Legal Decision Prediction and Classification of US Supreme Court Cases

Rym Mehdi
Muhammad arslan Ijaz
Clinton antony Rajasekar
Jaafar Saleh

EPITA Graduate School of Computer Science

Advisor
Professor Alaa Bakhti

Date

12/07/2023

Acknowledgements

We want to thank everyone who helped finish this research study.

First and foremost, we thank Professors Bill Manos and Alaa Bakhti for their great advice, persistent support, and important experience. Their guidance has shaped our research. We appreciate their wise advice, constructive criticism, and constant support.

Our advisors, provided important counsel, support, and mentorship. Their guidance and support helped us finish this project and progress academically. We appreciate their patience and dedication to our success.

We also thank our thesis committee for their insightful and constructive suggestions. Their skills and critical perspectives really improved our research.

EPITA provided the resources, facilities, and research environment needed to complete this project. The institution's support and possibilities helped us succeed.

We also thank our research colleagues and team members for their cooperation, stimulating conversations, and technical help. Their input has improved our research.

We also want to thank our family and friends for their constant support, understanding, and encouragement. Their confidence and motivation have helped us succeed.

Finally, we thank all participants who kindly gave their time, knowledge, and experience to this study effort. Their participation shaped our findings.

It's hard to thank everyone, but we're grateful for all who helped make this study aim a success.

Your direction, support, and efforts were invaluable.

Abstract / Executive Summary

This thesis addresses the problem of using natural language processing (NLP) techniques and the BERT model to assist judges with the classification of legal issues and the prediction of case outcomes. The dataset consists of 3,304 Supreme Court of the United States cases encompassing the years 1955 to 2021. This dataset contains detailed information that can be useful for NLP applications, whereas similar datasets frequently lack the incorporation of case facts.

This study's primary objective is to use NLP and BERT to construct a model capable of accurately classifying legal issues based on case titles and predicting case outcomes using case facts. Using BERT's contextualized representations, the model attempts to capture the nuanced semantic meaning of the case facts and use this information to make informed predictions.

The solution proposed entails pre-processing the dataset, extracting meaningful features from the case facts using BERT, and training a classification model. The model's performance on unseen data will be evaluated using standard metrics such as accuracy, precision, recall, and F1-score.

In addition, this thesis extends the solution by investigating the prognosis of judges' decisions based on their prior decisions. For each judge, separate models will be trained using the encoded case facts as input and the judge's decision as the objective variable. This strategy aims to provide judges with a greater understanding of how their decisions align with their historical patterns.

This research contributes to the fields of NLP and legal analytics by advancing the state-of-the-art in legal decision-making. Incorporating BERT and NLP techniques into the classification of legal issues and the prediction of case outcomes is a novel approach that could facilitate judges' decision-making. This study's findings have the potential to enhance the effectiveness and precision of legal analysis, which will benefit both justices and legal professionals.

1 Introduction

1.1 Motivations:
The research discussed in this thesis focuses on the urgent need for judges in the legal system to have access to effective decision-making support. Judges oversee applying and interpreting the law to settle legal disputes. Their judgements' impartiality and accuracy have a big impact on how justice is administered and how precedents are set in the law. It is essential to provide judges with tools that make decision-making easier and increase the potency of legal analysis.

The use of machine learning and natural language processing (NLP) techniques has opened new possibilities for the analysis and comprehension of textual data, including legal writings like court decisions and legal judgments. By utilizing these strategies,
By employing these techniques, we can extract valuable insights from immense quantities of legal data and equip judges with tools that aid in the categorization of legal issues and the prediction of case outcomes.

1.2 Context of the project:
This project's context resides at the intersection of natural language processing, machine learning, and the legal system. We intend to bridge the divide between textual data and judicial decision-making by employing NLP techniques. This research considers the unique requirements and challenges of the legal domain and strives to provide a comprehensive solution that improves the ability of judges to make decisions.

1.3 Objectives:
This research's primary objective is to develop NLP models, specifically utilizing the BERT (Bidirectional Encoder Representations from Transformers) model, to assist judges in two crucial aspects of their decision-making process: the classification of legal issues and the prediction of case outcomes. By accurately categorizing legal issues and predicting case outcomes based on case facts, judges can obtain valuable insights that can inform their decisions and enhance the efficiency of legal analysis.

In addition, this study seeks to broaden its scope by predicting the decisions of judges based on their historical patterns. By analysing prior decisions and identifying patterns, our models can offer judges additional insight into their own decision-making tendencies, thereby promoting consistency and fairness.

1.4 Overview of the Thesis:

Chapter 2 - Literature Review: This chapter provides an extensive review of the relevant literature in the field. It investigates prior research, surveys, and studies pertaining to legal judgment prediction, NLP techniques, machine learning models, and their applications in the legal domain. The literature review lays the groundwork for the research by highlighting existing knowledge, identifying knowledge deficits, and informing the design of the research.

Chapter 3 - State of the Art: In this chapter, the current state of the art in legal judgment prediction and NLP techniques is explored. It examines the most recent innovations, methodologies, and models employed in the field. In addition, the chapter investigates the limitations and difficulties of existing approaches, paving the way for the proposed system architecture.

