Review of Computer Engineering Research
2026 Vol. 13, No. 3, pp. 173-199
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v13i3.5147

Smart justice framework using artificial intelligence for ECHR article violation classification and case outcome prediction

Najeeb Saiyed[1]
Maha Veera Vara Prasad Kantipudi[2+]
Pradeep Kumar Nagalapura Shankar Murthy[3]

[1,2]*Symbiosis Institute of Technology, Symbiosis International, Deemed University, Pune, India.*
[1]*Email: officialnajeebsaiyed@gmail.com*
[2]*Email: prasad.kantipudi@sitpune.edu.in*
[3]*S. E. A College of Engineering and Technology, Bangalore, India.*
[3]*Email: pradii123@gmail.com*


(+ Corresponding author)

Article History
Received: 7 January 2026
Revised: 14 April 2026
Accepted: 26 May 2026
Published: 4 September 2026

Keywords
AI in law
Legal analytics
Legal case outcome prediction
Legal NLP
Legal summarization.

ABSTRACT

This study aims to build and analyse an Artificial Intelligence-driven pipeline for automating the analysis of legal cases in the European Court of Human Rights (ECHR), to address transparent, ethical, and responsible decision support systems in the legal domain. The study applies a Natural Language Processing (NLP) framework to a dataset of 4,178 legal cases in the ECHR related to Articles 3, 5, 6, and 8. The pipeline integrates multi-label classification of article violations in legal cases using Multinomial Naive Bayes, Random Forest, and BERT. SpaCy models were used for Part-of-Speech tagging and Named Entity Recognition (NER). Latent Dirichlet Allocation (LDA) and BERTopic for topic modelling of the legal texts. Summarisation of these large legal cases was applied using a fine-tuned BART-large-CNN, and generation of case outcome prediction was done using LSTM and bidirectional LSTM that contain attention mechanisms. In multi-label classification, Random Forest achieved the highest accuracy of 64%, where Multinomial Naive Bayes scored 59%, and BERT scored 52%. While BERTopic found eight semantically coherent thematic clusters within the dataset. For case summarisation, BART-large-CNN achieved a 45% ROUGE-1 score and a 24% ROUGE-L score. In case of outcome prediction (text generation), the Bidirectional LSTM with attention scored a ROUGE-L score of 17% compared to the LSTM with an 8% ROUGE-L score. Findings from the framework demonstrate the potential of NLP-based approaches in automating legal case analysis, offering a research foundation for developing AI-assisted systems that support article violation detection, case summarisation, and evidence-based jurisprudence in the legal domain.

Contribution/Originality: This study develops and evaluates an end-to-end NLP pipeline that integrates four tasks, such as multi-label article violation detection, topic modelling, case summarisation, and case outcome prediction, on the ECHR legal corpus, addressing a gap where prior work has largely focused on article classification, leaving the remaining tasks underexplored.

1. INTRODUCTION

A legal system, basically all over the world, follows some principles in its functioning so that the system functions properly and is just and fair. Prediction of judgment for a case has a vital role in ensuring not only a justified output but also an efficient, reliable, and ethical legal process [1]. In recent years, precedent and legal knowledge have been primary factors on which legal experts depend to predict results [2]. However, things are changing fast, all because of the emergence of Natural Language Processing (NLP), which involves machine learning to understand languages

and analyze large volumes of text data in legal documents to help lawyers [3, 4]. Within this context, our research paper embarks on a comprehensive exploration of the current state of NLP's influence within the legal domain [5, 6]. We investigate the methodologies used, the challenges encountered, and the ethical considerations while examining how NLP performs on legal data. We go one step further to cover all aspects of NLP, from multi-label classification to Named Entity Recognition, applying Part of Speech tagging, understanding the intricacies of topic modeling in legal cases, offering insights on summarizing models, and making groundbreaking judgment predictions, all within the legal domain [7]. That is because it does offer a very promising solution in legal cases by providing effective and efficient document analysis, case law, legal opinions, or legislation. This information is effectively processed and analyzed by NLP, saving the legal professionals from the tremendous amount of time that would otherwise be spent doing this work laboriously [8, 9]. It not only improves accuracy, reduces costs, and makes the analysis of cases quicker, but is also highly scalable, which makes it adaptable to a wide range of new cases and contexts. As data continuously grows, NLP can efficiently learn from this and generalize better for future cases [10, 11]. This paper is organized as follows. Section II presents the problem statement this study focuses on, which outlines key challenges of NLP in the legal domain. Section III provides a literature survey of recent trends in NLP in legal use cases. Section IV provides an overview of the historical context and applications of NLP in the legal domain. Section V details the dataset used and preprocessing steps applied to the legal corpus. Section VI describes the methodologies and techniques employed, like multi-label classification, POS tagging, Named Entity Recognition (NER), topic modelling, legal case summarization, and judgment prediction. Section VII dwells on the evaluation metrics used to assess the performance of models. Section VIII discusses the experimental results for each task employed in this study. Section IX presents the challenges faced during the implementation of this study. Section X discusses the limitations of this work, and Section XI provides a summary of findings and Section XII presents the conclusion of the study.

2. PROBLEM STATEMENT

The integration of NLP into the legal domain for predicting case outcomes is a breakthrough; however, challenges include handling bias and fairness, adapting NLP to specific legal tasks like multi-label classification, NER, POS tagging, summarization, and judgment prediction. Implementing NLP responsibly and transparently with explainability will ensure that justice and fairness are maintained within the legal system. Addressing these challenges is crucial for ensuring model integrity in the legal field.

3. LITERATURE SURVEY

The application of NLP in legal case analysis has evolved a lot over the last decade, transforming from the usage of simple keyword-based retrieval systems to deep learning architectures. This section reviews prior work and relates the present study to the existing landscape of NLP in the legal domain. Table 1 presents a literature survey of recent trends in the use of NLP for legal tasks.

Table 1. Literature review.

Title	Year	Key findings
Comparative analysis of legal outcome prediction with detailed and summarized Text [12]	2024	Different text lengths affect prediction quality; summarization could further improve accuracy.
Deep learning algorithm for judicial judgment prediction based on BERT [13]	2020	Using BERT + text mining lifted accuracy by 8–10%; combining multiple document types is a promising extension.
Automatic Judgment Prediction via Legal Reading Comprehension [14]	2018	Leveraging facts, pleas, and legal articles improved performance; document-structure integration remains open.
A Two-Phase Sentiment-Analysis Approach for Judgment Prediction [15]	2018	Sentiment signals help classify judgments, but real-world validation is still limited.

Title	Year	Key findings
LLMs – the Good, the Bad or the Indispensable? [16]	2023	LLMs show promise, yet fall short in legal contexts; further optimisation is required.
Legal Judgment Prediction Based on BERT + LSTM-CNN Fusion [17]	2021	Fusion architecture outperforms single backbones; more variants should be explored.
Multi-Stage Case Representation Learning in Real-Court Settings [18]	2021	Multi-stage learning boosts accuracy, but interpretability must improve.
Hybrid CNN + Bi-LSTM DSS for Judicial Decisions [19]	2022	Reached 91.5% accuracy; cross-jurisdiction testing is still needed.
Legal Judgment Prediction Based on Knowledge-Enhanced Multi-Task and Multi-Label Text Classification [20]	2025	Integrating external legal knowledge with multi-task and multi-label learning improved prediction.
Legal Judgment Prediction Using NLP and ML Methods: A Systematic Literature Review [21]	2025	Identified gaps in multi-jurisdiction generalisation, interpretability, and low-resource language coverage.
NLP for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges [22]	2025	Notes the lack of standardized benchmarks and cross-lingual resources as key bottlenecks.
LegNER: A Domain-Adapted Transformer for Legal NER and Text Anonymization [23]	2025	A domain-fine-tuned transformer outperformed general-purpose models on entity recognition.
Selective Retrieval-Augmentation for Long-Tail Legal Text Classification [24]	2025	Selective RAG strategy improved classification of rare legal categories by retrieving only high-confidence relevant precedents.
Analysing Similarities Between Legal Court Documents Using Transformer-Based NLP [25]	2025	Transformer similarity models outperformed TF-IDF baselines for document matching.
Evaluating Transformer Models for NER in Indian Legal Texts [26]	2025	Pre-trained transformers fine-tuned on Indian legal corpora showed strong NER performance; domain-specific pre-training yielded greater gains than general multilingual models.

Early approaches in legal judgment prediction relied on rule-based systems or the use of statistical models. Long et al. [14] demonstrated that integrating structured case components like facts, pleas, and legal articles into a reading comprehension framework improved output prediction quality. However, their model couldn't generalize to multi-article violation cases. Similarly, Liu and Chen [15] showcased that sentiment signals derived from a case could assist in judgment classification, though their approach was validated only on small-scale Taiwanese court data with limited cross-jurisdictional applicability. These classical pipelines, while foundational, do not scale well to multilingual, multi-label legal corpora such as the ECHR dataset used in this study.

The usage of BERT has transformed legal NLP research. Wang et al. [13] reported accuracy gains of 8–10% over baselines on fine-tuning BERT on Chinese judicial corpora, and Liu et al. [17] fused BERT with a Long Short-Term Memory–Convolutional Neural Network (LSTM–CNN) hybrid, demonstrating that ensemble architectures outperform single-backbone models. The present study addresses the gap by benchmarking BERT along with classical models across a multi-label classification task, showing that Random Forest (64% accuracy) actually outperforms fine-tuned BERT (52% accuracy) on the ECHR legal dataset. This finding challenges the implicit assumption that transformer models universally dominate in legal NLP.

Ma et al. [18] proposed a multi-stage case representation learning framework that was tested on Chinese court settings, achieving improvements in accuracy, though noting that interpretability remained a critical unresolved challenge. Ahmad et al. [19] reported a high accuracy of 91.5% via the usage of a hybrid CNN-BiLSTM architecture for judicial decision prediction. Varshini et al. [12] analysed how text length affects legal outcome prediction quality, and the finding was that usage of summarised inputs can improve model accuracy. This motivates the usage of (BART-large-CNN) for summarisation in our pipeline. However, their work did not evaluate summarisation with

downstream prediction tasks, which this study addresses. The relatively modest ROUGE-1 score of 0.45 achieved in the present work is consistent with the difficulty of summarisation on large legal texts.

Vats et al. [16] evaluated LLMs for Indian legal statute and judgment prediction. Their findings suggest that despite LLMs' general language capabilities, they underperform in legal contexts, which supports the fine-tuned transformer-based approach adopted in this study. Li et al. [20] proposed K-LJP, a knowledge-enhanced multi-task and multi-label Legal Judgment Prediction (LJP) framework, showcasing that incorporating label-level and task-level knowledge improves prediction outcomes, though its evaluation remains exclusive to the context of the Chinese legal corpus. Dina et al. [21] conducted a systematic review of NLP and Machine Learning (ML) methods for LJP across studies from 2015 to 2021, highlighting issues like class imbalance, small datasets, and multilingual complexity as primary ongoing challenges in the legal domain, which are also faced in this study.

