This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

JusticeAI: A Large Language Models Inspired Collaborative & Cross-Domain Multimodal System for Automatic Judicial Rulings in Smart Courts

NAGWAN ABDEL SAMEE[1], MAALI ALABDULHAFITH[1], S MUHAMMAD AHMED HASSAN SHAH[2,*], AND ATIF RIZWAN [3,*]

[1]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[2]School of Civil Engineering, Central South University, Changsha, 410075, China
[3]Department of Electronic Engineering,Kyung Hee University, Yongin, Republic of Korea

Corresponding author: Syed Muhammad Ahmed Hassan Shah (e-mail: syedmahmedhassan321@gmail.com), Atif Rizwan (e-mail: atifrizwan@khu.ac.kr)

This research was supported by Energy Cloud R&D Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT (2019M3F2A1073387), and this work was supported by the Institute for Information & communications Technology Promotion (IITP) (NO. 2022-0-00980, Cooperative Intelligence Framework of Scene Perception for Autonomous IoT Device). Any correspondence related to this paper should be addressed to Atif Rizwan.

ABSTRACT There has been a significant amount of attention in recent years toward the utilization of artificial intelligence (AI) in the realm of legal decision-making. This growing pattern reveals a higher interest among academics and legal professionals in utilizing AI technologies to enhance a number of legal system components. Artificial intelligence (AI) tools, such as machine learning and natural language processing, possess the capacity to analyze vast quantities of legal data, extract valuable insights, and facilitate decision-making processes. The primary aim of this study is to develop a sophisticated framework for judicial decision-making that incorporates methodologies from artificial intelligence and utilizes the dataset from the European Court of Human Rights (ECHR). The utilization of this methodology holds promise in improving the decision-making procedures of legal professionals and reducing the laborious task of manually analyzing legal documents. As a result, this can lead to the facilitation of more accurate predictions of court rulings. Our research introduces a hybrid ensemble model designed specifically for smart court rulings. This innovative approach harnesses the benefits of pre-trained embeddings and large language models to accurately predict court decisions. By utilizing the power of pre-existing embeddings and incorporating the capabilities of advanced language models, our proposed model demonstrates enhanced predictive accuracy and efficiency in the context of court rulings. We also focus on the models' feasible interpretability and highlight their ability to determine key factors in legal decision-making. We attain a notably high accuracy score of around 83%. Our research illuminates how large language models (LLMs) and advanced deep learning techniques can be utilized to predict legal outcomes.

INDEX TERMS Natural Language Processing, Multimodal Networks, Smart Courts, Deep Learning, Transformers

I. INTRODUCTION

Researchers in the field of law have used philosophical research methods for thousands of years. These methods involve outlining laws, feasible problem-solving, and adding philosophical comments to legislation and case law [1]. In contemporary times, numerous courts follow the need to advance accessibility and reusability of public sector information by publishing analyzed cases on the internet, thereby creating ample opportunities for automated analysis of legal data. Presently, computers are engaged in the task of automatically summarizing legal information, extracting relevant information, categorizing legal resources, and performing

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

statistical analysis.

The concept of automating and partially automating the legal field is common [2]. The application of NLP and Image processing have been a longstanding practice within the field of criminology, and forensics [3]–[10]. Text classification has been applied in the field of forensic linguistics. In contrast to previous eras, wherein manual analysis was conducted as exemplified by the Unabomber case, contemporary advancements have enabled the automation of numerous analytical tasks. Recent advancements in technology have led to the development of AI, and ML, that can accurately identify various attributes such as gender [11], age [12], personality traits [13], [14].

Predicting court decisions and analyzing legal texts are of growing interest among academics as means to improve the efficiency and accuracy of the legal decision-making process. In this research, we use artificial intelligence methods to make predictions about the outcomes of cases heard by the European Court of Human Rights (ECtHR) [14]. There has been a lot of interest in using AI in the judicial system as a decision-making tool in recent years. Predicting court decisions and analyzing legal texts are two areas where artificial intelligence has been the subject of several investigations. To anticipate the results of US Supreme Court cases, for instance, machine learning algorithms were utilized in a study by Daniel Martin Katz et al. [15], [16].

Manual study and interpretation of legal texts and judgments can be time-consuming and prone to error in traditional decision-making methods. With the aim of enhancing accuracy and efficiency, our focus is on implementing AI-driven technologies to automate and optimize decision-making procedures. The central objective of this research is to establish an intelligent framework for judicial decision-making by integrating artificial intelligence methodologies. This approach holds the potential to refine decision-making among legal practitioners, reducing the manual analysis of legal texts and thereby expediting court judgment predictions. The research presents an advanced multimodal system that incorporates language models and pre-trained embeddings, aimed at enhancing the predictive accuracy of the model. The key contributions of this paper are

1) This study presents an innovative method to improve the decision-making procedures within the court system by harnessing text-based case data within an intelligent framework. The system is carefully designed to improve decisions by using advanced computer methods and analyzing relevant text information.
2) This research introduces an innovative hybrid ensemble network known as the Cross-Domain Neural Knowledge Fusion System (CDKF). The label "cross-domain" signifies the incorporation of diverse components within the design, incorporating a pre-trained embedding module alongside a comprehensive language model module. These elements engage with text-based data using distinctive approaches, and the attributes extracted from each element are amalgamated through a process known as feature fusion. As a result, the CDKF model produces final court rulings.
3) In this study, we conduct a comparative analysis using different LLMs, namely BERT, ALBERT, RoBERTa, and Distilled BERT. We also perform a comparative analysis with pre-trained embeddings. This comparison helps us analyze the performance of different models individually.

The main goal of this paper is to utilize different state-of-the-art language models and pretrained embeddings to predict court rulings. In this research, we propose a new and novel system called the Cross-Domain Neural Knowledge Fusion System (CDKF) for court rulings prediction in smart courts. The system is based on an ensemble learning approach that uses both LLMs and pretrained embeddings. Firstly, we utilize pretrained embeddings to predict the probabilities of court rulings. We use four different embeddings: 50D, 100D, 200D, and 300D. After fusing all the features generated by these embeddings, we predict the outcomes of court cases. In the second phase, we employ LLMs, specifically four language models, to predict the probabilities. After obtaining the predictions, we combine the results of both language models and pretrained embeddings. The use of the ensemble approach and feature fusion enhances the efficiency of the model.

By leveraging state-of-the-art big language models and pre-trained language embeddings, this study seeks to improve the efficiency of legal decision-making processes. The development of a smart court decision-making system, inspired by AI, has the potential to assist legal practitioners and enhance the accuracy of court judgment predictions.

The rest of the paper is organized as follows. Section II describes the existing studies related to smartcout and natural language processing applications. Section IV briefly describes the proposed methodology and techniques used in the study. Section V shows the results of the proposed study and Section V-D illustrates the comparison of results with existing studies. Finally, Section VI concludes the findings of the paper.

II. RELATED WORK

In this section, we will review the existing techniques for AI on ECHR and SCOTUS. [17], presents a study on predicting rulings in the European Court of Human Rights using only documented text data. The study formulates a binary classification task, where the input is the textual content extracted from a case, and the target output is the judgment regarding the violation of human rights. By utilizing N-grams and topics to represent textual information, the models achieve an average accuracy of 79%.

In the domain of predicting judicial decisions, [18], conducted a comparative analysis involving different ML algorithms. Through their experiments, they discovered that the SVM model, with an accuracy of 79.5%, outperformed the other models across various settings. The study emphasized

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

the crucial role played by the semantic information extracted from case texts in feature selection for the predictive models.

In this study by [19], machine learning models were developed to predict violations of Articles in the Convention on Human Rights based on judgments from the European Court of Human Rights (ECHR). Textual features from ECHR Judgment documents were used, including N-grams, word embeddings, and paragraph embeddings. Models were constructed using auto-sklearn for 12 Articles, achieving an overall test accuracy of 68.83%.

[1], conduct a study to use big data analysis, statistical analysis and machine learning to analyze texts of court proceedings for automatic prediction of judicial decisions. They achieved an average accuracy of 75% in predicting violations of 9 articles of the European Convention on Human Rights. In another study, Deep Learning and NLP is applied for judgments prediction in the European Court of Human Rights (ECHR) [20]. State-of-the-art NLP techniques and pre-trained/custom trained Word Embedding text representations are used. CNN models achieve an average accuracy of 82%, outperforming SVM models (75%). Specifically, CNN models for four out of nine Articles achieve statistically significant higher accuracy than SVM models. [15], present a novel approach in the field of judicial prediction by constructing a time-evolving random forest classifier to forecast the behavior of the United States Supreme Court in a generalized, out-of-sample context. The model achieves an accuracy of 70.2% at the case outcome level and 71.9% at the justice vote level.

A novel approach in political science by combining two data sets and employing an AdaBoost decision tree regressor was introduce by [21]. Their research demonstrates that their AdaBoosted approach outperforms existing predictive models of Supreme Court outcomes that rely solely on a single data source or simpler modeling strategies. The improved predictive success, with an accuracy of 74%. The author discovered that Word score and Wordfish may produce judicial positions based on word recurrence in the text, enabling him to identify the text's characteristics [22]. There are several studies that use text mining to analyse court case arguments. able to predict decisions from the retrieved data by automatically analysing legal language and identifying arguments [23], [24].

