International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Prejudge: A Predictive Analytics System for Crime and Legal Judgments

Aastha Budhiraja[1], **Kamlesh Sharma**[2]

Submitted: 05/02/2024 **Revised**: 13/03/2024 **Accepted**: 19/03/2024

Abstract: The recent era has seen a substantial inflow of legal documents in the electronic format. Given the fact that data mining can be employed in the world of textual data to extract relevant knowledge, it is being prominently exploited in the domain of criminology and legal matters. With increasing crime rates day-by-day, it has become essential to readily impart justice to the victims. It takes a considerable amount of time for the lawyers to go through previous judgments for their research. The judicial process can be accelerated by decreasing the time spent on research work. Smart legal systems have enormous potential for providing significant insights to the legal community and the general public through the use of legal data. As a result, these systems can assist in the analysis and mitigation of a variety of societal concerns. By extracting numerous things from legal decisions, such as dates, case numbers, reference cases, person names, and so on, this work takes the first step toward realizing a smart legal system. The major research issues in the area of applying machine learning in jurisprudence are information extraction and analysis of legal texts. This study proposes an Machine Learning based framework to improve the user's query for retrieval of precisely relevant legal judgments in order to overcome these limitations. This work has been carried out in order to act as an aid to the legal advisors and the lawyers in framing arguments to make strong standpoints based on predictions given on their case pertaining to previous judicial outcomes for similar such cases. Logistic regression-based classification enables efficient retrieval and prediction by allowing inferences based on domain knowledge collected during the dataset development. According to empirical results obtained, the proposed methodology generates finer results than other traditional approaches.

Keywords: *Citation Analysis; Data Classification; Information Retrieval; Legal Domain; Natural Language Processing; Prediction; Similarity Search.*

1 Introduction

India has seen the rapid digitization of district and state courts during the previous decade. Many advances in this field have resulted from advancements in technology and computing power. Artificial Intelligence (AI)-powered technologies are becoming smarter, more efficient, and more accessible every day. This work shows how the implementation of AI technologies can converse with the end-user, comprehend the issue, and offer assistance by searching millions of solved instances. Artificial intelligence (AI) has gotten a lot of attention in the legal world as a method to save costs, improve access to justice, speed up document review, and eventually replace all people with robotic counterparts. This work aims to apply artificial intelligence (AI) into legal matters in order to make current jobs simpler.

In today's litigation, attorneys are frequently confronted with an overwhelming quantity of papers to evaluate and generate during the course of a case. Legal teams may be needed to produce millions of papers to opposing parties or regulators in large-scale litigation. Typically, these same teams must sift through a sea of documents to locate supporting evidence for their own claims. As more and more information are kept electronically, the expenses of manually examining papers have largely escalated. This procedure needs a significant amount of funding. Companies pay millions of dollars every year to provide relevant electronically stored records for legal proceedings. To react to regular discovery requests, attorneys spend endless hours analyzing documents, and these large expenses are ultimately passed on to clients. In the actual world of lawsuits, verdicts are made up of a number of sub-tasks that must be considered in their whole. A comprehensive review of data mining techniques that are used to forecast crime and make decisions is presented in this research. The phrase “legal research” refers to the act of locating and extracting relevant information from historical case records to assist in legal decision-making. The method is now primarily manual, though classic technologies such as keyword searches are sometimes utilized to speed it up.

Intelligent justice technology has gotten a lot of attention lately to check the existing limitations, because of the rapid growth of Natural Language Processing Technology (NLP). Natural language processing technologies can help lawyers speed up duties like legal research and contract

[1] *Aastha Budhiraja, Research Scholar, MRIIRS, Faridabad,121006, India.*
[2] *Dr. Kamlesh Sharma, PhD, MRIIRS, Faridabad,121006, India*

evaluations. Natural language processing, or NLP, is the process by which software understands written words. The field of automated court decision prediction is still in its infancy, and there are several research avenues to pursue. A vast amount of research links the challenge of multi-label text classification to crime prediction and related legal prediction problems. As machine learning technology advances, several researchers are attempting to employ machine learning models to complete judicial decision-making duties. At this point, the basic method is to manually gather the elementary characteristics of the crime fact text, model the problem, and then classify it using a classification model. Machine learning technology has considerably increased the accuracy of crime prediction and associated legislation prediction in this sector. The model, however, is only usable in particular circumstances due to the requirement to manually extracting features, and its generalization ability is limited. Intelligent justice is not often practiced. NLP might be used by legal practitioners to speed up document evaluation. You might, for example, run a contract through a natural language processing tool to help identify particular provisions, double-check the legal wording, and ensure the contract has all clauses necessary to comply with your standard operating procedures.