Ariai et al. [22] surveyed 131 legal NLP studies, highlighting interpretability and bias as important ongoing challenges in the legal domain. Karamitsos et al. [23] showed that domain-adapted transformers outperform generic models for legal Named Entity Recognition (NER), which proves the limitations observed with generic models (spaCy models) on ECHR data in this study. Mao [24] showed that Selective Retrieval-Augmentation (SRA) effectively handles long-tail label distributions in legal classification, which is relevant to the class imbalance problem. Oliveira and Nascimento [25] verified that fine-tuned domain-specific models perform better than general-purpose LLMs on legal document tasks. Bhandari et al. [26] suggested divergence-based dataset balancing for legal NER, offering a methodological approach that could guide future versions of such pipelines.

In contrast to prior research, this study introduces a comprehensive multi-task NLP approach applied to the ECHR legal corpus, which includes article-violation detection, entity recognition, topic modelling, case summarisation, and case outcome prediction.

4. NLP IN THE LEGAL DOMAIN

4.1. Historical Context

With the rapid progress in the legal domain, Natural Language Processing has provided novel solutions to the otherwise manual challenges. NLP's journey in the legal context harks back to the emergence of computational linguistics and textual analysis.

Initially, its applications were very basic in nature, such as text retrieval—training quicker searches for particular legal documents and references for legal professionals. Nevertheless, these systems had limited capacity with regard to grasping contextual details from data. Despite such limitations, these applications are just the very foundation of where NLP now is in the legal domain.

4.2. Applications of NLP in the Legal Domain

Quite recently, NLP in the legal domain received wide attention. NLP will become an important tool for conducting legal research and analyzing case law. Such rapid processing and extraction of insights from large volumes of legal texts are possible because of the notable capability of NLP, thus helping professionals to better equip themselves for more informed decision-making. NLP models have been proficient in the analysis of contracts and due diligence by extracting the clauses of the contracts, finding out the potential risks, and verifying legal standards. This ensures that there is a reduction in the likelihood of oversights and errors. NLP's predictive powers have increasing relevance to the legal profession.

By analyzing case data from the past, NLP algorithms will be able to predict the outcome of new cases. Lawyers and judges will be better equipped to make truly informed decisions about the 'chances of success' in any particular matter. This evidence-based practice can be enormously useful in formulating legal strategies and resource allocation.

4.3. Research Gap

With the increase in interest in applying NLP to legal corpora, current studies have largely focused on article classification. Tasks such as topic modeling, legal case summarisation, and case outcome prediction remain underexplored within the context of the European Court of Human Rights (ECHR). Furthermore, prior work rarely integrates multiple NLP tasks into an end-to-end pipeline on the ECHR dataset, making it difficult to assess how these perform in supporting legal decision-making for the corpus on ECHR. There is also a lack of studies that simultaneously evaluate traditional machine learning models, transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT), and sequence-to-sequence architectures on the ECHR legal corpus, and this study aims to address these gaps.

5. DATASET COLLECTION AND PREPROCESSING

The dataset used in this study is derived from the HUDOC database, maintained by the European Court of Human Rights (ECHR). HUDOC provides access to the full case law of the Court. For the purposes of machine learning research, the ECHR Open Data (ECHR-OD) repository, an openly published and peer-reviewed dataset derived from HUDOC to support AI and Data Science applications, was used as the primary data source [27]. ECHR-OD provides a clean, formatted, and standardised legal corpus. A sample short snippet of the article from the dataset is shown in Figure 1.

```
FOURTH SECTION

CASE OF TERGE v. HUNGARY

(Application no. 3625/15)

JUDGMENT

STRASBOURG

27 February 2018

This judgment is final but it may be subject to editorial revision.

In the case of Terge v. Hungary,

The European Court of Human Rights (Fourth Section), sitting as a Committee composed of:
```

Figure 1. Peek into our dataset.

This study uses open-source data for multi-label classification of article violation outcomes across Articles 3, 5, 6, and 8. Cases revolving around these four articles only were used for the creation of a dataset of 4,178 cases in which violation and non-violation instances are represented for each of the four selected articles, ensuring per-article label balance across the corpus. Unlike prior work that typically frames outcome prediction as a binary classification problem for a single article, this study adopts a multi-label classification approach in which a single case may simultaneously carry violation or non-violation labels across multiple articles, capturing the overlapping and co-occurring nature of human rights claims in ECHR judgments. In this present study, we extend that foundation by integrating classification with topic modelling, legal case summarisation using BART-large-CNN, and judgment prediction using LSTM and Bidirectional LSTM networks with attention mechanisms, which have not been addressed in past work. No effort was made to re-identify or profile individuals beyond what is disclosed in officially published court records.

It must be underlined that in this study, we have selected Articles covering 3, 5, 6, and 8 for both violated and non-violated type articles. These selected articles represent important facets of human rights law and have been chosen to understand how they perform when used with NLP. These articles reflect basic principles and rights, making the dataset a solid foundation for in-depth analysis and NLP-driven insights regarding human rights and legal case outcomes.

Reproducibility Note: To support full reproducibility of this study, the dataset was sourced exclusively from the publicly available ECHR-OD Dataset [27]. The dataset can be independently reconstructed by filtering ECHR cases under Articles 3, 5, 6, and 8 with both violation and non-violation outcome labels. No proprietary or restricted data was used at any stage of this work.

Article 3 - Prohibition of Torture: This article emphasizes protection from physical and mental abuse, highlighting the prohibition of torture and inhumane treatment of any kind.

Article 5 - Right to Liberty and Security: It guarantees the right to liberty and security, outlining conditions for deprivation of liberty, the rights of individuals in detention, and the entitlements of those in custody.

Article 6 - Right to a Fair Trial: This article safeguards the right to a fair trial and encompasses various procedural protections.

Article 8 - Right to Respect for Private and Family Life: It protects the right to privacy and respect for family life, including one's home and correspondence.

Table 2. Distribution of non-violation and violation articles in our dataset.

Article	Non-Violation	Violation
3	560	591
5	436	509
6	566	754
8	351	411
Total	1913	2265

Table 2 indicates the distribution of Non-Violation and Violation articles in the dataset. Surprisingly, Article 6 is highest in terms of violations, amounting to 754. This is actually the highest focus that the dataset has. Articles 3 and 5 have Violation cases of 591 and 509, respectively. Article 8 has the fewest violation cases, with 411 cases. Most of the main articles in Non-Violation would fall under Article 6, with the cases totaling 566, followed by Articles 3 and 5, with 560 and 436 cases, respectively. The sum of Violation and Non-Violation cases sums up to 2,265 and 1,913, accordingly, balancing the composition of the dataset and showing articles at the top of the dataset to be of particular interest for further analysis. As for balancing the distribution of the dataset, this paper ensures that the first 350 legal texts from each Violation and Non-Violation article were used to ensure balanced training for each label (article).

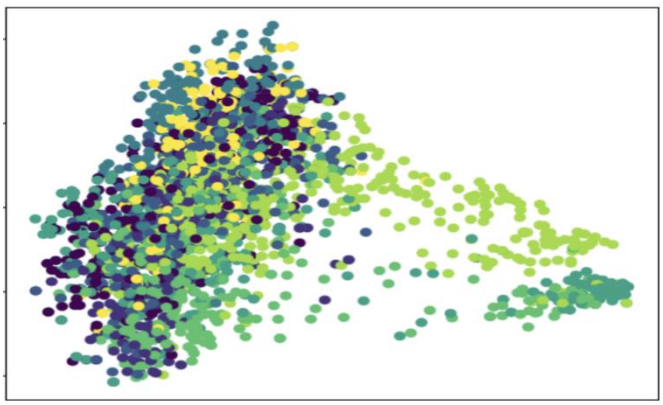

Figure 2. PCA visualization.

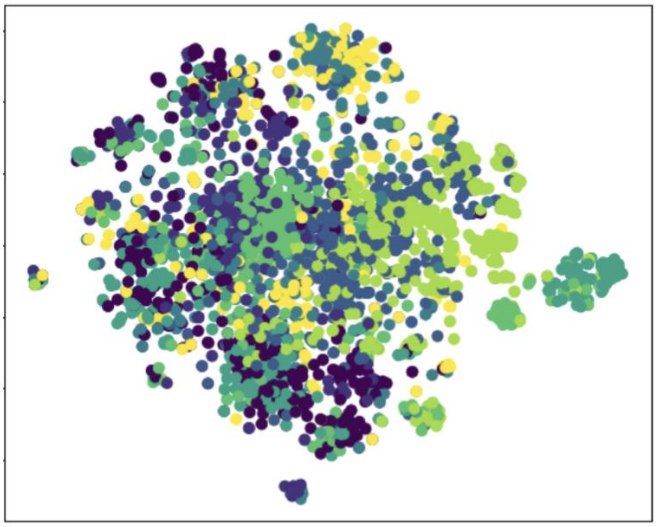

Figure 3. T-SNE visualization.

The messy clusters observed in the t-SNE and PCA visualizations are expected, given the high-dimensional and sparse nature of legal case text data, where overlapping topics make clean separation difficult, as seen from Figures 2 and 3. The fact that these clusters are messy indicates there are a lot of dimensions, normally words, in this text data. Furthermore, sparsity introduces extra complexity into these clusters. In addition, the presence of overlapping topics in the legal cases makes them not easily separable; this forms evidence that multiple cases contain several overlaps.

The data had a lot of flaws that needed to be cleaned, especially when exporting the text file to a dataframe; it displayed a lot \n as the data was raw, and even in normal paragraphs, there would be unorganised breaks during sentences. After cleaning, another major problem faced was differing from actual numbers in the data from line numbers displayed on each line in the cases. We noticed a pattern when it came to distinguishing line-number (number representing each line of a legal case) from that of the actual number (that has importance with respect to the legal case context itself); line-numbers were used in combination with a full stop. Using the regular expression library, we were able to remove such combinations of numbers and a full stop from line numbers. Using regular expressions and NLTK, this paper ensured all stop-words from our dataset were handled adequately. Preprocessing consists of text cleaning: newlines, digits, and extra spaces are unwanted patterns removed using regular expressions.

6. METHODOLOGIES AND TECHNIQUES

This study uses a multi-task NLP pipeline designed to sequentially address article violation classification, topic modelling, legal case summarisation, and case outcome (judgment) prediction in the ECHR context. The models in this pipeline are trained on tasks aligned with standard practices in the field. Each model is trained on four distinct article categories (Articles 3, 5, 6, and 8) containing both violation and non-violation labels, within a unified and reproducible framework.

Previous studies such as Wang et al. [13] and Ahmad et al. [19] have similarly used single-task models, but their scope is defined to a single jurisdiction and case category, with limited application across multiple articles. Moreover, they do not integrate the results into a broader, multi-task pipeline to understand the depth under various tasks.

This study's framework stands out by covering multiple articles, integrating tasks into a unified pipeline to understand how data and models behave across different tasks but on the same ECHR data. This study also excludes judgment outcomes from predictions to ensure predictions are based on case facts and legal provisions rather than outcome-driven patterns.

6.1. Multi-Label Classification

1. Multinomial Naive Bayes Model

It is a classifier that uses a probabilistic approach for classification. It works based on Bayes' theorem and considers the features to be independent. In this paper, we have proposed Multinomial Naive Bayes as it performs well with word-frequency-based features and handles sparse legal texts effectively. Additionally, its probabilistic framework offers transparency in decision-making, making it a strong baseline model for legal-text classification.