One of the most popular uses for text classification is natural language processing, which automatically classifies email, recognises spans of text, and classifies sentiment. By categorising and identifying the object, CNN has been demonstrated to be significant in the image processing area, but it is also extremely practical to use CNN in NLP domains. Robots are becoming closer and closer to humans thanks to NLP; by recognising human traits via feedback and processing it using a neural network, a robot would get the ability to think for itself. A vector is used to represent the complete text when CNN is used in text processing [25]. A document encoded as a matrix serves as the input for NLP jobs, and a convolutional layer will find any patterns before information is uniformly divided into each level. Convolutional neural networks are more powerful than traditional machine learning since they are comparable to the human brain. CNN has so produced significant advancements in NLP, including sentiment analysis, translation, and text prediction. Since the India Court is concentrated on, since it has the most extensive judicial legislation. We have a difficult problem in integrating this much data and predicting a decision that complies with the Indian constitution's laws and regulations. As a result, we begin our study by applying NLP and CNN methods with all of the available data and investigating how we may steadily increase their accuracy so that in the future 100% accurate judgements can be anticipated.

Akshay Khatri et al. proposed new technique just for the twitter [26]. These embeddings are obtained through unsupervised learning techniques, which enable us to acquire vector representations of words. They opted to employ the GloVe Twitter sentence embeddings to train our models. This choice was motivated by the fact that these embeddings capture the overall meaning of a sentence while occupying a relatively smaller amount of memory. Consequently, each input provided generated a list of size 200. Once they obtained the sentence embeddings, we combined the context and response in a specific manner. The context was placed before the response to ensure appropriate sequencing. It is worth mentioning that the context embeddings were generated independently of the response. By decoupling them in this way, we aimed to prevent the sentiment of the response from influencing the sentiment of the context.

To enhance learning results, these circumstances entail progressively optimising quantities. To address this, researchers use approaches that frame the issue as a generic constrained optimisation problem [27]. The goal is to minimise the objective function of a particular job, such as matrix factorization, while simultaneously taking into account additional constraints that reflect more information.

Researchers utilize the GloVe objective function and incorporate an extra condition in the training process to achieve this objective. This condition aims to enhance the similarity between the embedding vectors of individual words and the weighted average of their immediate "neighbors" within a semantic lexicon [28]. they make sure that related words have comparable embedding vectors by include this extra information while maintaining the distributional representations of the initial embedding job. They tested their method's performance against baseline and cutting-edge word vector models like GloVe [28], Retrofitting (Rf) [29], and Mittens in order to determine how effective it was.

A strategy for recommending laws based on the fusion of information in the area of judicial law is put out by Min Zheng et al. The suggested approach incorporates law rule extraction, law rule suggestion, and BERT training [30]. In order to collect legal knowledge, the legal knowledge extraction layer pulls keywords from the knowledge in the judicial data. Based on the Skip-RNN, the BERT model conducts semantic representation of legal knowledge. Con-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

sequently, it is possible to extract the semantic representation vector. The attention mechanism serves as the foundation for the integration layer for legal rule knowledge. The fusion of case description and legal knowledge features may be accomplished using the legal rule knowledge layer.

Although the BERT concept has been extensively applied in various domains such as text categorization and named entity identification, its potential in the field of legal advice remains largely untapped. To bridge this gap, the LawRec framework proposed in this research leverages the power of BERT, along with the Skip-RNN models. The framework aims to integrate legal knowledge with case descriptions, employing BERT to learn from the case description text and legal knowledge separately [30]. Through this integration, the LawRec framework is able to generate insightful suggestions and recommendations regarding laws and rules for specific instances, thereby offering valuable guidance in the realm of legal advice.

Furthermore, H Zhong et al. [31] investigated the use of big language models, including BERT and ROBERTA, in predicting judicial decisions. Their research demonstrated that these models outperformed traditional machine learning approaches and achieved high accuracy in predicting the outcomes of Italian Supreme Court cases. These findings align with the conclusions drawn in our research, which emphasize the superior performance of BERT and ROBERTA models in court decision prediction.

During the COVID-19 pandemic, online trials have become widespread through the implementation of smart courts. These smart courts utilize internet platforms to transform offline litigation activities into online processes. By conducting trials online, there is a reduction in the need for in-person presence, ensuring the continuity of trials [17]. The Supreme People's Court issued a notice to strengthen and standardize online litigation, creating a comprehensive framework for courts to conduct proceedings through smart courts. Clear regulations have been established for tasks such as online court hearings, electronic service, identity authentication, and material submission. Statistical data during the pandemic period showed a significant increase in online cases, court sessions, mediations, and electronic services, demonstrating the successful adoption of online litigation.

To improve trial efficiency, a collaboration between Zhejiang Higher People's Court, Zhejiang University, and Alibaba Group has resulted in the development of a full-process intelligent trial system (FITS). This system supports the construction of smart courts and has played a crucial role in financial and private lending cases. FITS enables the court procedures to be conducted on a network platform, providing judges with comprehensive case information and assisting in judicial decision-making [32]. The intelligent trial system performs various tasks, including extracting essential information from legal documents, summarizing key points from court debates, verifying evidence, recommending questions to judges, retrieving similar cases from historical data, and generating well-structured judgment documents. Researchers from Zhejiang University and Alibaba Group have conducted extensive research on these judicial tasks, proposing models and techniques such as named entity recognition, graph convolution, legal dispute judgment prediction, and controversy focus-based debate summarization [33], [34].

Another effort has been made by [35] to develop a Roberta-based model for determining the rhetorical function of a phrase for Judgemental. To obtain the text embedding, the authors applied the Roberta model. The Macro-F-scores for the three systems that the authors supplied were 0.468, 0.457, and 0.452, respectively.

III. METHODOLOGY

IV. PROPOSED METHOD

In this section, we delve into the details of the proposed model, its components, and the methodology of the developed architecture. The architecture is a hybrid and ensemble approach that incorporates various language models such as BERT, ALBERT, ROBERTA, and Distilled BERT, along with pre-trained embeddings like GloVe embeddings. The proposed model, known as CDKF (Cross-Domain Neural Knowledge Fusion), leverages these techniques and methods to extract informative features from the text data and fuse the results at the end. These fused features are then fed into a classifier for prediction. The working methodology of the proposed model is illustrated in Figure 1.

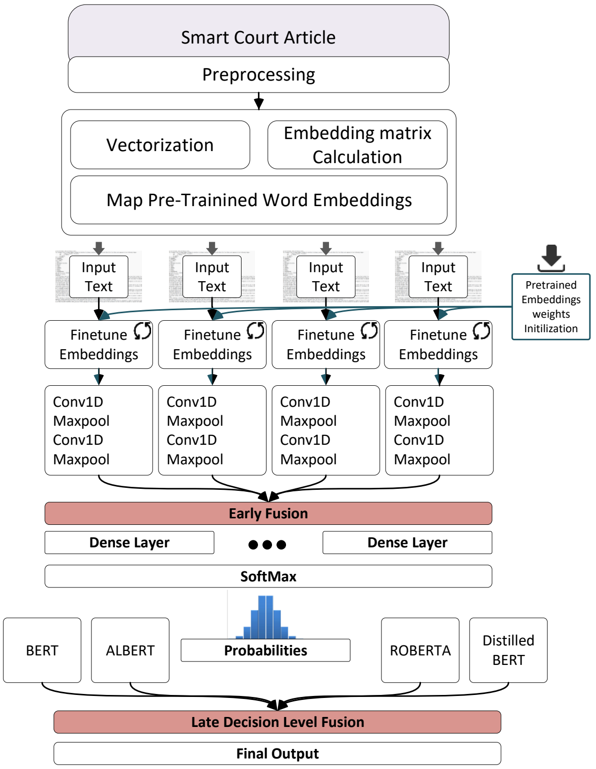


FIGURE 1. Proposed CDKF (Cross-Domain Neural Knowledge Fusion System)

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

The proposed Cross-Domain Neural Knowledge Fusion (CDKF) system is designed to enhance smart court rulings by integrating pre-trained embeddings and advanced transformer models. The architecture follows a hybrid and ensemble-based approach, combining models such as BERT, ALBERT, ROBERTA, and Distilled BERT, along with embeddings like GloVe. The primary aim of CDKF is to extract informative features from text data and fuse these features through multiple channels, ultimately leading to robust predictions.

The process begins with data preprocessing. The input text data $\mathcal{T}$ is tokenized and converted into numerical representations using vectorization. We denote the embedding matrix as Γ, where Γ is initialized using pre-trained embeddings, such as GloVe. For a given input text sequence $\mathcal{T}$, the corresponding embedding $E_{\mathcal{T}}$ is retrieved from Γ. Once the embeddings are generated, the data is split into batches of size 128, followed by a train-test split to ensure a balanced evaluation.

The CDKF architecture processes the text through two distinct phases. In the first phase, pre-trained GloVe embeddings of four different dimensions (50D, 100D, 200D, and 300D) are employed. Let $\Gamma^{(50)}$, $\Gamma^{(100)}$, $\Gamma^{(200)}$, and $\Gamma^{(300)}$ represent the four embedding matrices corresponding to each dimensionality. The text data is passed through these four parallel embedding layers, extracting features via 1D convolutional layers. For each embedding dimension d, the output feature map after applying convolution is represented as:

$$\Phi_d = \text{Conv1D}\left(\Gamma^{(d)}\right)$$

where Φ_d is the feature map for the d-dimensional embedding. Max pooling is then applied to reduce the dimensionality:

$$\Psi_d = \text{MaxPool}\left(\Phi_d\right)$$

The features from all four pathways are concatenated, yielding the fused feature map:

$$\mathcal{F} = \text{Concat}\left(\Psi_{50}, \Psi_{100}, \Psi_{200}, \Psi_{300}\right)$$

Dense and dropout layers are applied to the concatenated features to prevent overfitting, and a binary classifier is employed for the final prediction.