Chapter 4 - System Architecture: This chapter presents the system architecture designed for the automated prediction of legal judgment outcomes. It describes the system's components, modules, and overarching architecture. The system architecture incorporates natural language processing (NLP) techniques, such as BERT, and machine learning models to classify legal issues, predict case outcomes, and provide judges with relevant prior cases.

Chapter 5 - Methodology: The methodology chapter discusses the research development and execution of the proposed system. It describes the phases and procedures involved in the system's development and evaluation. The methodology includes techniques for dataset pre-processing, feature extraction, model training, and evaluation. It provides insight into the methodology employed to ensure accurate predictions and trustworthy results.

Chapter 6 - Results: The following section presents and analyses the results of the system's implementation. It discusses the performance of NLP models and machine learning algorithms with respect to classifying legal issues, predicting case outcomes, and recommending pertinent prior cases. This chapter contains quantitative analysis, evaluation metrics, and findings discussions.

Chapter 7 - Conclusions:
The final section concludes the thesis by summarizing the main findings, discussing their implications, and pointing out the contributions to the field. It evaluates the extent to which the research objectives have been attained. In addition, the chapter proposes areas for future research and enhancements to the automated prediction of legal decisions.

2 Literature Review

The topic or area of concern for research is to provide help to the judges the automated prediction of legal judgment cases outcomes based on the supplied data and suggest them previous relevant cases and to predict category of the legal cases using the cutting-edge technology like Natural Language Processing (NLP). The NLP provides the help to build and use the model like Bidirectional Encoder Representations from Transformers (BERT) which is domain specific language model. Utilizing the already provided data set on the Kaggle which is ruling of the Supreme Court of the United States (SCOTUS) which is in public domain knowledge. According to different research papers to predict the verdicts for the judges most of them taking of machine learning and NLP using domain specific language model (Legal-BERT) which is already trained model to generate the legal verdicts. Publications are also taking look at many methods, including neural models, manually created features, cross-task dependencies, and label semantics, highlighting systems that produced cutting edge outcomes. Researchers has built different models like LSTM, ChatGPT-2, Rouge-3, BLEU including BERT and Legal-BERT to check the performance and accuracy of the models related to the text generation problems. For model Evaluation and performance metrics researchers are proposing and comparing evaluation metrics to assess the performance of predictive models. Regarding the gaps major conflicts, for the researcher the challenging part is to strike a balance between model interpret ability and predictive accuracy, Lack of Transparency in Judicial Decision-Making reconciling model outputs with the intricate reasoning of judges poses challenges. The ability of models to generalize beyond the training data and apply to different contexts or jurisdictions is uncertain. After having look at different articles and research papers we get to points the importance of specific features and variables in improving the predictive power of the models, we explore the strengths and weaknesses of each method and highlight the most commonly employed models and we found that BERT is best in that case, also highlight the gaps in knowledge, suggest new methodologies, or propose ways to improve the accuracy and interpret-ability of legal case predictions.

To thoroughly evaluate the state of the art in LJP and offer insights into its difficulties, techniques, and future directions, several surveys and research have been carried out. Zhu, Liu, and Sun (2022) provide a thorough summary of automatic LJP in their survey. The authors go over several LJP research subtasks, assessment measures, and systems. They look at many methods, including neural models, manually created features, cross-task dependencies, and label semantics, highlighting systems that produced cutting edge outcomes. The research also discusses difficulties in LJP, such as label imbalance, training data biases, and prison term prediction. In addition, the authors underline the importance of fairness, prejudice reduction, and ethical considerations, and they suggest future research areas which are much related to our topic in a sense to label the data and specific features.

According to Nishchal, Mohand and Taoufiq (2022) from (IRIT), Toulouse, France, a model to accurately predict the best probable decision of a legal case from the facts is desired. The used the different modelling techniques like Legal-BERT, LSTM-GRU and their combination to evaluate the performance and they found Legal-BERT as the best for the performance.

Cui, Liu, Wang, Chen, and Huang (2021) conduct yet another survey that examines LJP from the viewpoint of datasets, metrics, models, and problems. The authors offer suggestions for improving LJP datasets and neural network models after comparing benchmark datasets and experimental findings from various techniques. They emphasize how crucial it is for LJP systems to have elements for complicated reasoning, admissibility of evidence, and task-specificity.

Tal, Matthew, Nikolaus and Christopher (2022) conduct the researcher for nine text generation models and they found that most human-consistent model tested was GPT-2 and BERT as they are per-trained models on large set of data and provide high performance and accuracy. Another research paper is conducted by authors Mihai, Ana, and Horia in Romania where they used the Romanian based BERT model to predict the legal jury verdicts on large, specialized corpus. The authors suggest due to low resource languages and for highly specialized tasks, transformer models tend to lag more classical approaches (e.g., SVM, LSTM).

Mohammed Alsayed, Shaayan Syed, and Mohammad Alali, (2021) conducted research in USA using the same data set as we are going to use and performed predictions using the LSTM model and KNN and they suggest to focused on researching which source of Supreme Court cases would provide the most appropriate data for our purposes. In Addition, their research has gaps in terms of model performance and accuracy. By analysing the results of this study, we gained a better understanding of how the dataset could be modified to improve and train more accurate models.