In the experiments, the range of hyperparameters was systematically explored by grid-search cross-validation in order to find the best combination for our Multinomial Naive Bayes model. After a great deal of experiments, the following set of hyperparameters turned out to be optimal.

Alpha (smoothing parameter): 0.1.

Fit prior: False.

Smoothing with a value of 0.1 is a relatively low smoothing, which would be appropriate with respect to our dataset. fit_prior is given to be False, which means not using prior probabilities of classes, and it will enable the model to make predictions based only on observed data.

2. Random Forest Model

A model that makes its predictions by combining the results of many decision trees. It is a flexible model in such a way that it can also be applied to classification tasks. In our research, we have chosen the Random Forest algorithm for making predictions on our law data, whether certain types of articles belong to one legal category or another based on their textual content. Each of these trees, at every given node, takes a random subset of features; hence the ensemble of trees provides robust and accurate predictions, and hence was chosen.

Hyperparameter tuning is one of the important tasks in optimizing any machine.

learning model. In our study, we systematically applied Grid Search CV in order to find out which combination of hyperparameters would give the best results for our Random Forest model. After enough experimentation using Grid Search CV, we concluded that the parameters listed in Table 3.

Table 3. Optimal hyper-parameter settings for random forest after grid search CV.

Hyperparameter	Optimal Value
Max depth	30
Max features	Auto
Min samples leaf	2
Min samples split	2
Number of estimators	200

The hyperparameters procured from Grid Search CV, as shown in Table 3, provide the best performance with respect to the classification task of our dataset. A maximum depth of 30 will enable the trees to be relatively deep, hence capturing complex relationships in the data. 'Max Features' is set to 'auto', meaning at every split the algorithm considers √(n_features) (the square root of the total number of features), hence reducing computation while retaining diversity across trees. Min Samples Leaf and Min Samples Split values, set to 2 and 2, respectively, allow the growth of trees and prevent overfitting. Finally, having 200 estimators in the ensemble helps ensure robust predictions.

3. BERT (Bidirectional Encoder Representations from Transformers)

BERT model was employed in this paper, which is a deep learning–based model for text classification. The specific details of the model are indicated in Table 4, with extra information concerning pre-processing, batch size, and other relevant details used in our study.

Table 4. Hyper-parameter settings for fine-tuning BERT for sequence classification.

Parameter	Value
Optimizer	Adam
Learning-rate	1 x 10^{-5}
Loss-function	Cross-entropy-loss
Number of epochs	15
Batch size	16

6.2. Part-of-Speech Tagging

Legal text documents are often perceived and analyzed in the legal domain. Proper perception involves proper understanding of textual law relations in case analysis, contract review, and legal research, among other legal processes. The process of assigning grammatical categories to words in text refers to part-of-speech tagging, naming nouns, verbs, adjectives, among other parts of speech.

The POS tagging captures entity extraction, such as party names, dates, venues, and contract clauses from legal documents. It becomes particularly useful in due diligence analysis, contract examination, and preparing legal briefs.

6.3. Named Entity Recognition

In this paper, we perform the NER approach by using the following available models on the spaCy library to understand how well existing models can work on legal cases:

en core web sm: The model is lightweight and moderately efficient; hence, it can be used for more memory- and computationally-bounded applications.

en core web md: Strikes a balance between accuracy and computational efficiency. It also includes word vectors, which can be useful for a variety of NLP tasks such as word similarity and document classification.

en_core_web_lg: Contains full word vectors and linguistic features. It is best used for higher-end NLP tasks like text summarization, understanding of the meaning of texts, and other uses that require context-rich applications, thus demanding deep linguistic analysis.

6.4. Topic Modelling

1. LDA (Latent Dirichlet Allocation)

In this study, we performed the LDA algorithm for topic modeling, a method of discovering latent topics that have driven the content of a corpus of legal documents. This paper describes, in detail, a topic-modeling process using LDA on a dataset. This comprises tokenization and stopword removal; the creation of the dictionary and document-term matrix; iteration to find the optimal number of topics based on coherence scores; building the LDA model with the chosen number of topics; and visualization and display of the topics inferred from the model.

2. BERTopic

A transformer-based embedding topic modeling technique combined with a clustering algorithm. An instance of KMeans with 8 clusters is initialized and passed to BERTopic; this processes the documents into a feature representation of each and creates the topic assignments and their probabilities for each document. Finally, visualize the top 31 identified topics that represent the thematic structure within.

6.5. Legal Case Summarization

1. BART Large CNN

We used the model architecture facebook/bart-large-cnn, available in the Hugging Face Transformers library. For the task of legal summarization, we have fine-tuned this model. The facebook/bart-large-cnn model specializes in the summary of texts. This is pre-trained on a huge corpus, enhancing its functionality to generate coherent and contextually relevant summaries. The model employs the usage of both bidirectional and autoregressive transformers,

hence capturing the context from the left and right sides in the text sequence to produce summaries of very high quality, similar to those written by humans, while retaining the actual meaning.

Preprocessing for training: First, the input and target sequences have to be tokenized and encoded. This was done with the Hugging Face Transformers library through the so-called AutoTokenizer class. We used a maximum length of input sequences of 750 tokens, and we set the target sequence to a maximum of 500 tokens. This will pad or truncate the input data to reach these maximum lengths to guarantee that the model can handle variable-length input.

This included tokenization, adding decoder input IDs, and decoder attention masks. The model was trained using a batch size of 4 for training and evaluation. Batch size is one of the hyperparameters that impacts the memory usage and training speed. We allowed the model to learn iteratively from the training data for 50 epochs. We use a fixed-number-of-steps evaluation policy, specifically every 1000 steps. This allowed us a constant gauge of the model's performance as it was training. Similarly, we save checkpoints of the model every 1000 steps and put a cap on the total saved checkpoints at 2.

6.6. Judgment Prediction

1. LSTM

The model architecture is defined as shown in Figure 4 and was employed in this study.

1. An input layer with a shape matching the padded sequence length.
2. An embedding layer that converts numerical word indices into dense vectors.
3. Two stacked LSTM layers, denoted as lstm layer1 and lstm layer2, each with hidden units and the option to return sequences.

We divided the dataset, 80% of the data was set for training, and the remaining for validation. During model training, batch size was set to 256 and epochs to 10.

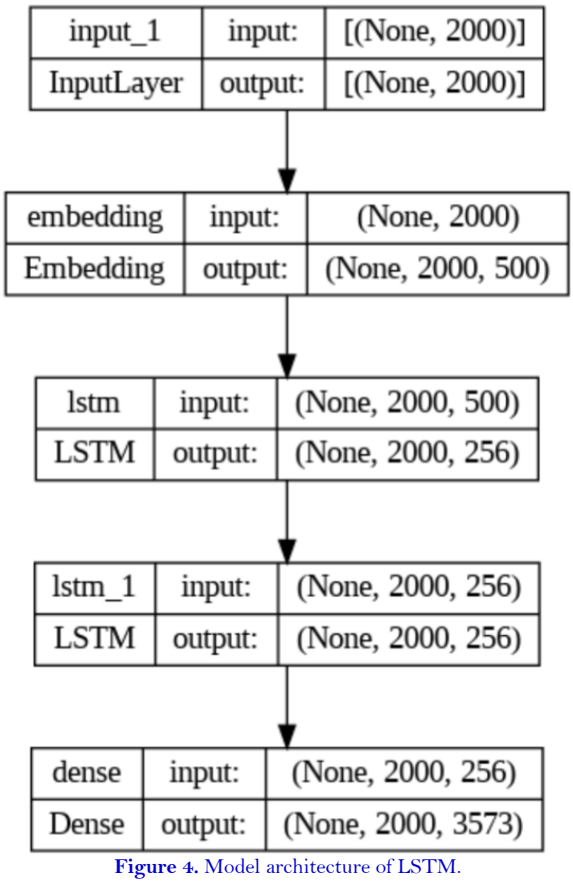

Figure 4. Model architecture of LSTM.

To allow teacher forcing, target sequences are prepared by shifting the original judgment_sequences one time step forward. This shifted version is called judgment_sequences_shifted. From this sequence, the training data (Judgment_train_shifted) and validation data (Judgment_val_shifted) are created.

During training, the model is fed with input_train and judgment_train_shifted data in batches for a set number of epochs. Additionally, callbacks for early stopping and model checkpointing are used to minimize overfitting and ensure that the best model is saved during the training process.

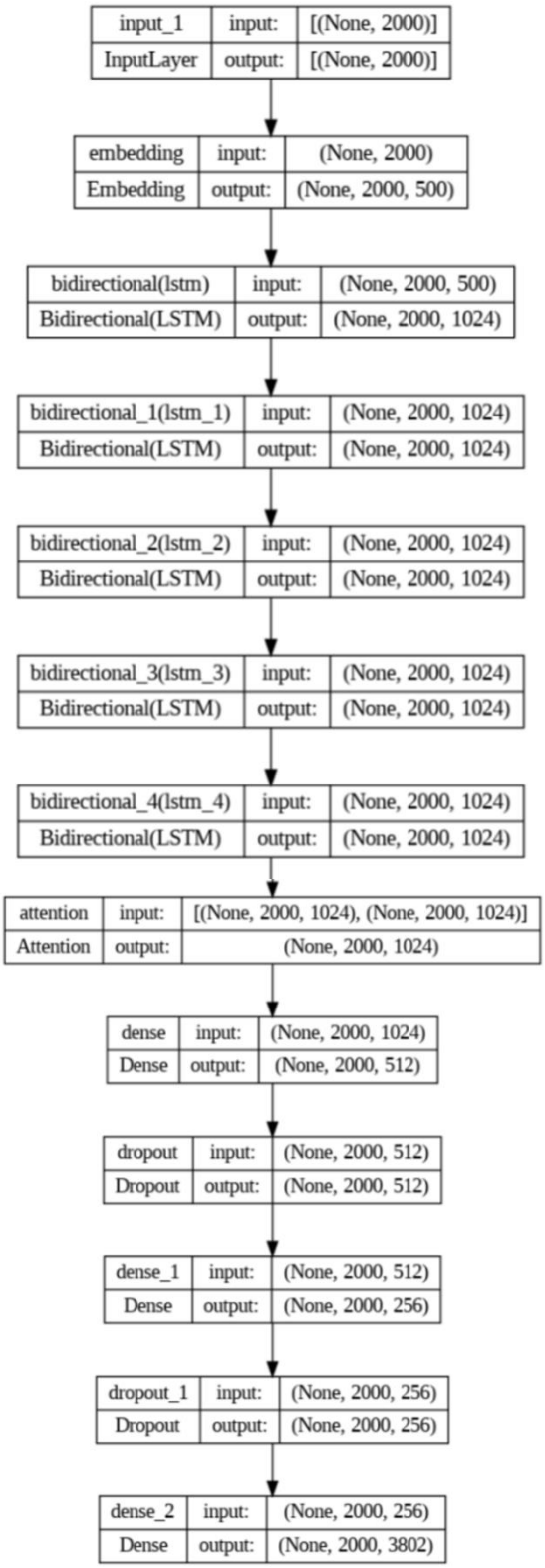

Figure 5. Model architecture of Bi-LSTM.