In the second phase, the architecture leverages transformer-based models, denoted by $\mathcal{M}_{\text{BERT}}$, $\mathcal{M}_{\text{ALBERT}}$, $\mathcal{M}_{\text{ROBERTA}}$, and $\mathcal{M}_{\text{DistilBERT}}$. These models process the same text data and generate label predictions $\mathcal{Y}_{\mathcal{M}}$ for each respective model:

$$\mathcal{Y}_{\mathcal{M}} = \mathcal{M}(T)$$

where T is the tokenized text input, and $\mathcal{M}$ represents the respective language model. Once predictions are obtained from both phases, the final label is determined using a voting mechanism, which combines the results from both the embeddings-based classifier and the transformer models. Let $\hat{Y}_{\text{CDKF}}$ be the final predicted label:

$$\hat{Y}_{\text{CDKF}} = \text{Mode}\left(\mathcal{Y}_{\mathcal{M}_1}, \mathcal{Y}_{\mathcal{M}_2}, \mathcal{Y}_{\text{Classifier}}\right)$$

The CDKF system incorporated a combination of static and dynamic embedding strategies, leveraging both GloVe and advanced large language models such as BERT, ALBERT, RoBERTa, and Distilled BERT. This hybrid ensemble approach combines the strengths of traditional embeddings with modern transformer-based architectures to improve the accuracy and robustness of predictions in smart court ruling applications. GloVe embeddings are inherently static, providing a global co-occurrence-based representation of words. In our approach, we initialized the model with GloVe embeddings and retained these embeddings as fixed during the early training stages. This allowed the model to use global semantic knowledge without altering the word vectors. However, as the embeddings are static and do not account for contextual information, we complemented them with dynamic embeddings from large pre-trained language models. For the dynamic embeddings, we employed BERT, ALBERT, RoBERTa, Distilled BERT, where specific layers of the pre-trained models were unfrozen. This allowed the model to adjust the pre-trained embeddings gradually by using backpropagation based on the task-specific data.

To control the extent of changes during fine-tuning, we carefully tuned the learning rate to a low value, which prevented significant deviations from the original embeddings while allowing the model to adapt to the nuances of our dataset. For GloVe embedding's, which are static, we employed embedding's pre-trained on large-scale datasets such as the Common Crawl or Wikipedia + Gig word corpus. For the dynamic embedding's from BERT, ALBERT, RoBERTa, and Distilled BERT, these models were pre-trained on diverse datasets like BookCorpus and English Wikipedia.

Through this combined use of static GloVe embeddings and fine-tuned dynamic embeddings from large language models, we effectively captured both global word associations and context-dependent relationships, resulting in a more powerful model tailored to our task's specific needs. The entire architecture is illustrated in Figure 1 (included in the PDF file), which visually depicts the data flow and integration across various embedding pathways and model architectures.

A. GLOVE EMBEDDINGS

Word embeddings that are specifically made to represent words as vectors are called GloVe Embeddings [28]. By storing the ratio of co-occurrence probabilities as vector differences, these embeddings are able to reflect the statistical link between words. We commence with a straightforward illustration that demonstrates how specific aspects of significance can be derived directly from probabilities of co-occurrence. Observing that the proportion P_{ik}/P_{jk} relies on three words i, j, and k, the most comprehensive model assumes the structure,

$$F\left(w_i, w_j, \widetilde{w}_k\right) = \frac{p_{ik}}{p_{jk}} \tag{1}$$

Although F might potentially entail a complex function defined by parameters such as a neural network, adopting this

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

TABLE 1. Detailed Explanation of the Methodology Steps

Step	Name of Step/Phase	Explanation of Steps	Mathematical Equation	Symbol Explanation
1	Data Preprocessing	Text-to-number conversion using vectorization and embedding initialization	$E_{\mathcal{T}} = \Gamma(\mathcal{T})$	$\mathcal{T}$: text input, Γ: embedding matrix, $E_{\mathcal{T}}$: embedding vector
2	Data Splitting and Batching	Data is split into training and testing sets, and batch size is set to 128 for training	$B = \text{Batch}(\mathcal{T}, 128)$	B: data batches, 128: batch size
3	Embedding Pathways	Text data is passed through four embedding layers with different dimensions	$\Gamma^{(d)} \quad d \in \{50, 100, 200, 300\}$	$\Gamma^{(d)}$: embedding matrix for each dimension
4	Feature Extraction	1D Convolution and Max Pooling are applied to extract features from each pathway	$\Phi_d = \text{Conv1D}(\Gamma^{(d)})\Psi_d = \text{MaxPool}(\Phi_d)$	Φ_d: convolutional output, Ψ_d: pooled features
5	Early Fusion	Features from all Glove pathways are concatenated for further processing	$\mathcal{F} = \text{Concat}(\Psi_{50}, \Psi_{100}, \Psi_{200}, \Psi_{300})$	$\mathcal{F}$: fused feature map
6	Dense and Dropout Layers	Dense and dropout layers are applied to prevent overfitting before classification	$D = \text{Dropout}(\text{Dense}(\mathcal{F}))$	D: output of dropout layer
7	Binary Classifier	A binary classifier is used to generate the output labels for the first phase	$\hat{y}_{\text{classifier}} = \text{Sigmoid}(D)$	$\hat{y}_{\text{classifier}}$: predicted label from the classifier
8	Transformer Models	Text data is passed through BERT, ALBERT, ROBERTA, and Distilled BERT to generate labels	$\mathcal{Y}_{\mathcal{M}} = \mathcal{M}(\mathcal{T})$	$\mathcal{M}$: transformer model, $\mathcal{Y}_{\mathcal{M}}$: predicted label
9	Late Fusion with Transformers	Features from BERT, ALBERT, ROBERTA, and Distilled BERT	$\mathcal{F}_L = \text{Concat}(\Psi, \mathcal{Y}_{\mathcal{M}})$	$\mathcal{F}_L$: late fused features, $\mathcal{Y}_{\mathcal{M}}$: transformer model output
10	Ensemble Voting	Results from the classifier and transformer models are fused using the mode function	$\hat{Y}_{\text{CDKF}} = \text{Mode}(\mathcal{Y}_{\mathcal{M}}, \hat{y}_{\text{classifier}})$	$\hat{Y}_{\text{CDKF}}$: final predicted label, Mode: voting mechanism

approach could obscure the linear framework we intend to capture. To address this concern, we can initially perform the dot product of the arguments.

$$F\left((w_i - w_j)^T . \widetilde{w}_k\right) = \frac{p_{ik}}{p_{jk}} \tag{2}$$

To maintain consistency in this exchange, we should not only swap $w \leftrightarrow w^{\sim}$ but also $X \leftrightarrow X^T$. Our final model should exhibit invariance under this relabeling, even though Equation 3 currently does not.

$$F\left((w_i - w_j)^T . \widetilde{w}_k\right) = \frac{w_i^T . \widetilde{w}_k}{w_j^T . \widetilde{w}_k} \tag{3}$$

To reduce the discrepancy between the logarithm of their co-occurrence counts and dot product of word vectors, GloVe uses a weighted least squares goal function.

$$J = \sum_{i,j=1}^{V} f(X_{i,j})(\omega_i^T \tilde{\omega}_j + b_i + \tilde{b}_j - \log X_{ij})^2 \tag{4}$$

Where ω_i and b_i represent the word vector and bias, respectively, of the word $i, \widetilde{\omega}_j$ and b_j represent the word j context word vector and bias, X_{ij} represents the number of times the word appears in the context, f represents a weighting function that give less weight to infrequent and frequent co-occurrence's. The figure 2 shows a sample Glove Embeddings based LSTM model for smart court decision prediction.

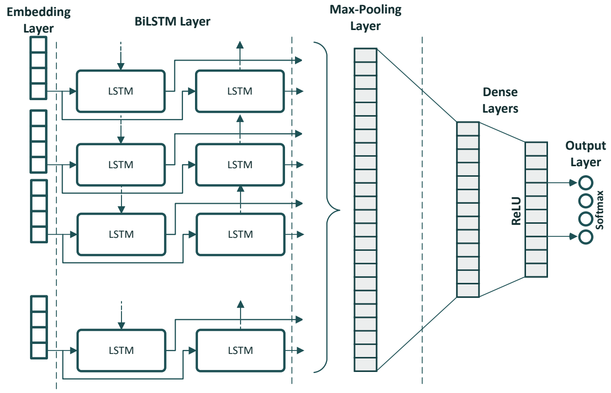


FIGURE 2. Illustration of Sample Glove Embedding Integrated LSTM Models

Unlike methods that focus solely on local context like Word2Vec, GloVe combines both the frequency of word pairs occurring together and the probabilities of their co-occurrences. It constructs a matrix representing word co-occurrences and then factorizes this matrix to generate word embeddings that effectively encode semantic relationships. GloVe embeddings have demonstrated their effectiveness in various NLP tasks due to their ability to capture rich linguistic information.

1) Semantic Information: GloVe embeddings encode semantic relationships.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

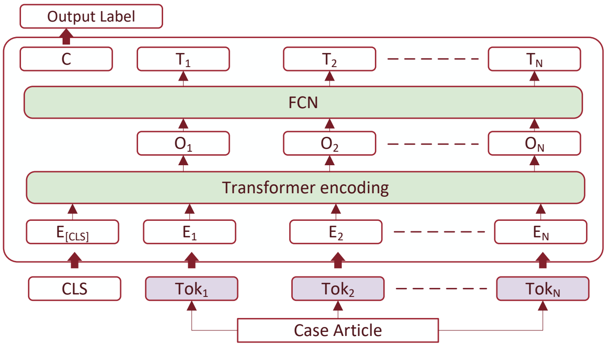


FIGURE 3. BERT architecture for text classification

2) Efficiency: GloVe pre-computes co-occurrence statistics, making training faster and requiring less memory.
3) Contextual: GloVe considers both local and global context, offering a balanced view of word meaning.
4) Limitations: Limited Context: GloVe relies on co-occurrences within a specific context window, missing more distant relationships.
5) Out-of-Vocabulary: Rare words or new terms not in the training corpus may not have accurate embeddings.
6) Fixed Size: GloVe embeddings are of fixed dimensions, which might not capture all nuances of words.