So, the question is how our study is different from the already conducted research and surveys, in terms of that we are using per-trained BERT model to not only for verdict generation but also for classification of Legal cases categories and suggesting the previous verdicts related to problem.

To conclude, the articles offer a thorough summary of the current state of LJP research, including its obstacles and potential future paths in terms of prediction models. The authors stress the significance of model's accuracies for predicting the verdicts and about ethical issues in LJP and suggest lines of inquiry for next studies to tackle these difficulties. The well-structured studies offer a concise and in-depth summary of the subject.

3 State of the Art – Results/Findings/Discussion Analysis

In this segment, we depict different works modelling the legal content representations and executing calculations to foresee lawful choices in numerous law domains, such as Human Rights, Protected Law and Choice Forecast of Incomparable Courts, Assess Law, Mental Property, and Authoritative Law and Criminal Law, managing with corpora in numerous dialects such as Chinese (C), English (E), Farsi (Fa), French (F), German (G), Iranian (I), Philippine (Ph), Portuguese (P), Spanish (S) and Turkish (T).

3.1. Human Rights

A few activities have been proposed to study the legal corpus constituted by the ECtHR judgments. The centre of the think about is on information mining investigation strategies for foreseeing infringement or encroachments of the European Tradition on Human Rights (ECHR) articles. For occasion, in 2016, a gather of American and British law and computer science analysts created a prescient calculation for the ECtHR. This explore was the to begin with to prepare the SVM demonstrate for prescient examination of the literary substance of court choices.

The calculation anticipated whether there had been an infringement of Article 3 (forbiddance of torment and brutal and corrupting treatment), Article 6 (right to a reasonable trial), and Article 8 (regard for family and private life) of the ECHR. The calculation was created to foresee infringement of these articles solely based on data extricated from the content of the judgments concerning the depiction of the realities, appropriate laws and the parties' contentions. The strategy, truths, circumstances, and the important law of the cases were extricated from 584 choices. Moreover, they utilize a law infringement of a given article of the Tradition. The proposition accomplishes between 62% and 79% of exactness within the choices of the cases examined. An important conclusion was that the foremost pertinent data for accomplishing the results was the depiction of the realities included within the circumstances of the

case variable.

Within the same vein, Medvedeva et al. utilized 11532 ECtHR judgment reports to prepare an SVM direct classifier to foresee future legal choices. An expectation precision of 75% was gotten for the infringement of nine ECHT articles. The approach emphasizes the potential of ML procedures within the legitimate field. In any case, they appeared that anticipating choices of future cases based on past cases adversely impacts execution (normal accuracy run of 58–68%).

At last, the authors compare different standard ML algorithms on corpus extricated from the ECtHR in this exertion. The authors propose a total ETL pipeline to create the benchmark scenarios.
to meet this objective. This pipeline incorporates the Bag of-Words and TF-IDF archives vectorization, as well as the utilize of n-grams some time recently the double classification errand. A few ML algorithms were compared, counting Decision Tree (DT), AdaBoost with DT, Bagging with DT, Naive Bayes (NB) (Bernoulli and Multinomial), Ensemble Extra Tree, Extra Tree, Gradient Boosting, K-Neighbours, SVM (Linear SVC, RBF SVC), Neural Network (Multi-layer Perceptron) and Random Forest. The algorithm's performance was compared utilizing quality measures, such as the precision, F1-score, and Matthew's correlation coefficient.

3.2. Constitutional Law and Ruling Prediction of Supreme Courts

For the constitutional law and ruling prediction of the Supreme Courts, this study relies on sentences issued by the Supreme Courts of various countries, such as Turkey, the Philippines, the United States, Spain and France. The sentences used in the experiments are related to debates over property rights, adoption, freedom and so on. The reports, in a few cases, were taken from regulation websites, such as CENDOJ (for the case of selection in Spanish courts) or a compilation of sentences from sacred courts (Turkey) or Cassation (France). For instance, Ruger et al. developed an algorithm that can predict the individual votes of the nine justices as well as the final direction of the court's decisions, that is, confirmation or revocation in the U.S. Supreme Court. This experiment was conducted with information obtained from the court for two years (2002–2003), and 628 cases were analysed. The algorithm was based on a classification tree of the following six variables: the federal circuit in which the case originated, thematic area, type of plaintiff, type of defendant, ideological direction, and whether the constitutionality of a rule or practice was challenged. The results obtained were compared with the predictions made by a group of academics and lawyers. The following results were obtained: of 78 cases reviewed in 2002–2003, the algorithm could predict 75% of the court decisions and 66.7% of the individual votes, whereas the experts correctly predicted 59% of the decisions and 67.9% of the individual votes. Katz, Bommarito and Blackman [42] designed a model based on the Random Forest algorithm to predict the behaviour of the U.S. Supreme Court using a time-evolving Random Forest classifier on a corpus of 200 documents, with an accuracy of 70.2%.