2. Bidirectional LSTM

Figure 5 displays the architecture of the Bidirectional LSTM model used for summarization. The model has shared the same batch size and number of epochs as the previously discussed LSTM model for maintaining consistency throughout.

1. Input Layer: An input layer with a shape matching the padded sequence length (2000 in this case).
2. Embedding Layer: An embedding layer with 1,901,000 as the input dimension and an embedding dimension of 500.
3. Bidirectional LSTM Layers: Five layers of Bidirectional LSTM, each with 512 units and configured to return sequences.
4. Attention Layer: is applied to the output of the Bidirectional-LSTM layers.
5. Dense Layers: Two dense layers were used with 512 and 256 units, respectively.
6. Dropout Layers: A dropout rate of 0.5 was used.
7. Output Layer: The output layer has 3,802 units and the softmax activation function.

The model is trained with the following parameters, which are consistent with the previous LSTM model, where the batch size was kept at 256 and 10 epochs.

7. EVALUATION METRICS

7.1. Multi-Label Classification

1. Precision

Precision evaluates the performance of a multilabel prediction model. It measures how correct the positive predictions of the model actually are. In particular, for each label i, it calculates the true positive predictions with respect to the total positive predictions it has made for that particular class.

2. Accuracy

Accuracy evaluates the overall correctness of each prediction in the multilabel classification model, which is computed based on the proportion of correct predictions against the total data points.

3. Classification Report

The classification report shows in detail how well the model is performing for each label, including precision, recall, F1-score, and support. Precision is a measure of how correct the predictions for positives are. Recall is the metric for how well the model can identify positive instances correctly. The F1-score is a measure that combines precision and recall into one for a balanced measure. Support refers to the counts of actual instances in each label.

7.2. Text Summarization and Text Generation

1. BLEU (Bilingual Evaluation Understudy)

BLEU is a metric commonly used by researchers in the field for evaluating machine-generated texts, either summaries or translations. It basically measures how similar the generated text is compared to the human-generated reference text. The higher the BLEU score, the better the quality of the generated text.

2. ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

A set of metrics has been specifically crafted to assess the quality of text summaries. Such metrics fall into recall and precision, calculating or estimating the overlap between one or more reference summaries and a generated summary. ROUGE scores are used for assessing the informativeness and quality of the generated summaries. In our experiments, BLEU and ROUGE metrics are used to evaluate the performance of text summarization and text generation models. Both measures are quite indicative of the quality of the generated text and thus can be used to compare different approaches.

8. RESULTS

8.1. Multi-Label Classification

It is noteworthy that Multilabel classification has been performed where data was passed to models as "facts" and "law." And not judgment to ensure the model understands the context and then predict a violation of the article, rather than simply passing the entire case with judgment for training.

1. Multinomial Naïve Bayes

On training Multinomial Naive Bayes on the legal data, we get the classification report as shown in Table 5.

Table 5. Classification report on multinomial Naive Bayes.

Class	Precision	Recall	F1-Score	Support
Nonviolated Art. 3	0.53	0.55	0.54	128
Nonviolated Art. 5	0.49	0.59	0.54	74
Nonviolated Art. 6	0.68	0.67	0.67	114
Nonviolated Art. 8	0.50	0.49	0.50	69
Violated Art. 3	0.56	0.63	0.59	118
Violated Art. 5	0.48	0.43	0.45	91
Violated Art. 6	0.81	0.71	0.75	139
Violated Art. 8	0.55	0.55	0.55	80
Accuracy			0.59	813
Macro-Average	0.58	0.58	0.58	813
Weighted-Average	0.60	0.59	0.59	813

On analyzing the Classification report from Table 5 given for Multinomial Naïve Bayes, there is a weighted average accuracy of 0.59 and a macro-average F1-score of 0.58 for the whole of the class. Noticeably, the model did much better in classifying violated articles than it did for the non-violated articles by displaying higher precision, higher recall, and higher F1-score for each class related to an article that was violated.

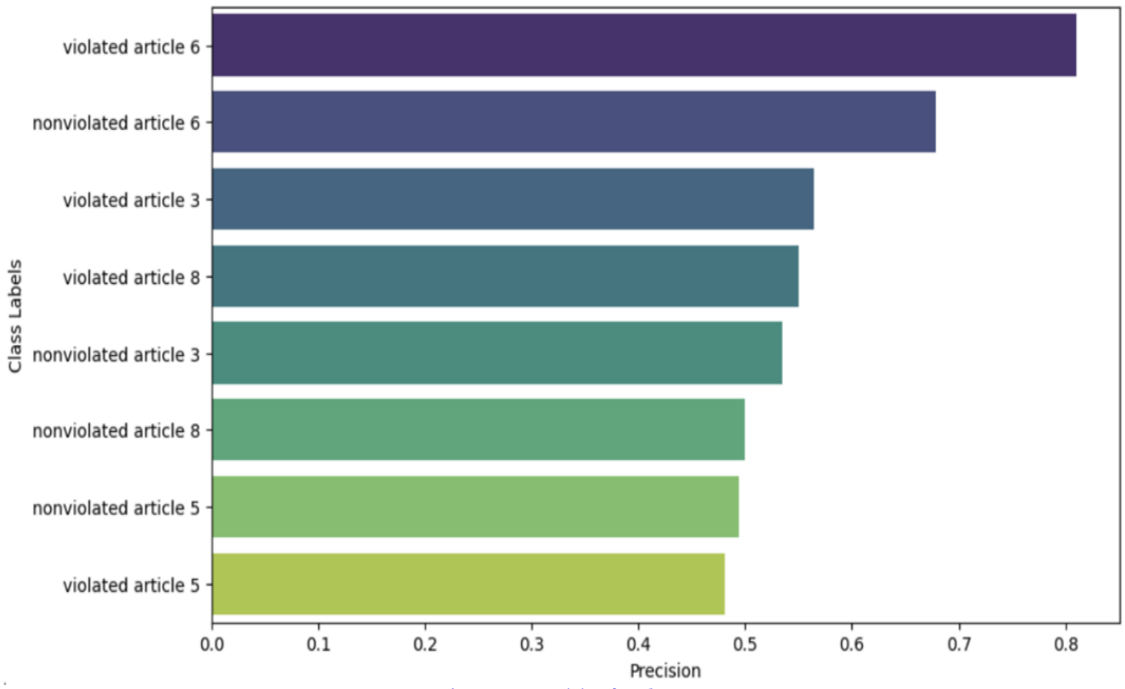

Figure 6. Precision by class.

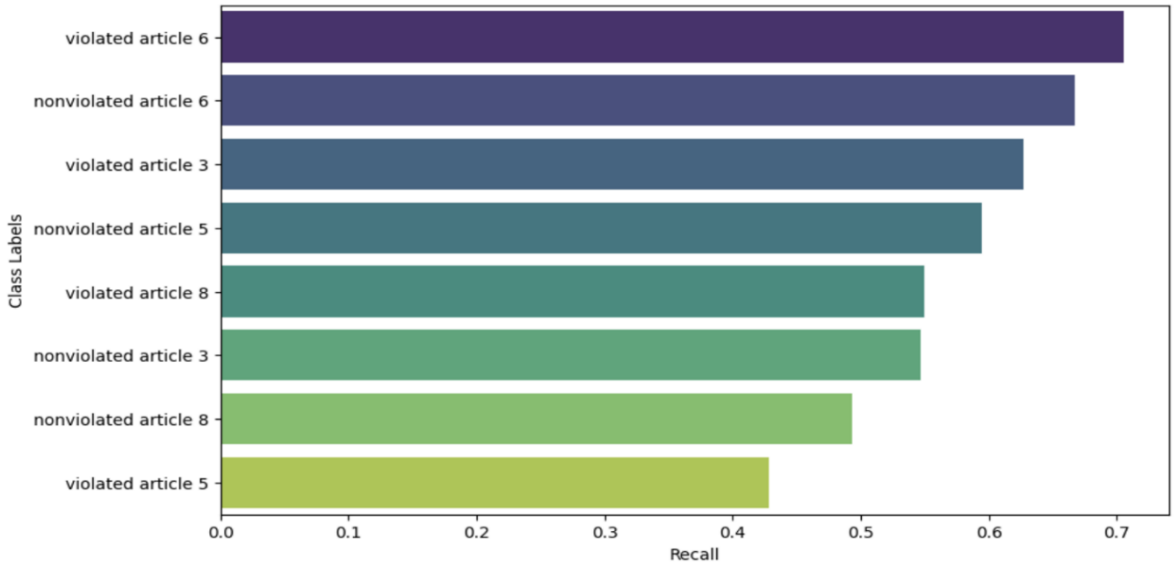

Figure 7. Recall by class.

From Figures 6,7 and 8, the best of them is 0.81 for the Violated Article 6 with respect to precision and 0.75 with respect to the F1-score. This does convey that in the case of Violated Article 6, the model is pretty confident in its predictions. The lowest performance is on Violated Article 5, with a precision of 0.48 and an F1-score of 0.45. This indicates that the model performs poorly in distinguishing violated Article 5 cases from the other classes.

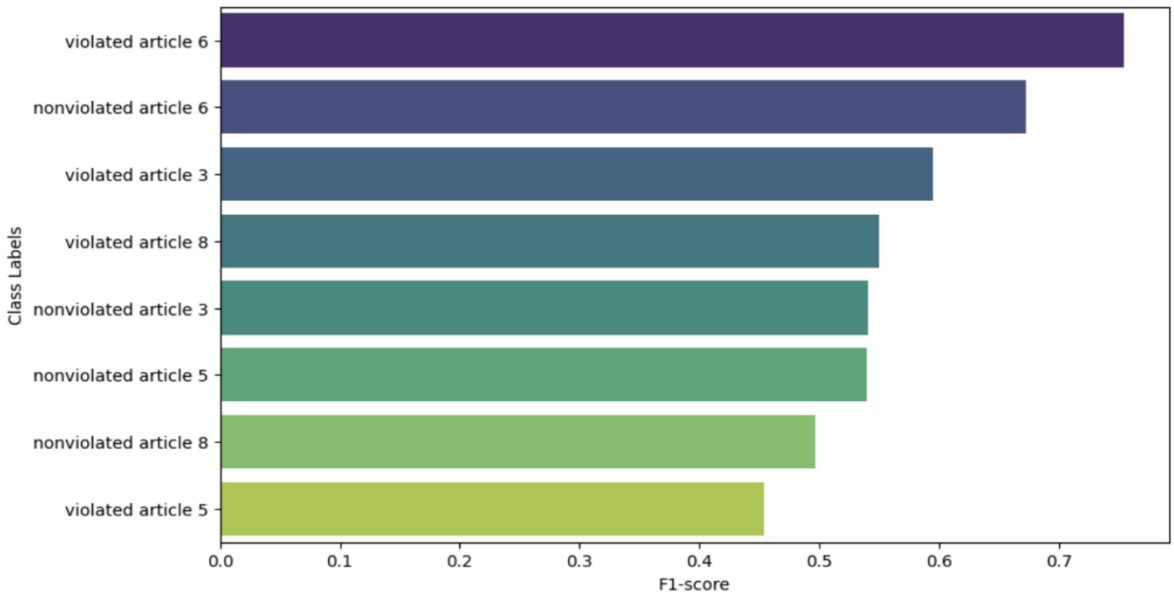

Figure 8. F1-score by class.