Unsupervised machine learning is used in the legal sector to evaluate text. For example, you might use an unsupervised machine learning algorithm to evaluate third-party contracts in order to utilize standard wording for limited liability provisions. The clustering algorithm will locate liability-related clauses and wording in other terms throughout a contract, making it easier (and faster) for a human to evaluate the data. Clusters may be used to not only presort information for individuals but also to provide training data for supervised machine learning algorithms. Multiple supervised and unsupervised algorithms are frequently connected together in practice to create a prediction or produce an output.

Artificial Intelligence (AI) developments, particularly in Natural Language Processing (NLP) and Machine Learning (ML), have recently provided us with the ability to automatically analyze legal texts in order to construct prediction models for court outcomes based on the semantics of law and case texts. Previous research on forecasting judicial decisions, on the other hand, has primarily relied on non-textual data. It's a strategy that might be used to provide appropriate judgment recommendations, such as charges, applicable law articles, and jail terms.

In this paper, we have put forth an empirical investigation on classification and prediction algorithms to predict judgments on cases based on historical records of similar such cases. This study has been taken up on a customized dataset curated using various judgment documents taken from Indian judicial websites. These textual documents have been pre-processed to make them qualified for further operation using NLP techniques. Once pre-processed Machine learning techniques have been implemented on these documents to extract features of interest and then exercise Logistic Regression on it for final classification and prediction. The effectiveness of the system design presented assures that the shortcomings in such approaches have been addressed, and model efficiency has been improved. In addition, based on the results of our comprehensive and comparative tests, we assess the advantage and capacity of each model by adding our novelty into it and explain the feature selections that account for their success.

2 Literature Review

We discovered just a modest amount of previous research on computerized text classification of legal documents. Support vector machines (SVMs), for example, have been used to categorize legal documents such as court docket entries [14] and non-English legal writings [13]. Despite the fact that our research looks at the application of machine learning to a corpus produced in the legal environment, we concentrate on categorizing legal opinions with very basic pre-processing. For example, before utilizing an SVM to classify texts using human-selected characteristics and labels, the Nallapati and Manning [14] method go through multiple rounds of pre-processing.

Brüninghaus and Ashley [15] describe IBP (Issue-Based Prediction), a multi-strategy approach that blends case-based and model-based reasoning for an interpretative CBR (Case-Based Reasoning) application that predicts legal case outcomes. It first employs an ad-hoc domain model to identify the issues posed in the case (which they refer to as a weak model), and then it reasoning using cases to reconcile contradictory evidence connected to each issue in the second phase. IBP employs evidentiary inferences and uses symbolic reasoning to determine the relevance of situations. Experiments with a set of historical examples demonstrate that IBP's predictions outperform those produced using its weak model or just cases. In comparison to typical inductive and instance-based learning algorithms, the authors claim that their method is more accurate.

The application of machine learning in construction lawsuit situations is described by Tarek and Kandil [16]. They use machine learning (ML) models to provide an automated litigation result prediction strategy for diverse site condition (DSC) conflicts. This paper compares the performance of three machine learning techniques, namely support vector machines (SVMs), nave Bayes, and rule induction, as well as neural network classifiers (decision trees, boosted decision trees, and the projective adaptive resonance theory) to develop the proposed method. The models were trained and evaluated on 400 DSC cases that were submitted between 1912 and 2007. The model's projections are based on important legal parameters that influence DSC verdicts in the construction sector. Among the nine ML models constructed, the third-degree SVM polynomial model fared the best, with a prediction precision of 98 percent.

Machine learning approaches were used to identify robbery and intimidation instances in [22]. Using 21 legal factor labels that were manually established, anticipate the statement. More recently, Aletras et al. [17] attempted to anticipate European Court of Human Rights rulings by using textual data such as N-grams and subjects to build Support Vector Machine (SVM) binary classifiers. Sulea et al. [18], [19] employed a linear SVM classifier to predict French Supreme Court legal areas and case judgments. For feature detection, Boella et al. [20] utilized the terms frequency-inverse document frequency (TF-IDF) and information gain. After that, an SVM classifier was created to determine the appropriate domain to which the provided legal text belonged. Liu and Chen [21] classified the judgment text using an SVM system based on relevant law articles, sentiment analysis of criminal facts, and sentence length. Despite the fact that these initiatives make full use of the supervised learning approach, they have a scalability difficulty since they rely largely on feature design and manual annotation.