In the phrase "The sun sets behind the mountains," GloVe computes the likelihood of co-occurrence. If "sun" and "sets" frequently occur together across various texts, their vectors become alike, capturing their connection. In essence, GloVe embeddings provide an effective and streamlined method to depict semantic links among words, yet they could miss certain nuanced contextual aspects due to their fixed dimensionality.

B. BERT

Bidirectional Encoder Representation from Transformers (BERT) [9], is an advanced tansformer model that enhances ordinary Transformers by eliminating the uni-directional constraints through the adoption of a masked language model (MLM) pre-training target. The primary objective of the MLM is to accurately predict the original identity of masked words by leveraging the contextual information provided by the surrounding tokens. Unlike traditional left-to-right language models, the MLM approach allows the deep bidirectional Transformer to effectively capture both the left and right contexts within its representations.

In addition to the MLM, BERT also utilizes the next sentence prediction task to enhance its understanding of text pairs during pre-training. By combining the MLM and the next sentence prediction, BERT is able to generate robust representations for text, empowering it with strong text comprehension and representation capabilities. The Figure 3 shows the BERT architecture for text related tasks.

BERT is a popular transformer model for NLP tasks. It comprises several self-attention layers and feedforward neural networks. The fundamental equation governing self-attention in BERT can be depicted as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{5}$$

Here, Q, K, and V are matrices representing queries, keys, and values, respectively. The attention mechanism calculates the relevance between each query and all keys to obtain an attention distribution. The numerator of the attention scores, QK^T, represents the dot product of queries and keys, and the division by $\sqrt{d_k}$ scales the scores to prevent them from becoming too large.

The softmax function is applied element-wise to the scaled scores, converting them into probabilities that sum up to 1. These probabilities are then used to weigh the values, which are linear combinations of the original input embeddings. The result is a weighted sum of values, capturing the contextual information relevant to each query.

In the BERT model, since the transformer architecture does not inherently consider the order of words, positional information needs to be incorporated explicitly. This is achieved using positional encodings:

$$\text{PosEnc}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right) \tag{6}$$

$$\text{PosEnc}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{7}$$

Here, $\text{PosEnc}(pos, 2i)$ and $\text{PosEnc}(pos, 2i + 1)$ are the positional encodings for the even and odd dimensions of the input embeddings. pos represents the position of the word in the input sequence, i denotes the dimension index, and d is the dimensionality of the embeddings.

By adding these positional encodings to the word embeddings, BERT can capture the sequential order of words and incorporate position-based information into its representations, enabling the model to understand the context of words within the input sequence.

Layer normalization is a technique used in BERT to enhanced the training stability and improve convergence. The equation for layer normalization is as follows:

$$\text{LayerNorm}(x) = \frac{x - \mu}{\sigma} \odot \gamma + \beta \tag{8}$$

Here, x represents the input tensor to the layer normalization, μ and σ are the mean and standard deviation calculated across the feature dimensions, γ is a learnable scaling factor, and β is a learnable bias term.

Layer normalization normalizes the activations within a layer, ensuring that they have a consistent mean and variance. This helps mitigate the vanishing gradient problem and accelerates convergence during training. The scaling and bias terms γ and β allow the model to learn the optimal normalization for each feature.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

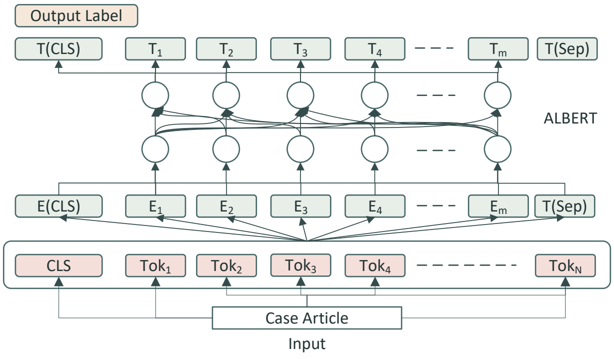


FIGURE 4. Workflow of ALBERT architecture

BERT, a revolutionary transformer-based model in NLP, leverages multi-head self-attention for context understanding. It compensates for the transformer's disregard of word order through positional encodings that capture position-dependent information. Additionally, layer normalization stabilizes training. This approach empowers BERT to grasp nuanced context, making it a pivotal tool for diverse language tasks.

C. ALBERT

ALBERT is a Transformer design that enhances BERT while lowering the number of parameters greatly [36]. It uses two ways to reduce the parameters. The first method is parameterized factorised embeddings. The expansive vocabulary embedding matrix is divided into two more manageable matrices using ALBERT. The size of the hidden layers might be different from the size of the vocabulary embedding because to this division. As a result, it becomes simpler to raise the concealed size without noticeably growing the vocabulary embeddings' parameter size. This aids in lowering the model's total parameter count. Sharing parameters between layers is ALBERT's second method. This method makes sure that the parameters don't increase along with the network's depth. ALBERT significantly lowers the amount of parameters needed in the model by sharing parameters between layers. The visual working flow of the ALBERT model is depicted in Figure 4.

ALBERT, which stands for A Lite BERT, is a special case of the BERT paradigm developed to overcome the efficiency and scalability issues of BERT. Although BERT has excelled in natural language processing tasks, its scale makes it difficult to use in resource-constrained settings or situations requiring real-time predictions.

To make the BERT architecture more lightweight and economical without compromising performance, ALBERT includes a number of significant changes. The adoption of parameter-sharing methods, notably cross-layer parameter sharing and embedding parameterization, is one of the main innovations in ALBERT [37]. These methods considerably lower the model's memory footprint and processing needs, improving its suitability for real-world use.

ALBERT introduces a novel training objective known as Sentence-Order Prediction (SOP), which replaces BERT's Next Sentence Prediction (NSP) task. The equation utilized to compute the SOP loss is as follows:

$$\mathcal{L}_{\text{SOP}} = -\sum_{i=1}^{N} \log \text{softmax}(W_{\text{SOP}} \cdot H_i) \tag{9}$$

Here, N is the number of sentence pairs, H_i represents the hidden representation of the i-th sentence in the pair, and W_{SOP} is the SOP task-specific weight matrix. The softmax function normalizes the predictions across possible sentence orders. The SOP objective encourages ALBERT to predict the correct order of sentences within a pair, helping the model capture more coherent context representations and improving the utilization of training data.

ALBERT uses layer parameter sharing for efficiency and effectiveness. The formula used to calculate a shared parameter matrix is as follows:

$$E^{\ell}_{\text{shared}} = \frac{1}{L}\sum_{k=1}^{L} E^{\ell}_{k} \tag{10}$$

Here, E^{ℓ}_{k} represents the embedding matrix of the k-th layer at depth ℓ, and L is the total number of layers. Cross-layer parameter sharing involves averaging the embeddings of all layers within the same depth ℓ. This strategy reduces the number of parameters and enables ALBERT to learn more efficiently, making it an effective solution for resource constraints.

ALBERT utilizes the Masked Inter-Sentence Objective (MISO) approach to enhance its modeling of relationships between sentences. The formula that computes the MISO loss is as follows:

$$\mathcal{L}_{\text{MISO}} = -\sum_{i=1}^{N} \log \text{softmax}(W_{\text{MISO}} \cdot H_i) \tag{11}$$

Here, the notation is similar to the SOP loss equation. W_{MISO} is the MISO-specific weight matrix, and H_i represents the hidden representation of the i-th sentence pair.

The MISO objective guides ALBERT to predict whether a sentence pair is consecutive or non-consecutive, enhancing the model's ability to understand and capture complex relationships between sentences.

Because of its effectiveness and scalability, the ALBERT model is a good fit for applications in the field of smart court prediction. The smart court prediction system can efficiently analyse and handle massive amounts of legal documents, including court cases, rulings, and precedents, by utilising the power of ALBERT.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

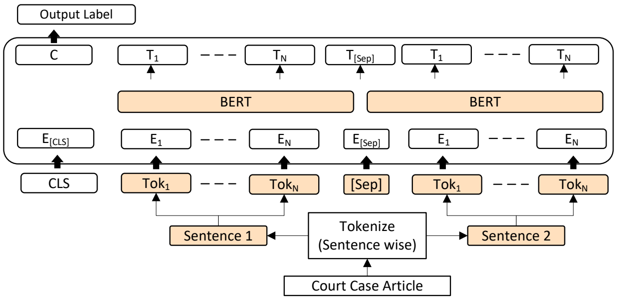


FIGURE 5. ROBERTA Architecture (enhanced variant of BERT)

D. ROBERTA

RoBERTa is a development of BERT that modifies the pretraining process [38]. The changes consist of:

- No NSP objective
- Training on longer sequences
- Training with larger batches.
- To more effectively account for training set size effects, the authors additionally gather a sizable new dataset that is equivalent in size to existing used datasets.

Unlike BERT, RoBERTa employs dynamic masking and omits the Next Sentence Prediction (NSP) task. In RoBERTa, dynamic masking is utilized during pretraining. The equation 12 defines the probability distribution $P_{\text{dynamic}}(t)$ for selecting tokens to be masked during training. If a token is already masked, its probability of being chosen is 0.1. For unmasked tokens, their probability is calculated as 0.9 divided by the sequence length minus 1. Additionally, RoBERTa differs from ALBERT by not using parameter sharing, leading to a more parameter-intensive model. The core equations (12) underlying RoBERTa's architecture are as follows:

$$P_{\text{dynamic}}(t) = \begin{cases} 0.1 & \text{if token is masked} \\ 0.9/(n-1) & \text{otherwise} \end{cases} \tag{12}$$

Here, $P_{\text{dynamic}}(t)$ represents the probability of each token being selected for masking, and n is the sequence length. The figure 5 shows the visual mechanism of ROBERTA architecture.