One possibility is that features for violated articles are more distinctive than the ones for non-violated articles. For instance, violated articles may contain some specific keywords or phrases that enable the model to understand violated articles much better.

In general, the multinomial Naive Bayes model has reasonable accuracy and F1-score in this data set. This multinomial Naive Bayes model is fairly reasonable on all classes in the dataset, though there is still margin for improvement in non-violated article classes.

2. Random Forest

On training Random Forest on the legal data, we get the classification report as shown in Table 6.

Table 6. Classification report on random forest.

Class	Precision	Recall	F1-Score	Support
Nonviolated Art. 3	0.56	0.59	0.57	128
Nonviolated Art. 5	0.47	0.53	0.50	74
Nonviolated Art. 6	0.77	0.68	0.72	114
Nonviolated Art. 8	0.70	0.55	0.62	69
Violated Art. 3	0.59	0.68	0.63	118
Violated Art. 5	0.47	0.44	0.45	91
Violated Art. 6	0.79	0.81	0.80	139
Violated Art. 8	0.73	0.74	0.73	80
Accuracy			0.64	813
Macro-Average	0.64	0.62	0.63	813
Weighted-Average	0.64	0.64	0.64	813

The classification report from Table 6 shows that the model has a decent hit value of 0.64 and a macro-average F1-score of 0.63 across all classes. It is obvious from the model performance that the articles that were violated do relatively better as compared to those that were not violated, since their respective precisions, recalls, and F1-scores are higher for all classes corresponding to the violated articles.

The model has a top accuracy and F1-score for the Violated Article 6 class with values 0.79 and 0.80, respectively. This indicates that this model can predict Violated Article 6 with a relatively high confidence level.

3. BERT

On training BERT, we get the classification report as shown in Table 7.

Table 7. Classification report on BERT.

Class	Precision	Recall	F1-Score	Support
Nonviolated Article 3	0.45	0.41	0.43	128
Nonviolated Article 5	0.43	0.42	0.42	74
Nonviolated Article 6	0.54	0.60	0.56	114
Nonviolated Article 8	0.45	0.45	0.45	69
Violated Article 3	0.51	0.54	0.53	118
Violated Article 5	0.39	0.38	0.39	91
Violated Article 6	0.65	0.71	0.68	139
Violated Article 8	0.63	0.50	0.56	80
Accuracy			0.52	813
Macro Avg	0.51	0.50	0.50	813
Weighted Avg	0.52	0.52	0.51	813

The classification report from Table 7 shows that the model has a reasonable accuracy of 0.52 and a macro average F1 score with a value of 0.50 across all classes. The model tends to perform better on classifying articles that were violated than on those which didn't, showing increased precision, F1-score, and recall across all classes related to articles that were violated.

The highest accuracy and F1-score for violated article 6 are 0.65 and 0.68, respectively, which shows that the model can identify violated article 6 with relatively high confidence. The lowest accuracy and F1-score for violated article 5 are 0.39 and 0.39, respectively, which means the model has difficulty distinguishing the instances of violated article 5 from the other articles.

8.2. POS Tagging

It could be done in a procedure called part-of-speech tagging, whereby one particular tag of parts of speech is assigned to each word in the case. Results from this can then be used in extracting valuable insights from the texts, but also for keeping in mind the enhancement in the actual understanding of legal texts.

Legal concepts identifiable from a document based on POS tags include nouns such as "party", "contract", and "tort"; verbs such as "breach", "commit", and "sue"; and adjectives, for example, "negligent", "intentional", and "fraudulent." This information may subsequently be used in the development of a structured representation from the document, a useful thought in carrying out tasks such as legal research and summarization.

The POS tags have very wide applicability in legal fact extraction, such as parties in a lawsuit, terms of a contract, and elements of a crime. Such information will go a long way in constructing the factual outline of the document that can be used for litigation support and legal analytics.

Table 8. Result from POS Tagging of a randomly selected legal case from the dataset.

POS Tag	Count
CountNN	842
NNS	286
NNP	365
NNPS	1
VB	178
VBD	299
VBG	88
VBN	280
VBP	26
VBZ	59

The results from Table 8 indicate the most frequent tags in the text are common nouns (NN): 842, plural nouns (NNS): 286, proper nouns (NNP): 365, past-tense verbs (VBD): 299, past-participle verbs (VBN): 280, base-form verbs (VB): 178, gerunds (VBG): 88, third-person singular verbs (VBZ): 59, and non-third-person singular verbs (VBP): 26.

We have chosen one legal case at random for this analysis to get an understanding of how the POS tags are distributed in the randomly chosen legal case. The relatively large number of nouns suggests that the text is probably going to be more descriptive.

The presence of a sizeable number of proper nouns suggests that the text might be talking about some specific individuals, places, or things. A large number of verbs naming actions and states of being, therefore, suggests that the given text may be action-oriented. The presence of a variety of adjectives suggests that the text might be descriptive of the qualities or characteristics of the nouns in the text.

Generally, the results from the POS tagging suggest that the text could be pretty descriptive and active; it can have some persons, places, or things involved and could describe some qualities or characteristics of the nouns within the text.

8.3. Named Entity Recognition

It is the technique of NER to identify and classify named entities in unstructured or textual data into predefined categories such as persons, organizations, locations, dates, and other custom categories of words or phrases.

As the legal cases in our dataset had some European words, which was a limitation, in this study we have applied pre-trained models available by spaCy so as to understand how well the existing models are doing in performance, which can help researchers of this domain to get to the next step.

On running these models on our dataset, we get the following output as shown in Figures 9,10 and 11.

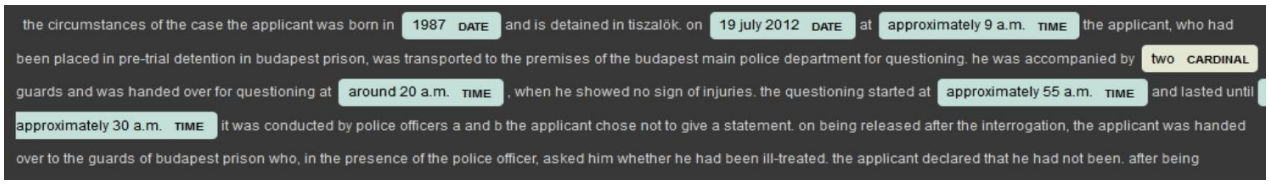

Figure 9. Output by en core web sm.

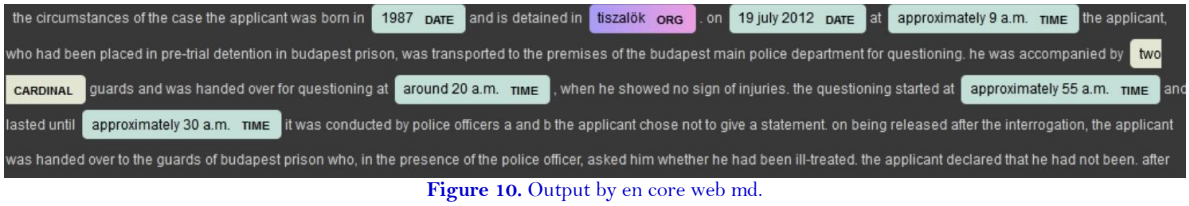

Figure 10. Output by en core web md.

Figure 11. Output by en core web lg.

By human evaluation, from the three model outputs from Figures 9,10 and 11, it can be seen that the large model performs better since it accurately detects "Budapest" as ORG, while the medium model fails to capture this. However, small and large models are able to accurately detect "Data" and "Time" from the data. The small model is not able to detect well, except it detects "Time" and "Date" upon finding in the dataset.

It is also to be noted that the Large and Medium pre-trained models have tagged 'tizalok' as ORG, whereas these, in fact, refer to places in Europe. This shows that proper custom training on these models could work even better on legal cases, as the names of Judges and places, and laws-art. being violated or not, can be tuned more accurately. Again, due to the existence of each country's own organizations and places, creating a common model that works on all legal cases will be difficult. So, configuring the model for this will require a huge amount of data preparation, but it can output quite an accurate entity recognition.

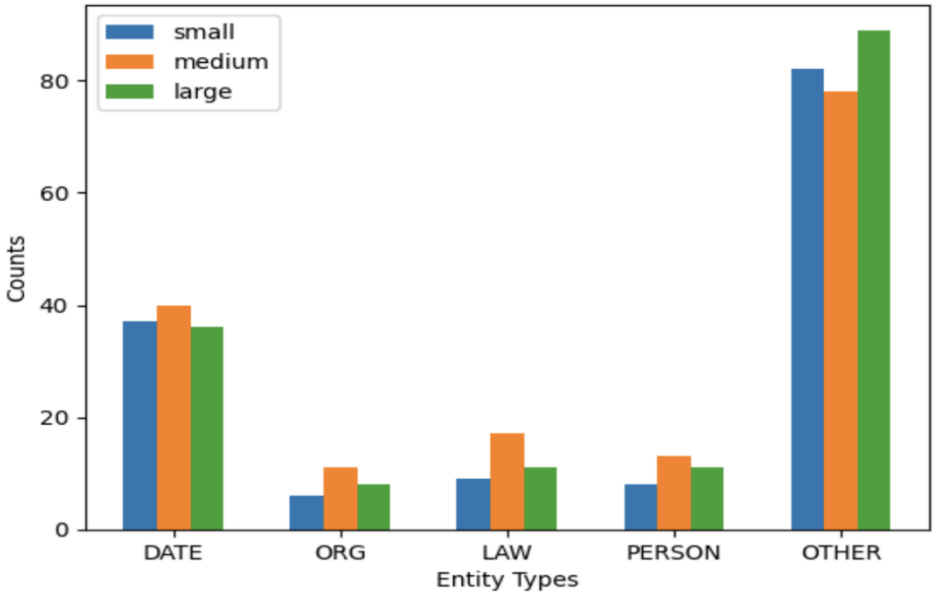

Figure 12. Entity counts in different models.

Figure 12 illustrates entity counts according to DATE, ORG, LAW, PERSON, and OTHER, classified by the small, medium, and large model sizes. From this bar chart, one generalization could be that the large model consistently outperformed both the smaller and the medium model across all entity types, especially for "OTHER," in which it showed a remarkable spike in counts. Overall, the "PERSON" and "LAW" categories have lower counts, which may indicate potential difficulties in entity identification within those particular types. On the other hand, both "DATE" and "ORG" seem relatively well-balanced across the three models, showing that while model size is crucial

for performance, entity types might need re-tuning in order to improve the accuracy of detection. On the whole, it points to the fact that the size of the model has some bearing on increasing the performance of entity recognition in natural language processing.

8.4. Topic Modelling

Topic Modelling is one of the most applied techniques in NLP, ideal for the automatic detection of hidden themes or topics, which can be realized when the information is spread within a document in a non-uniform fashion.