Table 1. *Summary of ruling decision prediction in Constitutional Law and Supreme Courts cases, where E = English, Fa = Farsi, F = French, Ph = Philippine, S = Spanish, and T = Turkish*

No	Reference	Corpus Size	Lang	Model	Repr.	Metric	Min (%)	Max (%)	Year
1	[5]	628	E	DT	Binary	Accuracy	68	75	2004

2	[4]	120,506	E	LSTM + CNN	Embeddin g	Accuracy	88	92.05	2015
3	[7]	28,000	E	RF	Term frequency	Accuracy	70.2	71.9	2017
4	[1]	430	T	SVM	ag-ofWords, 2-gram, 3-gram	F-measure	63	90.3	2017
5	[2]	27,492	PH	SVM, RF	BoW, n-grams	Accuracy	55	59	2018
6	[8]	39,157	T	GRU, LSTM, BILSTM	TF-IDF	F-measure	56	87	2021
7	[11]	1884	S	ANN	Binary	Accuracy	61	85	2021
8	[3]	430		MLP	Embedings, TF-IDF, BoW	F-measure	60	98.7	2021
9	[4]	50,000	E	RoBERTa, RF, MLP, BIGRULWAN, AttentionXML, APLC_XLNet, and XTransformer	Sequence of words	F-measure	74.5	80.2	2021
10	[6]	-	E	HCNN	Accuracy	77.65	-	81.13	2021
11	[10]	3072	E	XGBoost, ANN, SVM, and RF	TF-IDF	F-measure	60	76	2021
12	[14]	123,361	F	JuriBERT	Tokenized	Accuracy	70.38	83.28	2021
13	[9]	120,506	E	CNN, LSTM	Tokenized	F-measure	79	93	2022

3.3. Private Law

Sulea Et Al. propose predicting the law category, court ruling, and time of the choice of the French Supreme Court. The dataset utilized for these three forecast assignments was a diachronic collection of decisions from the French incomparable court (Court de Cassation) in XML organize, containing 126,865 one-of-a-kind reports after the cleaning stage. They utilize Bag-of-Words, 2-gram, and 3-gram as inputs for a direct SVM classifier actualized in Sckit-learn for the distinctive errands.

They report the comes about utilizing accuracy, review, exactness, and F-measure.

They examined 200 archives for the eight diverse classes within the law forecast assignment. The F-measure gotten for this errand is 90.3%. For the administering choice forecast, the

SVM calculation gotten an F-measure of 97% and 92.7% when anticipating 6 and 8 classes, individually. The creators utilized 1-gram and 2-gram representations for the straight SVM within the final assignment of transient expectation, accomplishing 73.2% and 73.9% when foreseeing 7 and 14 classes, separately. Alghazzawi et al. propose a determining administering choice employing a cross breed neural arrange show combining a long, short-term memory (LSTM) organize with a convolutional neural organize (CNN). The creators utilize 120,506 cases with 27 highlights from the US Preeminent Court administering. They partition the corpus into 80%, 10%, and 10% for training, approval, and testing, respectively. Furthermore, they utilized 10-fold cross-validation to choose the leading demonstrate. Douka et al. propose the JuriBERT pre-trained structure utilizing 123361 records from the Court of Cassation and Légifrance. The creators utilize distinctive forms of JuriBERT to classify lawful content into eight classes among the chambers and segments and to classify legitimate content into five categories. The precision ranges from 79.9% to 83.28% and 70.38% to 72.09%, separately. The work of Sivaranjani, Jayabharathy, and Teja predicts the Indian incomparable court choice on request cases employing a Progressive Convolutional Neural Organize (HCNN). The creators utilized the sentences spoken to utilizing Word2Vec between 2000 and 2019 to prepare the Progressive CNN, accomplishing a precision extending from 77.65% to 81.13%.

To the finest of our information, few reports have been found that report the utilize of machine learning and profound learning connected to private law. Undoubtedly, two fundamental commitments were found in this setting. Li et al. propose a Markov Rationale Systems (MLN) likelihood demonstrate to foresee the legal choice of sepa rate cases. The creators utilized 695 418 archives from China Judgments Online (China Judgments Online: wenshu.court.gov.cn/, gotten to on 16 September 2022). From these records, creators extricate affirmation of truths, articles, and legal choices to prepare the MLN. They seem anticipate the likelihood of an allowed separate (89.7%) and the offended party paying the expenses (87.06%), with an F1-measure between 73.58% and 77.74%. Other papers confront the issue of course limits between the prediction of one lesson instep another course within the classification issue. In this way, the creators propose a two-layered progressive fluffy calculation connected to a dataset containing Iranian standard contracts between a manager and a temporary worker gathered into three classes: building, semi-building, and non-building. Sometime recently applying the calculation, the creators performed a include choice step utilizing the normalized entropy degree. Afterward, the corpus was utilized as input to the calculation to extricate fluffy rules, and the precision was measured utilizing the 10-fold cross-validation technique. The creators handled a collection of court choices in French from the Régie du Logement du Québec (RDL) on genuine domain law case. Undoubtedly, on a corpus made up of 981,112 choices issued from 2001 to 2018 by 72 judges in 29 Quebec courts, the creators connected NLP apparatuses to uncover the inclinations that can impact expectation tests. The creators separated the decisions into

two wide categories: Proprietor versus Occupant (LvT)
and Inhabitant versus Proprietor (TvL). In this test, they utilized the
FlauBERT dialect demonstrate. They overseen to recognize more than a dozen
characteristics of each choice and distinguish inclinations contained within
the information sets, such as, for illustration, that proprietors tend to sue their inhabitants, with more likelihood of victory. The forecast comes about are 93.7% and 85.2% in LvT cases and 84.9% and 74.6% for TvL cases, separately.