In RoBERTa, the "Pretraining" phase involves training the model on a large text corpus. During this stage, RoBERTa learns to predict masked words within sentences, enhancing its ability to understand context. By exposing the model to diverse language patterns, structures, and relationships, the Pretraining phase equips RoBERTa with a foundational understanding of language, which is further fine-tuned for specific tasks in subsequent stages of training.

$$\mathcal{L}_{\text{pretrain}} = -\sum_{i=1}^{N}\sum_{j=1}^{L} \log P_{\text{dynamic}}(t_{ij}) \cdot \text{CE}(t_{ij}, t_{ij}^{*}) \tag{13}$$

This equation 13 represents the pretraining loss $\mathcal{L}pretrain$ for RoBERTa. It sums over all sentences (N) and all tokens (L). For each token, it calculates the negative log likelihood of predicting the masked token (tij) according to the dynamic masking probability and compares it to the true token (t_{ij}^{*}) using cross-entropy loss (CE). This loss function is important to learn efficient feature vectors.

A new type of loss function is used in RoBERTa called Masked Language Model (MLM) Loss. In MLM, some input text is randomly masked, and the model's task is to predict the masked text with the help of overall context of text.

$$\mathcal{L}_{\text{MLM}} = -\sum_{i=1}^{N}\sum_{j=1}^{L} \log P_{\text{MLM}}(t_{ij}|t_{ij}^{*}) \tag{14}$$

In equation 14, $\mathcal{L}_{\text{MLM}}$ represents the MLM loss used in RoBERTa. It sums over all tokens, similar to the pretraining loss. This loss helps RoBERTa capture contextual information and fill in masked tokens.

The Sentence Order Prediction (SOP) Loss in RoBERTa refers to a training objective that helps the model understand the sequential order of sentence pairs. By optimizing this loss, RoBERTa learns to accurately predict whether sentence pairs are in the correct order. This contributes to RoBERTa's ability to capture meaningful relationships and coherence between sentences, making it more effective in various language tasks.

$$\mathcal{L}_{\text{SOP}} = -\sum_{i=1}^{N} \log \text{softmax}(W_{\text{SOP}} \cdot H_i) \tag{15}$$

This equation 15 represents the Sentence Order Prediction (SOP) loss $\mathcal{L}SOP$ for RoBERTa. It sums over all sentence pairs (N) and calculates the negative log likelihood of predicting the correct sentence order using the softmax of the weighted hidden representation (H_i) and the SOP-specific weight matrix (WSOP). Unlike BERT, RoBERTa uses this loss to ensure coherent relationships between sentence pairs. RoBERTa boasts improved performance through dynamic masking and a larger training corpus, yielding nuanced language understanding. However, its resource-intensive nature demands substantial computational resources.

E. DISTILBERT

DistilBERT is a compressed version of the BERT model that aims to preserve performance while providing a smaller, quicker, and lighter alternative [39]. Hugging Face created it in 2019, and since then, it has grown in popularity among software programmes that use natural language processing. In order to make BERT acceptable for resource-constrained situations like mobile and edge devices, DistilBERT's main goal is to decrease the model's parameters and training time while maintaining its efficacy. The figure 6 shows the fully functional working of Distilled BERT architecture.

DistilBERT achieves a large parameter reduction of about 66% when compared to the basic model BERT-base. The

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

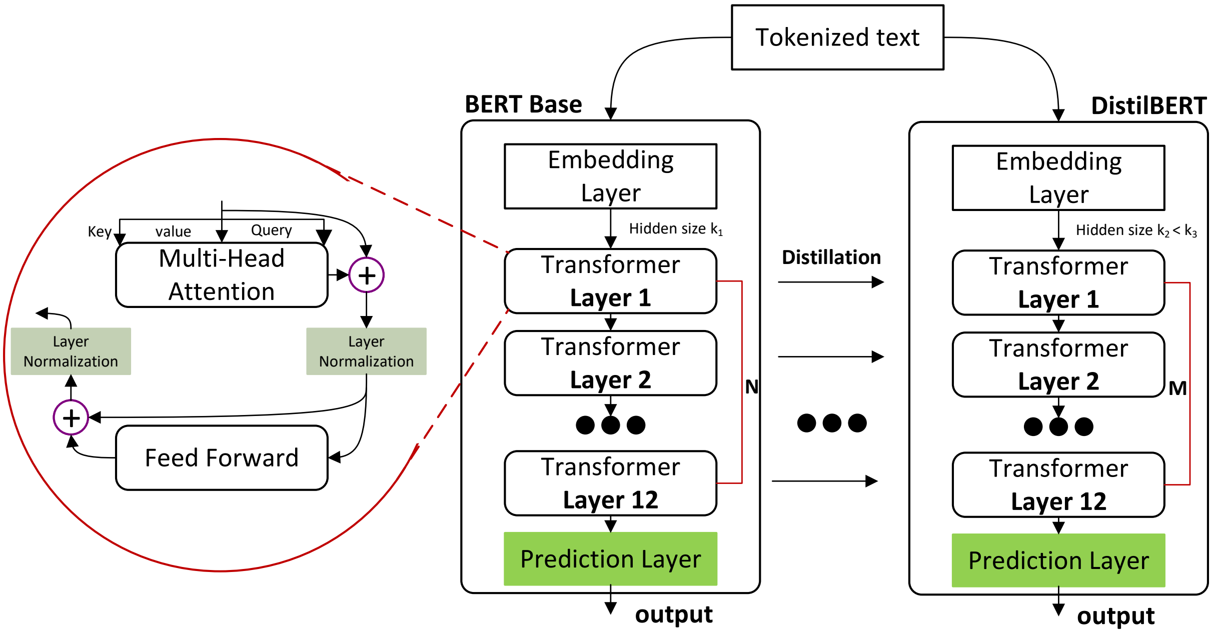


FIGURE 6. Wokflow mechanism of Distilled BERT

model's lower size and higher efficiency are both a result of the parameters' decrease. Although it has fewer parameters than BERT, DistilBERT aspires to perform at a level that is comparable to BERT.

DistilBERT employs knowledge distillation to compress the knowledge from a larger "teacher" model (such as BERT) into a smaller "student" model, while optimizing task-specific performance. The equation 16 represents the distillation loss, where CE is the cross-entropy loss. Equation 17 defines the teacher and student representations Z_i and H_i for each input x_i. The final equation 18 combines the task-specific loss with the distillation loss, where α balances their contribution.

$$\mathcal{L}_{\text{distill}} = \sum_{i=1}^{N} \text{CE}(\text{softmax}(Z_i/T), \text{softmax}(H_i/T)) \quad (16)$$

$$Z_i = \text{teacher}(x_i) \quad \text{and} \quad H_i = \text{student}(x_i) \quad (17)$$

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{task}} + \alpha \cdot \mathcal{L}_{\text{distill}} \quad (18)$$

DistilBERT uses the concept of knowledge distillation, which involves transferring knowledge from a large pre-trained model like BERT into a more compact model. This results in a smaller and faster model suitable for deployment on resource-constrained devices. Unlike ALBERT, DistilBERT doesn't modify the model architecture or use parameter sharing. It differs from RoBERTa by focusing on model compression rather than larger-scale training.

In our study, we opted for transformer-based models from the BERT family over larger language models like LLaMA and Claude due to several reasons. First, the task at hand—semantic classification of legal texts—does not necessitate the extensive text generation or reasoning capabilities of models like LLaMA. Instead, our objective is to capture the semantic nuances of legal language, which BERT-based models are specifically designed to do efficiently. Second, models like LLaMA, while state-of-the-art, are computationally expensive and require large-scale infrastructure for training and inference, making them impractical for real-time or large-scale applications in resource-constrained environments. Moreover, our architecture employs a multi-modal approach by incorporating different pre-trained embeddings (GloVe) and transformer models (BERT, ALBERT, ROBERTA, Distilled BERT) to enhance the model's ability to capture both syntactic and semantic nuances from the text. This multimodal strategy enables us to extract richer features from the legal texts without the need for excessively large models. Thus, we believe our approach offers a more pragmatic solution that balances performance, interpretability, and computational feasibility.

V. RESULTS AND DISCUSSION

In this section, we conduct a comparative analysis of the techniques utilized in this research. We analyze the effect of different dimensional pre-trained embeddings on the overall results, as well as compare the performance of various large transformer models. The experimental results presented in tables allow us to analyze the performance of each model

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

in terms of different classes. Furthermore, the following sections provide detailed information about the dataset used, a thorough comparative analysis of the pre-trained embeddings, a comparison of transformer models, and a comparison with state-of-the-art models. Table 2 presents the details of the parameter configurations and other essential information regarding the techniques employed in the training process.

Table 3 displays the text sequence lengths used in the training and testing of transformer models such as BERT, ALBERT, ROBERTA, and Distilled BERT. Since the data processing varies between the embedding phase and the transformer phase, the feature vector lengths differ for each phase, as indicated in Table 2 and 3.