Its applications vary from information retrieval to content summarization by keeping pace with the documents through which one gets a vast amount of data in textual form. It also helps in uncovering hidden relationships within the data.

1. By using LDA

LDA is a type of topic-modeling technique that has achieved widespread acceptance because of its capability to uncover latent thematic structures in a collection of documents. In this paper, we make use of the investigations by LDA. First, when we applied this technique to our dataset, we saw the following graph, as shown in Figure 13.

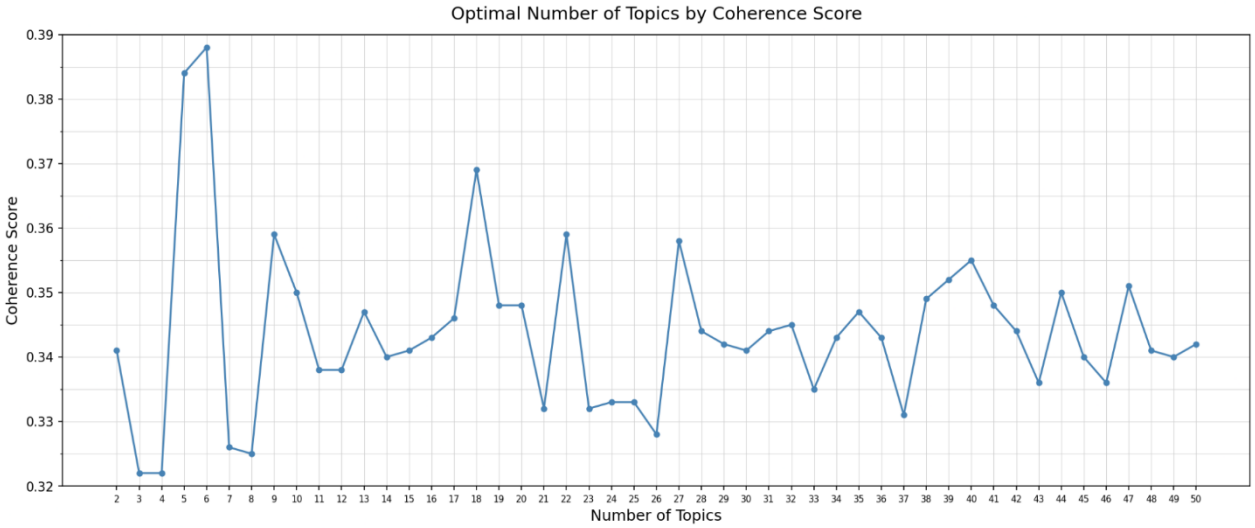

Figure 13. Deciding the optimal number of topics w.r.t coherence score.

Graphing the number of topics against coherence scores, one observes that the number of topics for which the maximum coherence score is returned is 6, with a coherence score of 0.387.

Table 9. Topics inferred by LDA.

Topic	Words
0	Legal Procedures
1	Surveillance and Data
2	Legal Entities and Court Proceedings
3	Detention and Convention
4	Police Investigation
5	Court Proceedings and Articles

Extracting topics using LDA from the legal corpus, we get the following topics (simpler shown in Table 9):

Topic 0: Legal Processes

This topic seems to be about the legal field because of the terms "applicant," "Court," "Article," and "§." Because punctuation marks like "," and "." are also on display, it may refer to the structure of a legal document.

Topic 1: Surveillance and Data

The topic appears to relate to surveillance and data, using words that span across "surveillance", "data", and "applicants". It also contains some of the common punctuation marks, which may be related to legal data matters.

Topic 2: Legal Entities and Proceedings in the Court

This topic would deal with such terms as "applicant, Court, and Article," indicating that it is a topic mainly dealing with entities and court proceedings. There are punctuation marks, which could mean that it might have to do with the structure of legal documents.

Topic 3: Convention and Detention

The third topic is connected to detention, the Convention, and legal terms. Most salient words under this topic are "detention", "Convention", and "applicant".

Topic 4: Police Investigation

The subject seems related to police investigations and the enforcement of law, as terms like "police," "investigation," "officers," and "prosecutor" were mentioned. It could, therefore, relate to cases involving legal courts for law enforcement agencies.

Topic 5: Court Proceedings and Articles

This topic appears to deal with court proceedings and articles of law, judging from the terms used: "Court, Article," and forms of punctuation.

2. By using BERTopic

The output of text data processing by the BERTopic model with KMeans clustering resulted in eight different topics. For each topic, the most frequent and representative words are given in Figure 14. At this stage, the top five words of a topic were determined, while the word scores signaled the relative importance of these words.

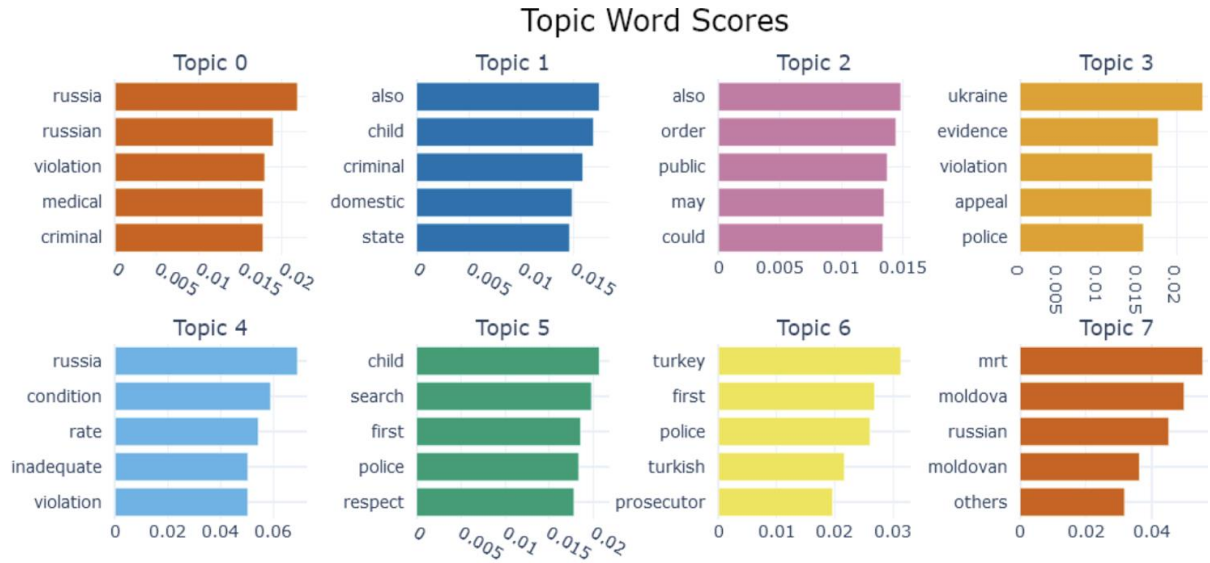

Figure 14. Topic Word scores for extracted themes using the BERTopic model.

So, topic 0 seemed to feature keywords such as "Russia," "Russian," and other important keywords included "violation," "medical," and "criminal." That does tend to suggest a topic of medical or criminal issues in Russia.

This was dominated by terms such as "also," "child," "criminal," and "domestic," suggesting that it mainly dealt with matters related to children and domestic crimes.

Among the few terms emphasized in Topic 2 are "order," "public," "may," and "could," which could probably show discussions related to public order or governance.

Topic 3 was heavily associated with Ukraine, and the leading terms include "ukraine", "evidence", "violation", "appeal", "police", which may suggest legal or police matters regarding Ukraine.

Topic 4 was about Russia again, with words added like "condition," "rate," "inadequate," and "violation"; however, it would indicate discussions of inadequate conditions or violations in Russia.

Topic 5 focused on "child," "search," "police," and "respect," which suggests a topic related to a police force and the protection of children.

Topic 6, which included words like "turkey," "first," "police," and "turkish," implied something to do with police-related issues in Turkey. Topic 7 included terms like "mrt," "moldova," "russian," "moldovan," and "others," and perhaps reflects discussions on Moldova, Russia, and associated geopolitical issues.

The word scores for each bar chart represent the relative importance of the word to the topic in which it appears. A higher score for some words in a topic would imply that those words are more associated with the core theme of that topic. The results provide structured insight into the main thematic clusters within these documents.

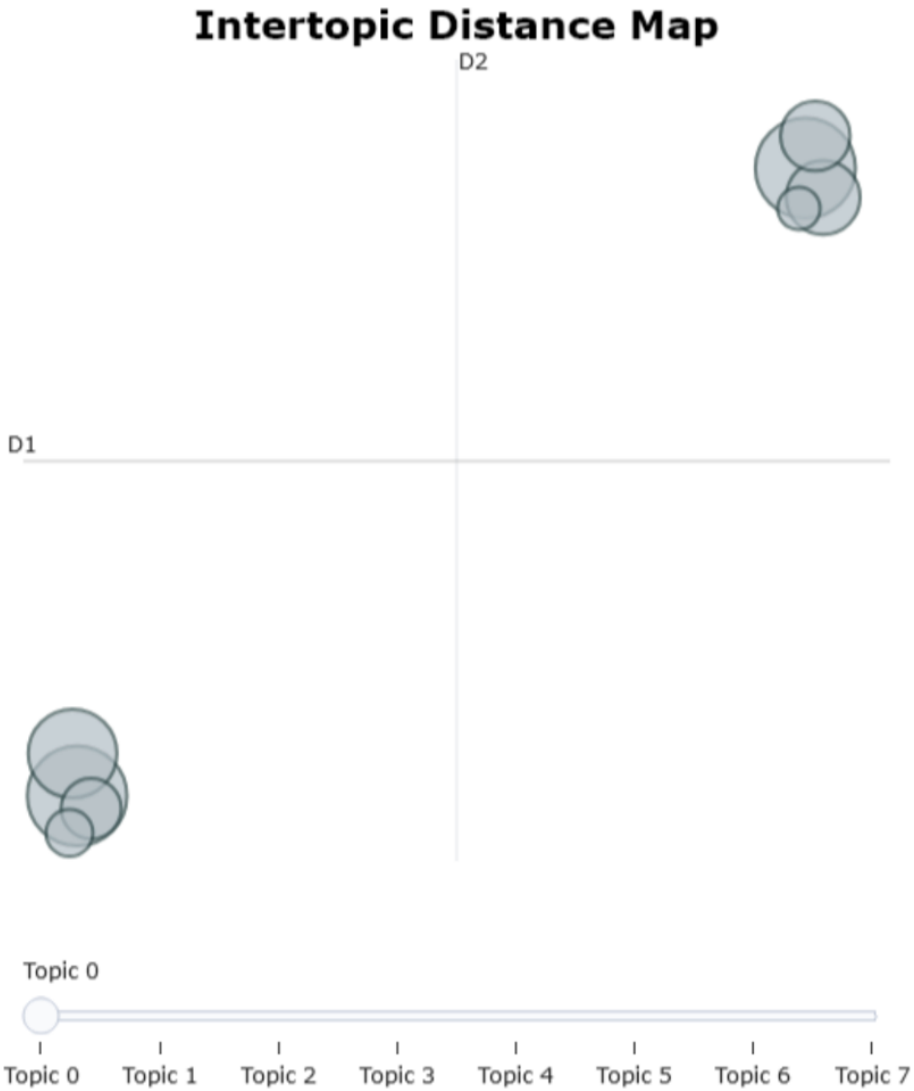

Figure 15. Intertopic distance map visualizing relationships between topics.