3.4. Tax Law, Intellectual Property, and Administrative Law

This section examines administrative, intellectual property, and tax legislation. Resolving tax disputes in nations like Germany, Brazil, and the Netherlands uses corpus linguistics for text classification in these fields. The works on intellectual property deal with disputes over domain names.
The classification of legal consultations was tested using various deep neural network types by the authors. Legal query comprehension, as the authors note, is a challenging topic that requires the simultaneous solution of two Natural Language Processing (NLP) tasks:
(i) determining user intent and (ii) detecting entities in inquiries. To examine each type of deep neural architecture separately, the authors also utilized recurrent neural networks, long short-term memory (LSTM), convolutional neural networks, and gated recurrent units (GRU). The models were contrasted with rule-based methods and machine learning (ML) techniques. These trials' findings demonstrate how challenging it is for DNNs to respond to lengthy queries.

Criminal Law:

For managing judicial files in criminal law, evaluated research employed text mining and supervised machine learning models, specifically for handling written criminal penalties in a way that enables case identification. like how they are classified. According to data gathered from criminal cases involving the murder of the Delhi District Court, the authors of [54] offer a 7-step approach to categorize the accused person's acquittal and conviction (classes). Thus, 86 examples were normalized with the min-max method and vectorized with the Bag of Words strategy.

Then, techniques like SVM, kNN, CART, and NB were employed. The model's performance is then evaluated using leave-one-out cross-validation, which yielded an F1 score of 86-92% and accuracy of 85-92%. Seron and Ferreira both conducted studies in Portuguese [55]. They used a dataset of 1562 murders and corruption in Brazilian court decisions to compare the Logistic Regression (LR), Latent Dirichlet allocation (LDA), KNN, Regression Tree (RT),

Gaussian NB (GNB), and SVM algorithms. Depending on the algorithm and court document, they got an F-measure score that ranged from 78% to 98%.
A document classification, clustering, and search methodology based on neural networks was created by Chou and Hsing [56] to assist law enforcement departments in handling written criminal judgements more effectively. As a result, judges can examine related instances and consider a variety of factors before passing judgment. 210 written criminal sentences from Taiwan's judiciary were utilized as the corpus for training and testing the models. Homicide, sex crimes, drug-related corruption, computer crime, theft, and fraud were the seven categories of crime that were chosen. 100 primary words with the highest usage frequency were chosen from a list of 2604 keywords to summarize the crucial terms used in the trials and documents.
There were 140 sample documents used for the neural network. The seven distinct criminal categories were divided into word segments, which were then expanded to 251 keywords. Based on the findings, the precision achieved in the training samples—which utilised all the written vectors' segments but weren't weighted as BPN inputs—reached 94%. Similar results were obtained with a model that employed all the written vectors' segments, but weighted vectors were equally used as BPN inputs. This model's accuracy was 67%. India has also engaged in text mining.

From the police department data that was downloaded from official pages, Kaur et al. [57] constructed a corpus. The K-Means clustering technique and KNN were employed as classifiers. The following cities' crime rates were assessed using these algorithms: New Delhi, Andhra Pradesh, Jammu, Kashmir, Daman, Diu, Jharkhand, Arunachal Pradesh, and Nhaveli. Delhi and Jharkhand were found to have the highest crime rates among these cities. Dehli's outcome can be explained by the city's multicultural character, the scarcity of preventive measures, and the ineffectiveness of law enforcement. The study concludes that Jharkhand's high crime rate is caused by the region's lack of police presence. The authors evaluated the performance of the KNN method using Root relative squared error (RRSE), and it scored 67.92%. According to Bingfeng et al. [58], accusations of crimes including fraud, theft, or homicide would be made. They made advantage of both judicial rulings and legal provisions found in statutes. To do this, they gathered legal documents from the Chinese government's website and extracted descriptions of the facts, legal provisions, and imputation allegations.

The corpus included 50,000 documents that were chosen at random, 5000 for validation, and 5000 for testing. The researchers suggest using a neural network to forecast the charges that will be included in the indictment as well as the pertinent legal texts that will be used to support those charges. The outcomes demonstrate the model's suitability for both objectives. The experiment made use of a BI-GRU sequence encoder, an SVM classifier for extracting legal articles, and SGD for training. CAIL2018, a tool that summarizes more than 2.6 million criminal cases on the website of the Chinese Supreme People's Court, is

presented by Xiao et al. [59]. This tool (PJL) performs NLP tasks using DL methods in conjunction with neural networks to anticipate trial outcomes. In their experiments, they combined the following three techniques: SVM with a linear kernel and TF-IDF to extract word features and train the classifier. Additionally, Fast Text was used to categorize texts using Hierarchical SoftMax and n-grams. CNN was also utilized to code fact descriptions and classify text. In terms of accurately forecasting imputation accusations and legal articles, many findings were made. Using over 1.2 million documents from criminal cases in China, Zhong et al. [60] compare text representations like TF-IDF and Embeddings with various classification algorithms like SVM, CNN, and HLSTM.

