, Jes´us Villalba, Yishay Carmiel, and Najim Dehak,"Hierarchical Transformers For Long Document Classification", arXiv:1910.10781v1 [cs.CL] 23 Oct 2019.

[27] Luo, B., Feng, Y., Xu, J., Zhang, X., & Zhao, D. (2017). Learning to Predict Charges for Criminal Cases with Legal Basis. arXiv preprint arXiv:1707.09168.

[28] Rafe Athar Shaikh,Tirath Prasad Sahu,Veena Anand,"Predicting Outcomes of Legal Cases based on Legal Factors using Classifiers",Procedia Computer Science,2020.

[29] D. Huang and W. Lin, "A Model for Legal Judgment Prediction Based on Multi-model Fusion," 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE), Xiamen, China, pp. 892-895,2019.

[30] Sajad Mousavi,Fatemeh Afghah,U. Rajendra Acharya,"HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks",Computers in Biology and Medicine,2020.

[31] Fa Li,Zhipeng Gui,Yichen Lei,"A hierarchical temporal attention-based LSTM encoder-decoder model for individual mobility prediction",Neurocomputing, 2020.

[32] Yirong Zhou,Jun Li,Luo Chen,"A spatiotemporal hierarchical attention mechanism-based model for multi-step station-level crowd flow prediction",Information Sciences,2021-33.

[33] Shuning Xing,Fang'ai Liu,Tianlai Li,"A hierarchical attention model for rating prediction by leveraging user and product reviews",Neurocomputing,2019 -34.

[34] Kongfan Zhu, Rundong Guo , Weifeng Hu, Zeqiang Li, and Yujun Li "Legal Judgment Prediction Based on Multiclass Information Fusion" ,in Hindawi Complexity Volume 2020, Article ID 3089189, 12 pages,https://doi.org/10.1155/2020/3089189.

[35] Jerrold sho, Legal Area Classification: "A Comparative Study of Text Classifiers on Singapore Supreme Court Judgments" Domains @agc.gov.sg @smu.ac.in.

[36] Hui Wang, Tieke He, Zhipeng Zou, Siyuan Shen, Yu Li "Using Case Facts to predict accusation based on deep learning",in 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C).

[37] Jin Wang, Liang-Chih Yu, Member, IEEE, K. Robert Lai, and Xuejie Zhang 11 December 2019, "Tree-Structured Regional CNN-LSTM Model for Dimensional Sentiment Analysis", Published in: IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 28) pp: 581 – 591,Date of Publication: 11 December 2019 doi: 10.1109/TASLP.2019.2959251,Publisher: IEEE.

[38] Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji,"Topic-Enhanced Capsule Network for Multi-Label Emotion Classification", Published in: IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 28) Page(s): 1839 – 1848 Date of Publication: 10 June 2020 doi: 10.1109/TASLP.2020.3001390 Publisher: IEEE.

[39] Zhang Yangsen, Zheng Jia1 , Jiang Yuru, Huang Gaijuan And Chen Ruoyu ,"A Text Sentiment Classification Modeling Method Based on Coordinated CNN-LSTM-Attention Model" Published in: Chinese Journal of Electronics (Volume: 28 , Issue: 1 , 1 2019)Page(s): 120 – 126,Date of Publication: 22 August 2019,doi: 10.1049/cje.2018.11.004,Publisher: IET.

[40] Hao Peng, Senzhang Wang,Lihon Wong,Qiron Gong,"Hierarchical Taxonomy-Aware and Attentional Graph Capsule RCNNs for Large-Scale Multi-Label Text Classification", Published in: IEEE Transactions on Knowledge and Data Engineering (Early Access)Page(s):1–1, 2019,doi: 10.1109/TKDE.2019.2959991,Publisher: IEEE.

[41] Zhongqing Wang , Qingying Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou, "Neural Stance Detection With Hierarchical Linguistic Representations", Published in: IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 28)Page(s):635–645,Date of Publication: 03 January 2020,doi: 10.1109/TASLP.2020.2963954,Publisher: IEEE.

[42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", Anthology ID:N19-1423 Volume:Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),June 2019, Minneapolis, Minnesota, Publisher:Association for Computational Linguistics, doi:10.18653/v1/N19-1423.

[43] Zhenzhong Lan, Mingda Chen, Sebastian Goodman1 Kevin Gimpel, Piyush Sharma Radu Soricut,"ALBERT: A Lite Bert For Self-Supervised Learning Of Language Representations" in arXiv:1909.11942v6 [cs.CL] 9 Feb 2020.

[44] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang, "How to fine-tune BERT for text classification?" , in arXiv:1905.05583v3 [cs.CL] 5 Feb 2020.

[45] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le2, Ruslan Salakhutdinov Carnegie Mellon University,GoogleBrain "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Count", arXiv:1901.02860v3 [cs.LG]2 Jun 2019.

[46] Mingjie Ling; Qiaohong Chen; Qi Sun; Yubo Jia,"Hybrid Neural Network for Sina Weibo Sentiment Analysis", Published in: IEEE Transactions on Computational Social Systems (Volume: 7, Issue: 4, Aug. 2020),Page(s): 983 – 990,doi: 10.1109/TCSS.2020.2998092 .

[47] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. "Deep contextualized word representations" in arXiv,preprint arXiv:1802.05365.

[48] J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity (2020), ICLR2020.