In conclusion, past research has improved various facets of the legal judgment prediction problem. Nonetheless, learning adequate semantic representations from diverse portions of a case description and performing the complete classification and Prediction task in a cohesive framework remains a difficulty.

3. Problem Identification

The goal of this article is to build a smart legal system based on a given legal corpus that allows for rapid access to important legal judgments.

The two basic and phenomenal terms of Law and Mankind are counterparts of each other. With the origin of human civilization, there began the necessity of a judicial system. In concern to the present society, an increase in the number of legal affairs has been witnessed, which proportionally escalates the demand for justice among the masses. While it has been noticed that most of the time goes into background research to form a sound case, it simultaneously increases the time taken to reach a judgment. Lawyers essentially spend a considerable amount of time reading the previous judgment papers, examining historical data related to their cases and drafting thoughtful proceedings. All of this can lead to a delayed judgment, eventually affecting the victim as well as having a consequential impact on other cases. Lawyers form their judgments about a case outcome based on complex cognitive processing steps which are established on the lines of experience in their law profession and intuitions which are difficult to define and chart out abstractly.

Given the amount of text data that has been flowing from judicial word lately can be used to develop algorithms that make the task of lawyers easier for research work. Law can utilize huge volumes of legal data (e.g., legal invoices, activities, and historical results from litigation cases) to mine for patterns that indicate the anticipated outcome of a freshly contested case through case outcome prediction. Machine Learning gives these algorithms the ability to learn/discover relevant decision-making patterns in rich data settings automatically. With vast volumes of digital data gathered every day by law firms, inducing finding patterns is a perfect fit for machine learning. At the same time, the lack of a universal model for process flow and its accompanying data definitions that can be adapted to all case litigation circumstances makes introducing machine learning to law difficult. There's also the issue of comprehending how lawyers attempt to resolve litigation issues from a cognitive standpoint (e.g., their reasoning process).

Through this work, it has been tried to design a smart system that can help in predicting outcomes of cases based on feature extraction and classification performed on previous legal records by employing Machine Learning techniques. The model architecture created throughout this process may be utilized to improve the user query and, as a result, get more relevant judgments that meet the user's needs. The retrieved judgments are utilized to create a document summary that will assist the user in quickly and effectively understanding the associated case histories. This will be helpful to attorneys when dealing with new cases, as well as judges who want to publish fresh case decisions. We've suggested a new structural framework for building an ontology that enables the depiction of complicated legal judgments.

4 Crime Data Classification Techniques and Approaches

We discovered just a tiny amount of previous research on automated text classification of legal documents. Support vector machines (SVMs), for example, have been used to categorize legal documents such as judicial docket entries and non-English legal writings. We curated our dataset utilizing different information sources, despite the fact that our research also looks at the application of machine learning to a corpus published in the legal environment.

There are several techniques that have been previously used for judgment prediction which have been discussed below:

Information Retrieval: The act of filtering down large unstructured materials into groups of documents relevant to a specific topic is known as information retrieval (IR). With enormous libraries of online legal knowledge, IR significantly reduces the number of papers to be analyzed, speeding up the analysis process. Domain-independent or domain-dependent information retrieval methods exist. Domain independent systems take a broad variety of input documents and extract just general information like names and dates.

Ontology-based technique: A structural framework for the creation of legal ontologies was presented in one of the research works. Persons, objects, events, facts, and acts were the top-level components addressed in that work. The ontology also had the ideas of query improvement and case histories for legal concepts that alter their identity and category via procedures, as well as a comprehensive concept hierarchy and the concept of comprehending the terms that were significant to the major concepts. As a result, a knowledge engineering strategy was taken. This system showed how the generated knowledge base may be utilized to improve query enhancement system outcomes by employing inference methods that leverage ontology information.

ML Algorithms used for classification task:

The identification and classification of legal documents, which is a time-consuming procedure, is a crucial aspect of legal judgment prediction. Many researchers have been working on solutions that can automatically detect and organize requirements in papers. In recent years, these approaches have been built on machine learning (ML) technologies, which have yielded encouraging outcomes. There is also a need for a comprehensive knowledge of novel methods, which is currently lacking in the literature.

This paper analyses existing machine learning techniques that are commonly employed for the categorization of judicial judgments —

SVM (Support Vector Machine) - SVM is a supervised learning method used for both regression and classification. It is widely used because of its capacity to handle many continuous and categorical variables. This approach divides the dataset into classes in order to determine the greatest marginal hyper plane. The initial stage is to locate similar data points (support vectors) to the class. The dividing line can be defined using support vectors. To construct the classifier, SVM finds a linear function.