The Intertopic Distance Map from Figure 15 has resolved two clear-cut clusters of topics. These would then mean that the themes within one cluster are semantically related but quite different from those in the other cluster. The largest bubble in the lower-left, Topic 0, suggests it is the most dominant topic in the data. Topics in the upper-right are smaller and more evenly distributed. Overlapping bubbles within each cluster are suggestive of some thematic overlap among the closely related topics.

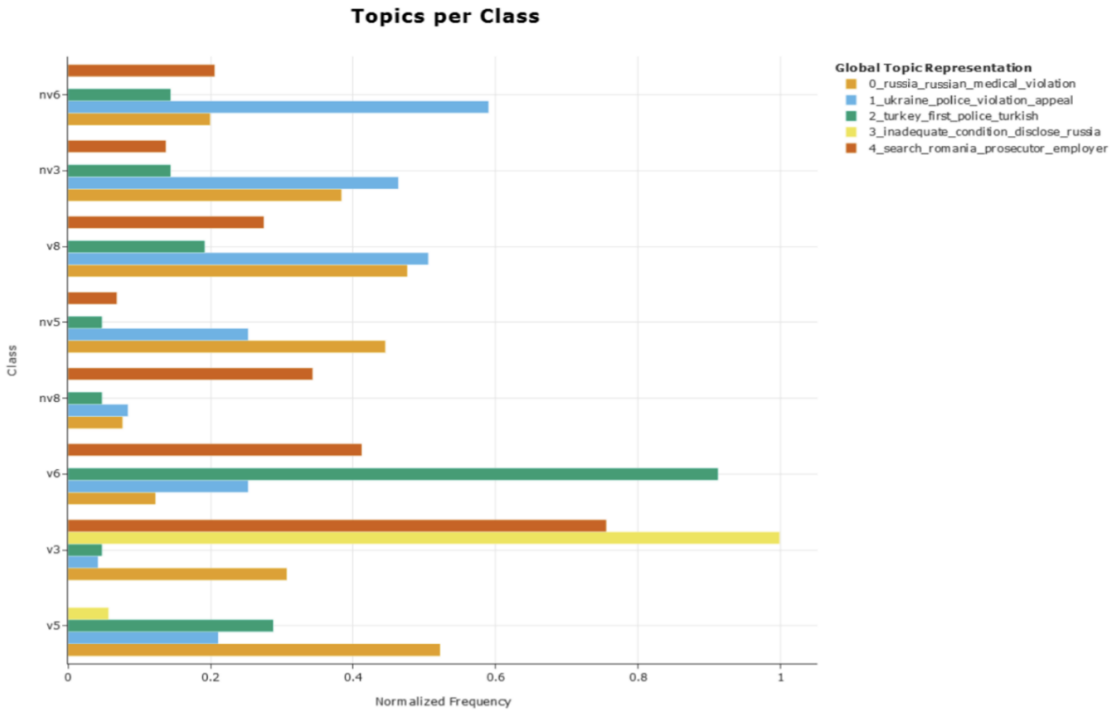

Figure 16. Topics per class bar chart.

The chart in Figure 16 shows how the five main topics are spread across different classes, using normalized frequency on the x-axis. It reveals that some topics are much more common in certain classes, suggesting clear thematic differences in the text data. For instance, Class "v3" is mostly linked to Topic 3 ("inadequate_condition_disclose_russia"), Class "v6" is strongly tied to Topic 2 ("turkey_first_police_turkish"), and Class "nv6" is mainly connected to Topic 1 ("ukraine_police_violation_appeal"). This means each class focuses on specific themes, reflecting distinct issues or ideas present in the grouped documents.

8.5. Legal Case Summarization

Summarization is similarly an important step that not only involves condensing large volumes of legal documents but also summarizes them into concise information.

It saves time and more easily conveys an understanding of the legal arguments and decisions to attorneys, judges, and even to the public in any consideration of complex legal issues.

1. BART LARGE CNN

In this work, "Bart large cnn" was used, and the same model name was used for creating the tokenizer. The data was then divided into 80% for training and 20% for validation before training. Further, this data is tokenized on both the input text, which constitutes our legal cases, and the summaries for each case. The pre-processing stage included padding, truncation, and adjusting the maximum token length for both input text and summaries.

After prepping the data and initializing the model with a batch size of 4 and 50 epochs, we further fine-tuned the model for summarization of the legal cases. Thus, the following results were obtained on the training of the model on the Bart large CNN.

Table 10. BART summarization results.

Metric	Score
BLEU Score	0.04
ROUGE-1	0.45
ROUGE-2	0.14
ROUGE-L	0.24

Table 10 indicates that a BLEU Score of 0.04 implies a very low level of precision in the generation of summaries. Normally, a score close to 1.0 from BLEU would mean a very high level of similarity with the reference text. On the other hand, the ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, metric suite includes ROUGE-1, ROUGE-2, and ROUGE-L—specifically aimed at the content overlap of the generated summary with the reference text. The values for ROUGE-1 Score, ROUGE-2 Score, and ROUGE-L Score are 0.45, 0.14, and 0.24, respectively. These scores are indicative of the system's capability regarding recall of information from the source text. Curiously enough, ROUGE scores give more emphasis to recall compared to precision.

```
Generated Summary: FOURTH SECTION CASE OF TERGE v.
3625/15) against Hungary lodged with the Court under Article 34 of the Convention for the Protection of Hu
man Rights and Fundamental Freedoms (“the Convention”) by a Hungarian national, Mr István Terge (“the appl
icant”), on 9 January 2015.
After being transported back to Budapest Prison,.e applicant was handed over to the guards of Bud
apest Prison who, in the presence of the police officer, asked him whether he had been ill-treated.
 unrecognised a report was drawn up on the incident and photos were taken of the injuries. Budapest Main p
olice Department for investigation, from where he was taken to the military hospital by the police officer
s conducting his interrogation and allegedly assaulting him.
11. The medical report noted bleeding on the lower lip, jaw sensitive to pressure, and
```

Figure 17. Generated summary by BART.

```
'In the case of Terge v. Hungary (Application no. 3625/15), the European Court of Human Rights (ECtHR) add
ressed allegations of ill-treatment and the adequacy of the investigation conducted in response to those a
llegations, under Article 3 of the European Convention on Human Rights (ECHR). Mr. István Terge, a Hungari
an national, claimed he had been physically ill-treated during police interrogation while in pre-trial det
ention at the Budapest Main Police Department, alleging that a police officer had beaten him, resulting in
injuries. Medical examinations confirmed his injuries, but the investigation was discontinued due to insuf
ficient evidence. The ECtHR found no substantive violation of Article 3, as there was insufficient evidenc
e to prove ill-treatment. However, they determined a procedural violation of Article 3 due to an inadequat
e investigation characterized by delayed witness questioning and failure to collect essential evidence. As
a result, the Court awarded the applicant €6,000 for non-pecuniary damage and €2,000 for legal costs and e
xpenses, requiring Hungary to make these payments within three months and imposing default interest for de
lays. In summary, while no substantive ill-treatment was proven, Hungary was found to have violated the pr
ocedural aspect of Article 3 by failing to conduct an effective investigation into the allegations of ill-
treatment.'
```

Figure 18. Actual summary.

Figures 17 and 18 display the generated summary from BART and the actual summary, respectively.

8.6. Judgment Prediction

1. LSTM (Long-Short Term Memory)

Table 11 provides the results obtained from training an LSTM for judgment prediction.

Table 11. LSTM results.

Metric	Value
BLEU Score	0.02
ROUGE-1 recall	0.7099
ROUGE-1 precision	0.0454
ROUGE-1 F1 Score	0.0854
ROUGE-2 recall	0.4976
ROUGE-2 precision	0.0175
ROUGE-2 F1 Score	0.0338
ROUGE-L Recall	0.6641
ROUGE-L Precision	0.0425
ROUGE-L F1 Score	0.0799

From Table 11, the model performs worse in text generation, as can be concluded from a BLEU score of only 0.02, which reflects a low degree of similarity between the generated text and the reference text. With closer looks at the ROUGE scores, rouge-1, rouge-2, and rouge-l, one can notice that deficits are observable for both precision and recall, yielding quite low levels of F1 scores. For instance, the F1 score is 0.0854 for 'rouge-1', 0.0338 for 'rouge-2', and 0.0799 for 'rouge-l'; these are indicative of the inferior quality of generated text in relation to the reference

text. Evidently, enhancement in model performance should also improve the quality and coherence of the generated text.

Figure 19 displays the output generated by LSTM.

```
expense takes self concurrent 42332 low investigations 77617 runs 323 percentage
episodes 3456 being fmi" yard demonstrate legality watt 66 "compatible pain cases
privacy ground declare partition mass forth nails shërbim required clearly
representation with notified extra 000 asked detergent deficiency belt latest simply
becoming mai disruptions disproportionate consisted discussion discretion sleeping
changed" 128 grill 66 remedied review shpejtë disproportionate deprived 25664 above
```

Figure 19. Small Snippet of the generated output from LSTM.

2. Bidirectional LSTM (Bidirectional Long-Short Term Memory)

From Table 12, the BLEU score is very low, amounting to 0.04, which is too low and, hence, generally poorly similar to the reference.

Table 12. Bi-LSTM results.

Metric	Score
BLEU score	0.04
ROUGE-1 Recall	0.6364
ROUGE-1 Precision	0.1148
ROUGE-1 F1 Score	0.1944
ROUGE-2 Recall	0.4237
ROUGE-2 Precision	0.0506
ROUGE-2 F1 Score	0.0905
ROUGE-L Recall	0.5620
ROUGE-L Precision	0.1013
ROUGE-L F1 Score	0.1717

Both scores denote partial unigram and bigram overlap with the reference text in the generated text, and further improvement in the content and fluency is required for high quality. Figure 20 shows the output generated by BiLSTM.

```
damp additionally reliance injuries a reliance injuries injuries injuries reliance
under reliance injuries damp disperse injuries other cc damp reliance with reliance
photographs injuries injuries reliance assessment §\xa01 assessment with under
respondent delay damp cc clashes reliance damp reliance damp damp assessment reliance
with reliance injuries reliance applicant's injuries reliance damp remand injuries
reliance reliance disclose damp a damp allegedly assessment damp threshold where
```

Figure 20. Small Snippet of the generated output from Bi-LSTM.

9. CHALLENGES

Legal case summarization has several challenges, mainly because the legal texts are themselves complicated, and the cases vary a lot. These include:

Ambiguity and Interpretation: Legal language is often wordy and sometimes open to a number of different interpretations. Summarizers need to avoid this incoherence within their efforts at summaries, accuracy, and fairness.

Context Preservation: Summarization must also preserve the context and subtlety of the case in question. Over-simplification may close the door to misinterpretation. Privacy and Confidentiality: One of the most important issues that every attorney must consider is a significant challenge when representing sensitive or confidential legal documents without accidentally releasing sensitive information.