The accuracy that the authors utilized to gauge the effectiveness of the algorithms ranged from 38.3% to 94.4%. Furthermore, Li et al. [61] assess the target case's judgment rationality based on the judgment outcomes of situations that are like it. The 41 418 Chinese documents utilized as the input for a GRU were represented by the authors using Doc2Vec. The F-measure, which has a performance range of 73.6% to 78.7%, was utilized by the authors to assess the performance of the proposal.
Choosing cases on demand using a Progressive Convolutional

4 System Architecture

4.1 Anticipated System Architecture:

Two important machine learning models, LSTM (Long Short-Term Memory) and BERT (Bidirectional Encoder Representations from Transformers), are at the core of our suggested system design. These models are chosen based on their distinct qualities, which make them particularly well-suited to the duty of assisting judges in classifying cases and making decisions considering prior instances and their rulings.

LSTM (Long-Short-Term Memory) Model: A Recurrent Neural Network (RNN) with an LSTM has been developed to learn and recall data over lengthy sequences, solving the long-term reliance issue. LSTM is a perfect option for processing the sequence of events inside a case, the court processes, and the timing of the crime because of its potential to remember over an extended period. The LSTM can learn from and predict future events using this sequence data, which may be described as a time-series dataset. The LSTM's inherent nature, which is significant, enables it to capture patterns across time, such as a certain series of events leading to a particular verdict. When forecasting outcomes, this pattern recognition over temporal data can offer critical insights and make a big difference.

Transformers' Bidirectional Encoder Representations (BERT) Model Contrarily, BERT is a machine learning model based on transformers created especially for Natural Language Processing (NLP). It is intended to comprehend a word's context by looking at its surroundings (words that come before and after it), making it bidirectional.

This feature is essential to our program because court proceedings and legal papers frequently contain jargon and sophisticated linguistic structures. BERT will significantly contribute to forecasting the category of a case and verdict by being able to comprehend the context of these terms.

5 Methodology

5.1 Anticipated Methodology:

These models are then put into practice in a variety of ways that are intended to optimize their usefulness for making decisions.

Data Gathering: The procedure starts with gathering case data, which includes information about the offense, the court proceedings, and the judgement. Since the data is labelled, it is possible to determine the category and the outcome for each case. This labelled data is priceless since it serves as the foundation for our models' learning. The system's capacity to create precise predictions can be considerably influenced by the nature and volume of this input.

Data Preprocessing: Preprocessing is the next step after collecting the data. Cleaning up the data to get rid of any discrepancies or missing numbers is a crucial step. By following this procedure, we can guarantee that the data that goes into our models is as accurate and consistent as possible, avoiding any incorrect forecasts brought on by faulty data.

The data is cleaned before being translated into a format that the LSTM and BERT models can understand. The text is tokenized (divided into words or phrases that the models can understand), the sequences for LSTM are padded (to maintain consistency in input length), and word embeddings (words are transformed into numerical representations) are created.

LSTM and BERT Models: The pre-processed data is split into a training set and a test set. After that, utilizing the course of events and the decision, the LSTM model is trained on the training set. The LSTM can identify patterns and anticipate verdicts with accuracy because it can comprehend the temporal dynamics of the case data.

The BERT model is also trained using the training set. The BERT model, on the other hand, is more concerned with the case data's context, such as the circumstances of the crime and the defence’s arguments. After both models have been trained, their performance is assessed on the test set to determine whether any revisions are required.

Fusion of Outputs: The predictions from the LSTM and BERT models are combined to provide the final verdict prediction. More weight may be given to a model's prediction in the decision depending on how consistently accurate it is relative to the other models.

The system's goal is to aid in the process rather than take the place of a judge's power of decision-making. The system will produce a predicted verdict based on the learned patterns from the LSTM and BERT models and the data from prior cases once the specifics of a case are provided via a user-friendly interface.

5.2 Updated Methodology

An updated methodology will be provided during the Action Learning Week. It will provide the opportunity to evaluate the LSTM and BERT models' performance, the effectiveness of data preprocessing methods, and the accuracy of the fusion of outputs.

Any improvements or changes will be documented in this updated methodology. The focus will be on areas that require refinement, including more advanced techniques for model tuning, different data preprocessing methods, or even the exploration of complementary models to the existing LSTM and BERT models.