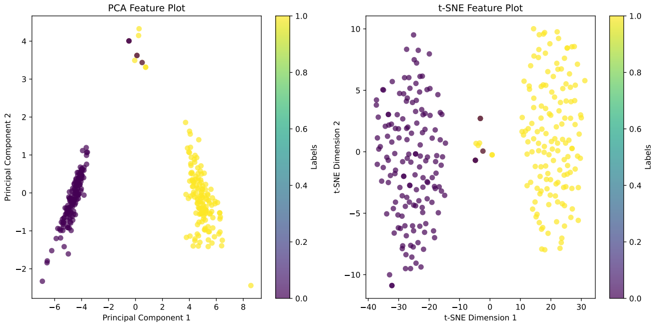


FIGURE 7. Illustrating the PCA and TSNE-transformed features obtained from the final layers of the model (Violation vs Non-Violation)

A. DATASET

We make use of the ECtHR's publicly available data [14]. The European Convention on Human Rights (ECHR) created the ECtHR as an international court in 1959. Its duty is to decide on complaints made by people or independent nations that the Convention's civil and political rights have been violated. In order to protect human rights in European democracies that maintain the ideals of the rule of law, the ECHR serves as an international accord. The Council of Europe Treaty, which was first drafted in 1950 by the ten founding member states, now includes 47 members, with a combined population of around 800 million people. Being a party to the Convention is a requirement for joining the Council of Europe, and all new members are required to ratify the ECHR as quickly as feasible. In its own right, the Convention went into force in 1953. The Court has been in operation as a permanent judicial body since 1998, and individuals have the right to approach it directly if they can show that they have used all domestic legal systems' available legal remedies to address their human rights complaints before national courts.

A typical judicial decision rendered by the ECtHR consists of several key components, including:

1) Introduction: This section provides the title of the case (e.g., Lawless vs. Ireland), the date of the decision, the Chamber responsible for the case, and the composition of the Court, including the judges, president, and registrar involved.
2) Procedure: The procedure section outlines the sequence of events from the initial lodging of the application to the final judgment by the Court. It covers the various stages and processes involved in the case.
3) Facts (Circumstances): This part provides relevant background information about the applicant and the events or circumstances that prompted them to seek justice, claiming violations of their rights under the European Convention on Human Rights (ECHR).
4) Facts (Relevant Law): This part includes references to legal provisions from documents other than the ECHR. Typically, these provisions encompass domestic laws, European treaties, and international agreements that are pertinent to the case.
5) Law: In this section, the Court presents its legal arguments. Each alleged violation is discussed separately, and the Court provides its reasoning and analysis for each claim.
6) Judgment: The judgment section contains the Court's decision regarding each alleged violation. It states whether the Court finds a violation of the rights protected by the ECHR or if the application is dismissed.
7) Dissenting or Concurring Opinions: This part includes additional opinions expressed by individual judges. Concurring opinions explain why judges support the majority decision, while dissenting opinions outline the reasons why judges disagree with the majority's ruling.

It's important to note that the structure and content of a judicial decision may vary depending on the specifics of the case and the preferences of the Court. Table 4 illustrates the overall data distribution.

In this section we perform comparative analysis on smart court decision data. We use different pre-trained embeddings and deep transformer models. We experiment with different dimensional embeddings for example 100D and 200D. We also compare proposed model with previous state of the art presented models.

B. VISUALIZATION OF EXTRACTED FEATURES

After training the models, we extract features from them by accessing the second-to-last layer in the model architecture. These features are represented in a 128-dimensional space, and to visualize them effectively, we employ both the TSNE and PCA techniques. Figure 7 illustrates the visualization of the reduced-dimensional features.

Following the visualization of PCA and TSNE-transformed features, we further employed a heatmap to depict the inter-feature correlations. The heatmap, displayed in Figure 8, offers insight into the correlation structure among all the features.

C. COMPARATIVE ANALYSIS

The objective of our work with respect to this topic is to carefully analyze a variety of court articles in order to provide

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

TABLE 2. Overview of the Hyper-Parameter Configuration

CDKF System Configuration	
Utilized Techniques	BERT, ALBERT, ROBERTA, Distilled BERT and Pre-trained embeddings
Phase -1: Trainable Parameters of Pre-trained embeddings network	299794
Non-Trainable Parameters of Pre-trained embeddings network	13001300
No of epochs of Pre-trained embeddings network	100
No of epochs for other transformer models	10
Learning rate for embeddings models	Deafult
Learning rate for transformer models	1e-5

TABLE 3. Preprocessed Input Sequence Length for Each Transformer Model

Transformers	**Train Sequence Mean**	**Train sequence 95 percentile**	**Train sequence 99 percentile**	**Train Sequence Mean**	**Train sequence 95 percentile**	**Train sequence 99 percentile**
BERT	11034	25743	45537	11199	30323	48620
ALBERT	11034	25743	45537	11199	30323	48620
ROBERTA	11034	25743	45537	11199	30323	48620
Distilled BERT	11034	25743	45537	11199	30323	48620

TABLE 4. Dataset Distribution

		Training Dataset		**Total**	
Article	**Title**	**Violation cases**	**Non-Violation cases**	**Training Data**	**Testing Data**
2	Right to life	57	57	114	398
3	Prohibitito torture	284	284	568	851
5	Right to Liberty and security	150	150	300	1118
8	Right to respect for private and family life	229	229	458	496
10	Freedom of expression	106	106	212	252
13	Right to an effective remedy	106	106	212	1060
14	Prohibition to discrimination	144	144	288	44

a thorough comparison study. That was already indicated, the aim is to predict human rights court verdicts using a variety of artificial intelligence techniques. In particular, in section I of our research, we compare various pre-trained embeddings applied for the purpose of predicting court decisions. These embeddings have been thoroughly labelled as "violation" or "non-violation," allowing us to evaluate how well they predict the results of human rights-related court cases. We evaluate memory-based deep transformers for smart court decision prediction in section II. In part III, we conduct a comparative analysis of cutting-edge architectures that have been cited in the literature.

1) Comparison Between Pre-trained Embeddings on Each Articles

This section's main objective is to evaluate the performance of several pre-trained embeddings on a per-article basis. We want to evaluate their effectiveness and examine whether they are appropriate for the purpose of our research. It is important to note that some articles might not have enough data to properly train our models. To ensure the validity and accuracy of our conclusions, we thus exclude these specific articles from our examination.

When considering the classification of violation and non-violation, we are able to clearly observe distinct variations in the performance of different embeddings across various articles when looking at the Table 5 above. Using Glove 200D embedding, we achieve a highest accuracy of 88.10% for Article 13. The highest accuracy levels, on the other hand, are 83.33% (with Glove 300D), 78.53% (with Glove 100D), 72.88% (with Glove 300D), 72.03% (with Glove 200D), 71.43% (with Glove 50D), and 70.00% (with Glove 200D) for Articles 13, 13, 3, 3, 13 and 5.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

TABLE 5. Performance (%) outcomes after training pretrained embedding on individual articles

Method	Measure	Article2	Article3	Article5	Article8	Article10	Article13	Article14
4***Glove 50D**	Accuracy	56.25	64.64	70.00	62.07	52.38	71.43	58.33
	Recall	53.12	71.19	75.00	58.62	57.14	78.57	56.25
	Precision	54.84	68.85	63.38	61.82	53.33	73.33	65.85
	F1-Score	53.77	70.00	68.70	60.18	55.17	75.86	60.67
4***Glove 100D**	Accuracy	68.75	60.17	66.67	62.01	57.14	78.53	83.33
	Recall	78.12	57.63	65.00	62.07	66.67	83.33	83.33
	Precision	65.79	60.71	67.24	61.01	58.33	71.43	80.00
	F1-Score	71.43	59.13	66.10	61.54	62.22	76.92	83.63
4***Glove 200D**	Accuracy	59.38	72.03	70.00	53.45	59.52	88.10	75.00
	Recall	50.00	63.56	80.00	37.93	40.48	88.10	85.42
	Precision	57.14	72.82	64.86	56.41	60.71	86.05	63.08
	F1-Score	53.33	67.87	71.64	45.36	48.57	87.06	72.57
4***Glove 300D**	Accuracy	59.38	72.88	66.67	58.62	57.14	83.33	60.42
	Recall	40.62	72.88	58.33	70.69	64.29	95.24	72.92
	Precision	61.90	73.50	67.31	52.56	56.25	76.92	58.33
	F1-Score	49.06	73.19	62.50	60.29	60.00	85.11	64.81

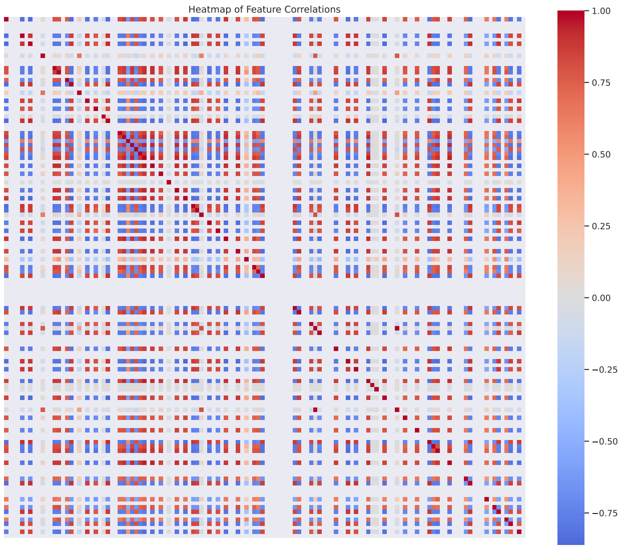


FIGURE 8. Illustrating Heatmap of features obtained from the final layers of the model (Violation vs Non-Violation)

2) Comparison Between Transformers on Each Articles

After looking at pre-trained embeddings for classifying violations in the articles, we now turn our focus towards evaluating the effectiveness of various pre-trained transformer models in the context of smart court decision classification. Four unique transformer variants are utilized into account in this analysis: BERT, ALBERT, Distilled BERT, and ROBERTA. Table 6 provides a comprehensive performance analysis of transformers for individual articles.