Diversity of Cases: The cases indeed range over a wide variety of topics and levels of complexity. Any summarization techniques should be able to adapt to this diversity.

10. LIMITATIONS

The present study, while contributing meaningfully to the growing body of literature on legal NLP, is subject to certain limitations that the authors duly acknowledge. The dataset of 4,178 cases, confined to Articles 3, 5, 6, and 8 of the ECHR, represents only a limited portion of the Court's broader jurisprudence and may not be readily generalisable to other legal corpora or jurisdictions, particularly those outside the European human rights framework. It is further to be noted that enforcing class balance by capping 350 cases per label, though methodologically necessary, imposes an artificial constraint on the natural distribution of cases, thereby restricting the representativeness of the training data.

With respect to data quality, the ground truth violation labels as assigned by the Court are reflective of nuanced and multi-layered legal reasoning, which surface-level textual features may not adequately capture, thus introducing a degree of noise into model supervision. Additionally, as a considerable portion of ECHR case documents are translated from French and other European languages, linguistic inconsistencies are inevitably present in the corpus. Pretrained English-only models such as BERT are not well-suited to handle such multilingual variation, and this is understood to be a contributing factor to the comparatively lower performance observed for BERT in the present study.

The issue of model interpretability and explainability is another area that the present work has not been able to address in sufficient depth. While Random Forest and Bidirectional LSTM with attention mechanism demonstrated superior predictive performance, neither model offers the level of transparency that would be required for responsible and accountable deployment in actual legal decision-support settings. The absence of explainability tools such as SHAP values or attention visualisation renders the model outputs difficult to audit, which is a concern of considerable importance in the legal domain.

It is also to be acknowledged that temporal factors governing legal outcomes have not been accounted for in the present study. Judicial interpretation of Convention articles evolves, and a static dataset may not adequately reflect such shifts, potentially affecting model reliability on more recent cases. Lastly, the possibility of algorithmic bias, particularly given the dominance of cases from specific jurisdictions such as Russia, Turkey, and Ukraine, as identified through topic modelling, remains an area that has not been formally evaluated and must be taken up as a priority in future research endeavours.

11. SUMMARY

This study presented a comprehensive NLP pipeline applied to the ECHR legal corpus across Articles 3, 5, 6, and 8, covering multi-label article classification, Part-of-Speech tagging, Named Entity Recognition, topic modelling, abstractive legal case summarisation, and judgment prediction. On classification tasks, Random Forest (accuracy: 0.64, macro F1: 0.63) consistently performed better in comparison with both Multinomial Naïve Bayes (0.59) and BERT (0.52). All three models performed better on violation classes than non-violation classes for Articles 3, 6, and 8, with Article 5 being a consistent exception, where non-violation classification outperformed violation classification across all three models. BERT's poor performance is linked to domain mismatch, token length constraints, and insufficient training data for proper fine-tuning, which partially contradicts [13], where BERT had scored better results on Chinese court corpora, indicating that transformer effectiveness may vary by jurisdiction and language.

In NER, spaCy's en_core_web_lg resulted in the most accurate entity identification in the large legal text. BERTopic combined with KMeans clustering revealed eight coherent thematic clusters, indicating an interpretable structure within the dataset. BART-large-CNN achieved a ROUGE-1 of 0.45 and ROUGE-L of 0.24 for legal case summarisation, and Bidirectional LSTM with attention (ROUGE-L: 0.17) performed significantly better than

standard LSTM (ROUGE-L: 0.08) in case outcome (judgment) prediction, indicating that the bidirectional context in modelling sequential legal text is useful.

12. CONCLUSION

This study introduced a multi-task NLP framework for ECHR legal case analysis by integrating multi-label article classification, topic modelling, case summarisation, and legal case outcome (judgment) prediction under a unified pipeline. By framing the problem as multi-label rather than binary classification, this research more accurately represents the complexity of the ECHR corpus.

The key findings confirm that Random Forest outperforms BERT in this low-resource environment, domain-specific setting, indicating the limitations of general-purpose transformers in the absence of domain adaptation. This hints at Legal-BERT as a promising direction for future research to focus on. The topic modelling demonstrated that unsupervised methods can surface legally meaningful structure without labelled supervision, as it provided geographically and thematically coherent clusters. In legal case summarisation, BART-large-CNN showed strong summarization performance, while the BiLSTM model with attention significantly outperformed the standard LSTM in judgment prediction.

Though listing these contributions, there are several limitations, such as the size of the dataset containing 4,178 cases. Though balanced across articles (classes), it represents only a narrow slice of ECHR jurisprudence and may not generalise to other jurisdictions and legal systems. The complexity of legal reasoning remains an open challenge, and the absence of full model interpretability, especially for Random Forest and LSTM, limits the transparency required for ensuring responsible and ethical deployment of such use cases in production to aid legal decision-support contexts.

Going further, studies should prioritise the adoption of domain-adapted legal language models trained on larger ECHR corpora. Integration of explainability systems to support judicial accountability is much needed, along with ethical evaluation of model outputs with respect to fairness and bias criteria before deploying in real-world use. This work serves as a foundational contribution to AI-assisted jurisprudence and provides a reproducible and extensible baseline for broader legal NLP research.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable.
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, "Predicting judicial decisions of the European Court of human rights: A natural language processing perspective," *PeerJ Computer Science*, vol. 2, p. e93, 2016. https://doi.org/10.7717/peerj-cs.93

[2] I. Chalkidis and D. Kampas, "Deep learning in law: Early adaptation and legal word embeddings trained on large corpora," *Artificial Intelligence and Law*, vol. 27, no. 2, pp. 171-198, 2019. https://doi.org/10.1007/s10506-018-9238-9

[3] K. Shirsat, A. Keni, P. Chavan, and M. Gosavi, "Study of deep learning-based legal judgment prediction in internet of things era," *International Research Journal of Engineering and Technology*, vol. 8, no. 4, pp. 1139–1142, 2021.

[4] X. Chen, Y. Wu, H. Chen, W. Lu, and Y. Ding, "Prediction of legal case outcomes using machine learning and natural language processing: A systematic review," *arXiv preprint arXiv:2207.00947*, 2022.

[5] D. Hendrycks, C. Burns, A. Chen, and S. Ball, "CUAD: An expert-annotated nlp dataset for legal contract review," *arXiv preprint arXiv:2103.06268*, 2021. https://doi.org/10.48550/arXiv.2103.06268

[6] N. Limsopatham, "Effectively leveraging BERT for legal document classification," in *Proceedings of the Natural legal Language Processing Workshop 2021*, 2021, pp. 210-216.

[7] R. Sil and A. Roy, "A novel approach on argument based legal prediction model using machine learning," in *2020 International Conference on Smart Electronics and Communication (ICOSEC), IEEE*, 2020, pp. 487-490.

[8] N. Xu, P. Wang, L. Chen, L. Pan, X. Wang, and J. Zhao, "Distinguish confusing law articles for legal judgment prediction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3086-3095.

[9] L. Xiao, J. Hu, Z. Liu, K. Tu, and M. Sun, "Lawformer: A pre-trained language model for Chinese legal long documents," *arXiv preprint arXiv:2106.08345*, 2021.

[10] H. Yıldız and S. Çakır, "Natural language processing in law: Prediction of outcomes in the higher courts of Turkey," *Expert Systems with Applications*, vol. 166, p. 114124, 2021.

[11] Y. Zheng, A. Guha, M. Anderson, M. Henderson, and D. Ho, "Law and the CaseHOLD dataset of 53,000+ legal holdings," *arXiv preprint arXiv:2105.03507*, 2021.

[12] S. H. Varshini, G. S. Varma, P. Prabhakar, and P. B. Pati, "Comparative analysis of legal outcome prediction with detailed and summarized text," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT), IEEE*, 2024, pp. 1-6.

[13] Y. Wang, J. Gao, and J. Chen, "Deep learning algorithm for judicial judgment prediction based on BERT," in *2020 5th International Conference on Computing, Communication and Security (ICCCS), IEEE*, 2020, pp. 1-6.

[14] S. Long, C. Tu, Z. Liu, and M. Sun, "Automatic judgment prediction via legal reading comprehension," presented at the China National Conference on Chinese Computational Linguistics, Springer, 2019.

[15] Y.-H. Liu and Y.-L. Chen, "A two-phase sentiment analysis approach for judgement prediction," *Journal of Information Science*, vol. 44, no. 5, pp. 594-607, 2018. https://doi.org/10.1177/0165551517722741

[16] S. Vats *et al.*, "Llms–the good, the bad or the indispensable?: A use case on legal statute prediction and legal judgment prediction on indian court cases," presented at the Findings of the Association for Computational Linguistics: EMNLP 2023, 2023.

[17] L. Liu, D. An, Y. Wang, X. Ma, and C. Jiang, "Research on legal judgment prediction based on Bert and LSTM-CNN fusion model," in *2021 3rd World Symposium on Artificial Intelligence (WSAI), IEEE*, 2021, pp. 41-45.

[18] L. Ma *et al.*, "Legal judgment prediction with multi-stage case representation learning in the real court setting," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 993-1002.

[19] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and Y. D. Al-Otaibi, "A hybrid CNN+ BILSTM deep learning-based DSS for efficient prediction of judicial case decisions," *Expert Systems with Applications*, vol. 209, p. 118318, 2022. https://doi.org/10.1016/j.eswa.2022.118318

[20] A. Li *et al.*, "Legal judgment prediction based on knowledge-enhanced multi-task and multi-label text classification," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 6957-6970.

[21] N. Z. Dina, S. D. Ravana, and N. Idris, "Legal judgment prediction using natural language processing and machine learning methods: A systematic literature review," *Sage Open*, vol. 15, no. 2, p. 21582440251329663, 2025. https://doi.org/10.1177/21582440251329663

[22] F. Ariai, J. Mackenzie, and G. Demartini, "Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges," *ACM Computing Surveys*, vol. 58, no. 6, pp. 1-37, 2025. https://doi.org/10.1145/3777009

[23] I. Karamitsos, N. Roufas, K. Al-Hussaeni, and A. Kanavos, "LegNER: A domain-adapted transformer for legal named entity recognition and text anonymization," *Frontiers in Artificial Intelligence*, vol. 8, p. 1638971, 2025. https://doi.org/10.3389/frai.2025.1638971

[24] B. Mao, "Exploring selective retrieval-augmentation for long-tail legal text classification," *arXiv preprint arXiv:2508.19997*, 2025. https://doi.org/10.48550/arXiv.2508.19997

[25] R. S. D. Oliveira and S. E. G. Nascimento, "Analysing similarities between legal court documents using natural language processing approaches based on transformers," *PloS One*, vol. 20, no. 4, p. e0320244, 2025. https://doi.org/10.1371/journal.pone.0320244

[26] A. Bhandari, P. Giriyan, P. Gawade, and A. Sahitya, "Evaluating transformer models for named entity recognition in Indian legal texts," presented at the 2025 3rd International Conference on Disruptive Technologies (ICDT), IEEE, 2025.

[27] A. Quemy and R. Wrembel, "ECHR-OD: On building an integrated open repository of legal documents for machine learning applications," *Information Systems*, vol. 106, p. 101822, 2022. https://doi.org/10.1016/j.is.2021.101822