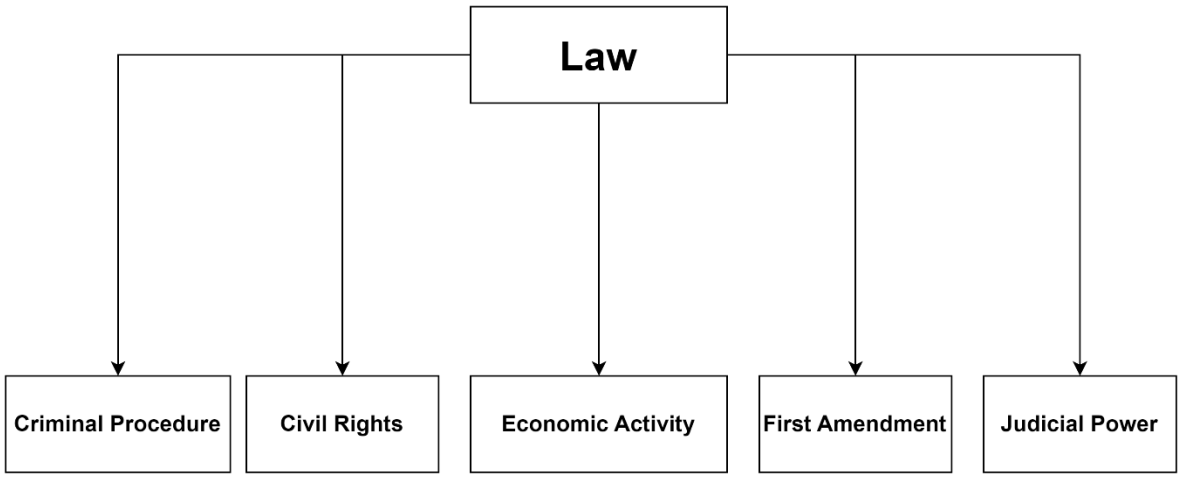

***Figure 1**: Types of issue areas from data set*

6 Results

The relevant dataset was acquired from the Supreme Court of the United States (SCOTUS) using reputable sources such as Kaggle. The dataset was then cleansed and pre-processed to ensure its suitability for the study's aims.

Model Selection and Experimentation: The BERT model, along with its variants such as Legal-BERT, was selected as the primary model for predicting legal case outcomes based on the findings of the literature review. Other models identified in the literature, such as LSTM and GPT-2, were also considered for comparison and experimentation.

Training and Evaluation of Models: Using the collected and preprocesses dataset, the selected models were trained and fine-tuned. They were trained to predict the outcomes of legal cases and to categorize legal case types. The F1 score, accuracy, precision, and recall were used to assess the models' performance.

Analysis and Comparison of Performance: The performance of the trained models was examined and compared. The accuracy and efficacy of the models in predicting the outcomes of legal cases and classifying legal cases were evaluated. To validate or expand existing knowledge in the discipline, the results were compared with those from the reviewed literature.

Efforts were made to resolve the identified research and knowledge gaps. Proposed methodologies and techniques aim to enhance the accuracy and interpretability of models. A balance was struck between model interpretability and predictive accuracy, and model outputs were reconciled with the complex reasoning of judges.

Documentation and Reporting: The entire research procedure was meticulously documented, including data acquisition, model selection, training, evaluation, and analysis. This documentation is an integral element of the thesis, detailing the research methodology and findings.

6.2 Development updated. (Submitted end of Action Learning Week)
Subsections to be added by you as needed.

7 Conclusion

This study's literature review shed light on the current state of Legal Judgment Prediction (LJP) research, providing valuable insights into the field's extant methodologies, challenges, and opportunities. This conclusion emphasizes the significance of the literature review's main findings and contributions to the current project.

The significance of precision in predicting legal case outcomes has been emphasized throughout the review of relevant literature. Numerous models, including Legal-BERT and LSTM-GRU, have demonstrated their ability to achieve high prediction precision. By employing a pre-trained BERT model, this research endeavour seeks to improve the precision and dependability of legal judgment prediction.

Ethical considerations have emerged as an integral component of LJP research. The literature review has highlighted the importance of addressing potential biases in training data and ensuring impartiality and reducing prejudice in prediction models. This initiative recognizes the importance of these ethical considerations and seeks to contribute by developing models that are not only accurate but also fair and transparent in their decision-making processes.

In legal judgment prediction, transformer models, particularly BERT and Legal-BERT, have demonstrated promise. For low-resource languages and highly specialized tasks, it is acknowledged that transformer models may still lag traditional approaches. By investigating

the performance of these models on the Supreme Court of the United States (SCOTUS) dataset, this research endeavour aims to shed light on their applicability in the legal context.

Knowledge deficits present opportunities for further investigation. Exploration must focus on striking a balance between model interpretability and predictive accuracy, reconciling model outputs with the complex reasoning of judges, and ensuring the generalizability of models across contexts and jurisdictions. This initiative intends to address these gaps by proposing methodologies to enhance interpretability and performance, thereby contributing to the field's advancement.

In conclusion, this research endeavour aims to improve legal judgment prediction by leveraging advances in NLP and BERT models. The project seeks to develop accurate and interpretable models to aid judges in predicting case outcomes and classifying legal cases. The literature review has set the groundwork for this endeavour by providing insights into existing research, emphasizing the importance of precision and ethical considerations, and identifying avenues for future research.

Based on this literature review, the subsequent chapters of this thesis will examine the system architecture, methodology, and experimental results. By incorporating the insights garnered from the literature review and applying them in a practical setting, this research project aims to contribute to the advancement of automated legal judgment prediction and provide judges with valuable assistance in making decisions.

Overall, this study seeks to bridge the gap between legal research and cutting-edge NLP techniques by proposing a novel approach to predict legal case outcomes while adhering to ethical considerations and promoting legal domain transparency.