Examining the provided table indicates several patterns and trends regarding the effectiveness of various transformer model types on various articles. Notably, when used for expressing article 13, the Distilled BERT model shows amazing accuracy of 88%. Additionally, Distilled BERT performs well in regard to Articles 2 and 3. While ROBERTA achieves the maximum accuracy of 66% for article 8, ALBERT seems to be very effective in the case of article 5. It is notable that Distilled BERT again shows impressive results for articles 10 and 14, reaching accuracy rates of 62% and 77%, respectively.

Tables 5 and 6 show the results of each approach on specific articles. On the other hand, in table 7 below, we compare the effectiveness of embeddings and transformers across all articles as a whole. In order to distinguish between violations and non-violations, it is necessary to aggregate all the articles and perform a decision classification of court. By looking at this combined study, we may learn more about how well various algorithms handle the classification task over the full dataset.

It is clear from the presented table that the Transformer base models outperform the pre-trained embeddings in terms of performance. When comparing the Transformer base models to the pre-trained embeddings, it is clear from the presented table that the Transformer base models perform better. In particular, using the BERT model, we achieve the maximum accuracy of 82.14%. The performance of the proposed CDKF system is consistently strong and stable across various performance measures.

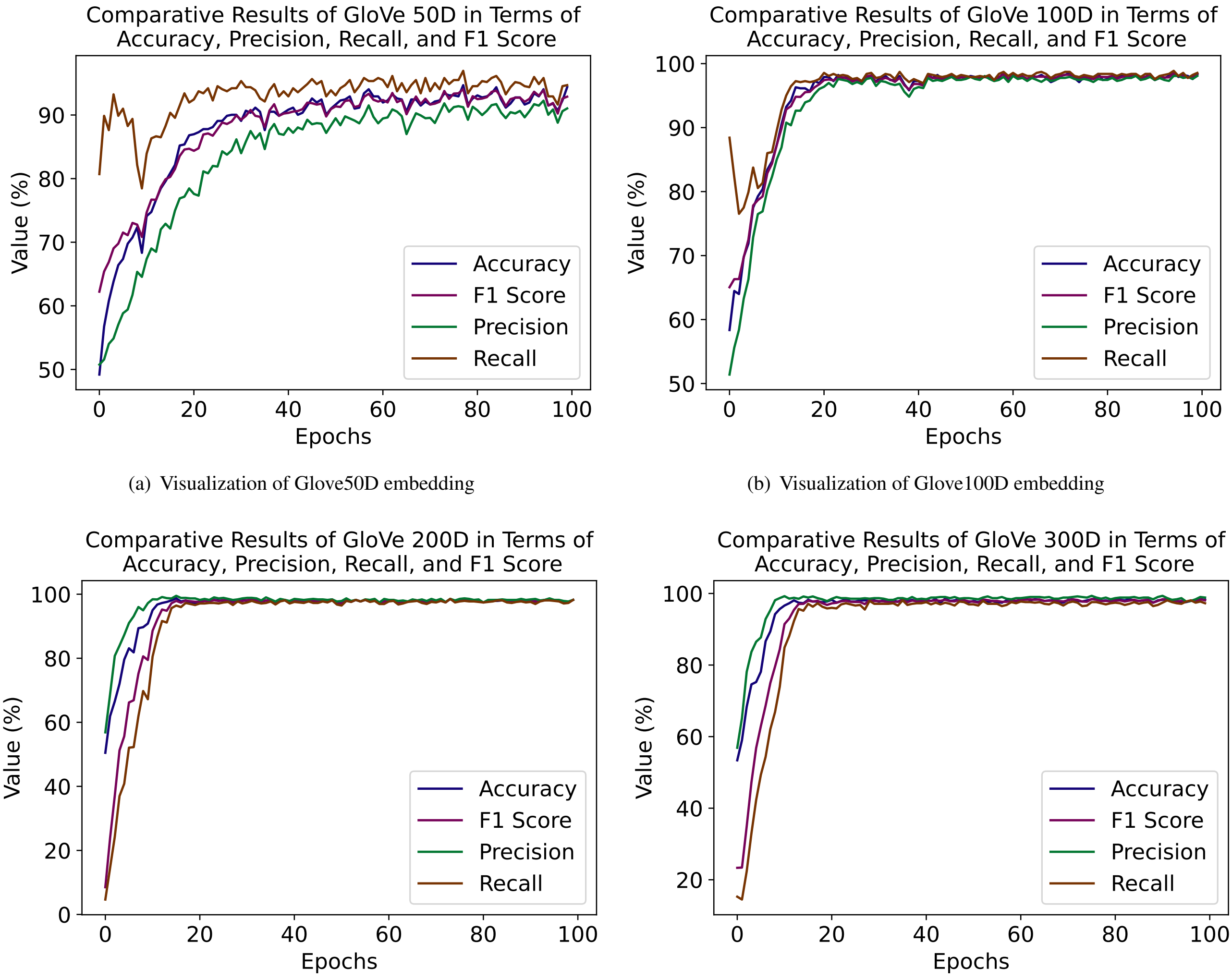
D. COMPARISON WITH STATE OF THE ART MODELS ON EACH ARTICLES

We perform a comparative analysis in this section, comparing our suggested CDKF and BERT model which achieves the best accuracy—to cutting-edge architectural designs that may be found in published literature. A comprehensive evaluation of the performance of several models on the dataset for smart court classification is provided in Table 9.

It is clear from the conclusions drawn from the Table 7 that our proposed CDKF and BERT model for smart court decisions has performed exceptionally well. Table 8 presents the overall performance of the proposed CDKF model in terms of violation and non-violation cases. Additionally, in

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775



(a) Visualization of Glove50D embedding

(b) Visualization of Glove100D embedding

(c) Visualization of Glove200D embedding

(d) Visualization of Glove300D embedding

FIGURE 9. Visualization the performance of all embedding across all articles

TABLE 6. Comparing Transformer Models: Analyzing Each Individual Article

Method	Measure	Article2	Article3	Article5	Article8	Article10	Article13	Article14
4***BERT**	Accuracy	62	70	67	57	45	74	65
	Recall	63	69	67	61	45	74	65
	Precision	62	73	67	61	43	74	65
	F1-Score	62	69	67	57	41	74	65
4***ALBERT**	Accuracy	50	73	63	59	60	76	62
	Recall	50	73	63	57	60	76	64
	Precision	50	73	63	58	60	76	63
	F1-Score	49	73	63	57	60	76	62
4***Distilled BERT**	Accuracy	67	70	62	59	62	88	77
	Recall	69	70	62	59	62	88	77
	Precision	69	70	62	59	62	88	77
	F1-Score	67	70	62	59	62	88	77
4***ROBERTA**	Accuracy	50	69	58	66	36	74	60
	Recall	56	70	61	65	50	73	61
	Precision	56	70	61	66	18	75	61
	F1-Score	50	69	58	65	26	73	60

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

TABLE 7. Comparative Analysis: Transformers vs Pre-trained Embeddings across All Articles

	Pre-Trained Embeddings				Transformers				
Techniques	Glove 50D	Glove 100D	Glove 200D	Glove 300D	BERT	ALBERT	ROBERTA	Distilled BERT	Proposed CDKF
Accuracy	64.96	60.95	67.52	65.33	82.14	75.00	78.57	71.43	76.28

TABLE 8. Comprehensive Performance Analysis of the Proposed CDKF Models

Proposed Model	Accuracy	Recall	Precision	F1-Score
CDKF	76.28	81.02	74.00	77.35
Classification Report				
Violation	76.28	72.00	79.00	75.00
Non-Violation	76.28	81.00	74.00	77.00

Table 8, we conduct a thorough class-wise analysis of the CDKF model's performance, providing a deeper insight into its efficiency.

Legal texts are often contextually rich, and small linguistic differences can significantly alter outcomes. Therefore, we employed a hybrid model that integrates multiple embeddings and transformer-based architectures (BERT, ALBERT, ROBERTA, and Distilled BERT). These diverse techniques help capture different aspects of the legal language, allowing for a more robust representation of the input text. By introducing early and late fusion mechanisms, we ensure that features from both pre-trained embeddings and transformers are utilized to their full potential. This approach allows us to compare the performance of various models, yielding a more accurate and comprehensive prediction of court rulings. Our architecture is specifically designed to address the inherent complexity of legal language and improve model interpretability and performance.

VI. CONCLUSION

Artificial intelligence has found use in many areas of life today, including the legal sector. In this study, we want to predicted court decisions using machine learning strategies. The dataset consists of many articles on human rights, classifying the rulings as violations or non-violations. In this study, we introduce a novel CDKF system designed for predicting court rulings. The system incorporates advanced pre-trained embeddings and state-of-the-art transformer models. It follows a hybrid and ensemble approach, leveraging multiple techniques from different domains to enhance the accuracy of court ruling predictions. By integrating these diverse methods, the CDKF system offers a comprehensive and robust framework for decision-making in legal contexts.. To find the most efficient method, we trained a variety of transformers and pre-trained embeddings and conducted comprehensive analysis and comparison. Our findings show that CDKF, BERT and ROBERTA, with accuracy rates of 76.28, 82.14% and 78.57%, respectively, had the highest performance. Although the proposed model exhibits slightly lower accuracy, it demonstrates higher confidence due to its multiple decision-makings and ensemble approach. On the other hand, the performance of BERT and ROBERTA models may fluctuate depending on the data. Nonetheless, this work holds significant value for future researchers in the smart court domain, providing a foundation for further exploration and study.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

CONTRIBUTIONS

Nagwan Abdel Samee: Writing – original draft, Supervision, Visualization, Formal Analysis, Editing and Review, **Maali Alabdulhafith**: Writing – original draft, Supervision, Visualization, Formal Analysis, Editing and Review, **Syed Muhammad Ahmed Hassan Shah**: Writing – original draft, Conceptualization, Investigation, Methodology, Implementation, Experimentation, Visualization and Validation, **Atif Rizwan**: Supervision, Conceptualization, Visualization, Formal Analysis, Validation, Writing, Review & Editing,

REFERENCES

[1] Masha Medvedeva, Michel Vols, and Martijn Wieling. Using machine learning to predict decisions of the european court of human rights. Artificial Intelligence and Law, 28:237–266, 2020.