Bibliography/References

Must use APA style and preferably included in the MS Word bibliography as taught in class.

TITLE OF THE PAPER	YEAR	No. OF CITATIONS	KEYWORDS	WHY DO WE NEED THIS?	CUTTING EDGE POINT / BREAKTHROUGH
Legal Judgment Prediction: A Survey of the State of the Art	2022	7	automatic legal judgment prediction, natural language processing, research	Different subtasks of LJP using BERT	

			challenges, multiple jurisdictions,		
jurBERT: A Romanian BERT Model for Legal Judgement Prediction	2021	8	Natural Language Processing, transfer learning, state-of-the-art results, low resource languages, Romanian BERT model	Using the model BERT in details fore dataset	same model that we are going through -problem solving
A Survey on Legal Judgment Prediction: Datasets, Metrics, Models and Challenges (arxiv.org) ("A Survey on Legal Judgment Prediction: Datasets, Metrics ... - NASA/ADS")	2022	0	Natural Language Processing, state-of-the-art results, low resource languages, BERT model	Have comparision between different models	legal judgment prediction using Bert and other models with comaprisions
JUSTICE: A Benchmark Dataset for Supreme Court's Judgment Prediction | Papers with Code	2021	2	Natural language processing, Legal text classification, Benchmark dataset	Provides code and Dataset Using BERT model	legal judgment prediction using Bert using the dataset
Testing the limits of natural language models for predicting human language judgments	2022		Linguistics, Sentiment Analysis, Language Translation, Performance Evaluation, Benchmarking.	Humans process language and make judgments, leading to advancements in fields such as cognitive science and linguistics.	natural language processing and machine learning to improve the accuracy and robustness of language models
"Effect of Hierarchical Domain-specific Language Models and Attention in the Classification of Decisions for Legal Cases" ("Effect of Hierarchical Domain-specific Language Models and Attention in ...")	2022	8	Classification, BERT, NLP, Legal	For classification of legal cases	Classification using Bert model

References

Masala, M., Iacob, R., Uban, A. S., H., Rebedea, T., & Popescu, M. (2021). jurBERT: A Romanian BERT Model for Legal Judgement Prediction. University Politehnica of Bucharest, University of Bucharest, BRD Groupe Societe Generale.

Alsayed, M., Syed, S.& Bodala, H. (2021). JUSTICE: A Benchmark Dataset for Supreme Court's Judgment Prediction. University of Southern California.

Golan, T., Siegelman, M., Kriegeskorte, N., & Baldassano, C. (2023, page46). Evaluating the limits of natural language models for predicting human language judgments. Zuckerman Mind Brain Behavior Institute, Columbia University, Department of Cognitive and Brain Sciences.

Prasad, N., Boughanem, M., & Dkaki, T. (2022, p 43). Effect of Hierarchical Domain-specific Language Models and Attention in the Classification of Decisions for Legal Cases. ("Effect of Hierarchical Domain-specific Language Models and Attention in ...") Institut de Recherche en Informatique de Toulouse (IRIT), Toulouse, France. ("Institut de recherche en informatique de Toulouse — Wikipédia")

Source : Zhu, X., Liu, Z., & Sun, M. (2020). A Survey of Automatic Legal Judgment Prediction. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). ("Proceedings of the 58th Annual Meeting of the Association for ...")

Cui, J., Liu, Z., Wang, S., Chen, M., & Huang, Y. (2021). A Survey on Legal Judgment Prediction. Preprint submitted to Elsevier.

Ruger, T.W.; Kim, P.T.; Martin, A.D.; Quinn, K.M. The supreme court forecasting project: Legal and political science approaches to predicting supreme court decisionmaking. Columbia Law Rev. 2021, 104, 1150–1210.

Katz, D.M.; Bommarito, M.J.; Blackman, J. A general approach for predicting the behavior of the Supreme Court of the United States. PLoS ONE 2020, 12, e0174698.

Sert, M.F.; Yıldırım, E.; Ha˸slak, ̇I. Using Artificial Intelligence to Predict Decisions of the Turkish Constitutional Court. Soc. Sci. Comput. Rev. 2021, 93, 08944393211010398

Alghazzawi, D.; Bamasag, O.; Albeshri, A.; Sana, I.; Ullah, H.; Asghar, M.Z. Efficient prediction of court judgments using an LSTM+ CNN neural network model with an optimal feature set. Mathematics 2022, 10, 683.

Muñoz Soro, J.F.; Serrano-Cinca, C. A model for predicting court decisions on child custody. PLoS ONE 2021, 16, e0258993

Song, D.; Vold, A.; Madan, K.; Schilder, F. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. Inf. Syst. 2021, 106, 101718.

Sivaranjani, N.; Jayabharathy, J.; Teja, P. Predicting the supreme court decision on appeal cases using hierarchical convolutional neural network. Int. J. Speech Technol. 2021, 24, 643–650.

Appendices

Figures:

Figure 1: Types of issue areas from data set 17

Tables:

Table 1. Summary of ruling decision prediction in Constitutional Law and Supreme Courts cases, where E = English, Fa = Farsi, F = French, Ph = Philippine, S = Spanish, and T = Turkish 9