[2] David Okore Ukwen and Murat Karabatak. Review of nlp-based systems in digital forensics and cybersecurity. In 2021 9th International symposium on digital forensics and security (ISDFS), pages 1–9. IEEE, 2021.

[3] Dongming Sun, Xiaolu Zhang, Kim-Kwang Raymond Choo, Liang Hu, and Feng Wang. Nlp-based digital forensic investigation platform for online communications. computers & security, 104:102210, 2021.

[4] Neil Shah, Nandish Bhagat, and Manan Shah. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. Visual Computing for Industry, Biomedicine, and Art, 4:1–14, 2021.

[5] Flora Amato, Aniello Castiglione, Giovanni Cozzolino, and Fabio Narducci. A semantic-based methodology for digital forensics analysis. Journal of Parallel and Distributed Computing, 138:172–177, 2020.

[6] Flora Amato, Giovanni Cozzolino, Vincenzo Moscato, and Francesco Moscato. Analyse digital forensic evidences through a semantic-based methodology and nlp techniques. Future Generation Computer Systems, 98:297–307, 2019.

[7] Maxime Bérubé, Thuc-Uyên Tang, Francis Fortin, Sefa Ozalp, Matthew L Williams, and Pete Burnap. Social media forensics applied to assessment of post–critical incident social reaction: The case of the 2017 manchester arena terrorist attack. Forensic science international, 313:110364, 2020.

[8] Zeinab Shahbazi and Yung-Cheol Byun. Nlp-based digital forensic analysis for online social network based on system security. International Journal of Environmental Research and Public Health, 19(12):7027, 2022.

[9] J Atiyah. Image forensic and analytics using machine learning. International Journal of Computing and Business Research, 12:69–93, 2022.

[10] Haydn D Kelly. Forensic gait analysis. CRC Press, 2020.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

TABLE 9. Comparative Analysis with Existing Models in the Literature

Author	Court	Target Variable	Model	Year	Accuracy
[17]	ECHR	Violation, non-violation	SVM	2016	79
[18]	ECHR	Violation, non-violation	SVM	2017	79.5
[19]	ECHR	Violation, non-violation	auto-Sklearn	2019	68.83
[1]	ECHR	Violation, non-violation	SVM	2019	75
[20]	ECHR	Violation, non-violation	CNN	2019	82
[15]	SCOTUS	Affirmed, Reversed	Random Forest	2017	70
[21]	SCOTUS	Affirmed, Reversed	AdaBoost Regressor	2017	74
Proposed	ECHR	Violation, non-violation	BERT	2023	82.14
Proposed	ECHR	Violation, non-violation	ROBERTA	2023	78.57
Proposed	ECHR	Violation, non-violation	CDKF	2023	76.28

[11] Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. N-gram: New groningen author-profiling model. arXiv preprint arXiv:1707.03764, 2017.

[12] Mart Busger op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva, Chris Pool, Johannes Bjerva, Hessel Haagsma, and Malvina Nissim. Gronup: Groningen user profiling. Notebook for PAN at CLEF, 2016.

[13] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting personality from twitter. In 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing, pages 149–156. IEEE, 2011.

[14] Alexandre Quemy and Robert Wrembel. Echr-od: On building an integrated open repository of legal documents for machine learning applications. Information Systems, page 101822, 2021.

[15] Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. A general approach for predicting the behavior of the supreme court of the united states. PloS one, 12(4):e0174698, 2017.

[16] Mónika Márton. Romania: derogation from the european convention on human rights–freedom of expression during public emergency. Toruńskie Studia Polsko-Włoskie, pages 39–54, 2021.

[17] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. PeerJ computer science, 2:e93, 2016.

[18] Zhenyu Liu and Huanhuan Chen. A predictive performance comparison of machine learning models for judicial cases. In 2017 IEEE Symposium series on computational intelligence (SSCI), pages 1–6. IEEE, 2017.

[19] Conor O'Sullivan and Joeran Beel. Predicting the outcome of judicial decisions made by the european court of human rights. arXiv preprint arXiv:1912.10819, 2019.

[20] Arshdeep Kaur and Bojan Bozic. Convolutional neural network-based automatic prediction of judgments of the european court of human rights. In AICS, pages 458–469, 2019.

[21] Aaron Kaufman, Peter Kraft, and Maya Sen. Machine learning, text data, and supreme court forecasting. Project Report, Harvard University, 2017.

[22] Word Fish. BWorld Robot Control Software. http://www.wordfish.org/. [Online; accessed 12-April-2023].

[23] Raquel Mochales and Marie-Francine Moens. Study on the structure of argumentation in case law. In Proceedings of the 2008 conference on legal knowledge and information systems, pages 11–20, 2008.

[24] G Veena, Deepa Gupta, Akshay Anil, and S Akhil. An ontology driven question answering system for legal documents. In 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), volume 1, pages 947–951. IEEE, 2019.

[25] S Sandhiya and U Palani. An effective disease prediction system using incremental feature selection and temporal convolutional neural network. Journal of Ambient Intelligence and Humanized Computing, 11(11):5547–5560, 2020.

[26] Akshay Khatri et al. Sarcasm detection in tweets with bert and glove embeddings. arXiv preprint arXiv:2006.11512, 2020.

[27] Stavros J Perantonis, Nikolaos Ampazis, and Vassilis Virvilis. A learning framework for neural networks using constrained optimization methods. Annals of Operations Research, 99:385–401, 2000.

[28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543, 2014.

[29] Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. arXiv preprint arXiv:1411.4166, 2014.

[30] Min Zheng, Bo Liu, and Le Sun. Lawrec: Automatic recommendation of legal provisions based on legal text analysis. Computational Intelligence and Neuroscience, 2022, 2022.

[31] Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. Legal judgment prediction via topological learning. In Proceedings of the 2018 conference on empirical methods in natural language processing, pages 3540–3549, 2018.

[32] Xin Zhou, Yating Zhang, Xiaozhong Liu, Changlong Sun, and Luo Si. Legal intelligence for e-commerce: multi-task learning by leveraging multiview dispute representation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 315–324, 2019.

[33] Jiaming Gao, Hui Ning, Zhongyuan Han, Leilei Kong, and Haoliang Qi. Legal text classification model based on text statistical features and deep semantic features. In FIRE (Working Notes), pages 35–41, 2020.

[34] P Nineesha and P Deepalakshmi. Automated techniques on indian legal documents: A review. In 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), pages 172–178. IEEE, 2022.

[35] Soumayan Bandhu Majumder and Dipankar Das. Rhetorical role labelling for legal judgements using roberta. In FIRE (Working Notes), pages 22–25, 2020.

[36] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019.

[37] Riya Sil and Abhishek Roy. A novel approach on argument based legal prediction model using machine learning. In 2020 International Conference on Smart Electronics and Communication (ICOSEC), pages 487–490. IEEE, 2020.

[38] Yinhan Liu, 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.

[39] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3491775

NAGWAN ABDEL SAMEE received the B.S. degree in computer engineering from Ein Shams University, Egypt, in 2000, and the M.S. degree in computer engineering and the Ph.D. degree in systems and biomedical engineering from Cairo University, Egypt, in 2008 and 2012, respectively. Since 2013, she has been an Assistant Professor with the Information Technology Department, CCIS, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Her research interests include data science, machine learning, bioinformatics, and parallel computing. Her awards and honors include the Takafull Prize (Innovation Project Track), Princess Nourah Award in Innovation, the Mastery Award in predictive analytics (IBM), the Mastery Award in Big Data (IBM), and the Mastery Award in Cloud Computing (IBM).

MAALI ALABDULHAFITH was born on September 21st, 1985 in Saudi Arabia. In 2018 she received her Doctor of Philosophy Degree (PhD) in the field of Computer Science from Dalhousie University, Halifax, Canada. In 2014, she joined the College of Computer and Information Science (CCIS) in Princess Noura University (PNU) as a Lecturer and was promoted to Assistant Professor in 2018. Her research interests lie in the area of machine learning, data analytics, emerging wireless technology, technology applications in health care. Currently, she is the Director of Data Management and Performance Measurement at CCIS at PNU overlooking and managing the strategy of the college.

S MUHAMMAD AHMED HASSAN SHAH received his B.S. degree in Computer Science from COMSATS University Islamabad in 2022. He previously served as a Research Assistant at the National Center of Artificial Intelligence (NCAI), Pakistan, focusing on advancements in Medical Imaging, Computer Vision, and AI technologies. Currently, he is pursuing an M.Phil. in Civil and Transportation Engineering at Central South University, Changsha, China, with a focus on advanced computing technologies such as AI, IoT, and smart infrastructure management. His research interests include Artificial Intelligence, Deep Learning, Point Cloud Intelligence, Digital Twins, IoT, 3D Computer Vision, and Cyber-Physical Infrastructure.

ATIF RIZWAN is currently working as a Postdoctoral Researcher at Kyunghee University, Yongin, Korea. He received his M.S. degree in computer science from COMSATS University Islamabad, Pakistan, in 2020, where he also worked as a Research Associate. He completed his Ph.D. degree in Computer Engineering at Jeju National University, Jeju, Republic of Korea, in 2024. He has substantial industry experience in mobile and web application development and testing. His research interests include federated learning, algorithm optimization, applied machine learning, and IoT-based applications.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4