Decision Tree (DT) - The most popular and extensively used supervised learning method is the Decision Tree (DT). This technique may be used for both classification and regression. The two most essential things are decision nodes and leaves. The data is split at the decision node, and the outcome is represented by leaves. When many characteristics are utilized to assess the goal value of a given instance, the feature with the most information must be found as the feature on which the data may be separated. The measure information gain provides this feature information. ID3 is used to classify requirements in many research projects. SVM is outperformed by DT.

K-Nearest Neighbors (KNN) - It employs feature similarity to forecast the values of new data points, which implies that the new data point is given a value depending on how closely it matches the points in the training set. The K value, i.e., the nearest data point, must be chosen. For each point in the test data, the distance between the test data and each line of training data is calculated using one of the methods, such as Euclidean or Hamming distance. These computed values are used to sort them in ascending order. From the sorted results, the top rows of K are picked. The most common class of these rows is used to assign a class to the test point. This method has been used to categorize text by researchers. SVM and Multinomial Naive Bayes outperform KNN.

Naive Bayes (NB) - This classification technique is based on the premise that each characteristic in the same class is independent of the others. Even while this method takes less time to train and can handle big datasets, it has one drawback: it assumes that all characteristics are independent of one another. Real-life examples cannot include aspects that are unrelated to one another. Even though it is quick and scalable, it does not outperform other machine learning methods.

There are several unsupervised learning methods available, including LDA, K-means, and the single link clustering technique. The Bi-Term and Hierarchical Agglomerative Clustering algorithms are frequently used to find legal documents and forecast outcomes. According to research, the accuracy of these algorithms is similarly low.

5 Comparative Study of Various Machine Learning Approaches for Case Judgment Prediction with added Novelty:

It has been tried to work out the existing techniques by adding some novel tweaks to obtain better results. We have compared it through several different baseline works. We use the following models and judgment prediction methods as baselines for comparison. A comparative analysis has been charted out in

Below table.

Table 1. Analysis on various prediction techniques

Prediction Techniques	Novelty Added	Average Precision %	Average Recall %	Average Accuracy %	Average F1- Score
Linear Support Vector Machine [1]	Random Forest + SVM	79.12	82.15	80.63 (+2)	0.80
SVM + Neural Network [2]	CNN	87.20	89.86	88.53 (+3)	0.88
KNN + SVD + Linear Regression Classifier [3]	Fuzzy C-Means Clustering + SVM	81.14	83.54	80.89 (+2)	0.82
Decision Tree [4]	Decision Tree + Cosine Similarity	83.23	85.02	84.125 (+2)	0.84
SVM + Random Forest + Linear Regression [5]	Random Forest + SVM + Jaccard Similarity	87.96	89.11	88.53 (+3)	0.88
Naïve Bayes classifier + SVM [6]	Naïve Bayes Classifier + Decision Tree	74.45	77.26	75.86 (+2)	0.75

In [1], Linear SVM has been used as a classifier along with clustering techniques. Contiguous word sequences, such as N-grams and themes, are used to express textual information. The text content serves as the primary input for N-gram prediction and classification, as well as the subjects. ECHR Dataset [1] (3 datasets): Article 3, Article 6 and Article 8 have been used. The use of Random Forest in combination with SVM significantly increases the accuracy because Random Forest is a very efficient and powerful classifier than the other contemporaries.

In [2], Artificial Intelligence has been used to predict legal judgments based on the information presented in case files. Three different datasets were used to carry out the experimentation CJO [9], PKU [10], CAIL [11]. The subtasks are represented as graphs, and the Top-Judge architecture is built on the Directed Acyclic Graph format. Under this notion, two separate case studies are conducted. For multiple defendant judgments, this method is ineffective. The incorporation of conventional CNN in the architecture yields better results because CNN is a very strong classifier on its own. Temporary classification factors are taken into consideration as well which again adds to the model efficiency.

In [3], Machine Learning was used to speed up the estimate of slow-moving judgments. An ensemble of various techniques has been incorporated which leads to the system imbalance. The dataset used was prepared by three law students to annotate the accusations in over 100 criminal cases and from news websites [12]. When Fuzzy C-means clustering is combined with SVM, a significantly higher score for evaluation metrics is achieved, and transforming the categorical dataset for clustering integrates a large number of mathematical functions as well as transfer functions, implying a drop in the error rate advocates for the novelty added in the model.

In [4], a system has been proposed that uses an improved decision tree technique to detect suspicious texts about crimes. Anticipated emails were considered for the dataset. An enhanced ID3 method is linked to an advanced highlighting approach and an asset significance factor to build a quicker and better decision tree that is based on data entropy, which fluctuates unevenly as the process of creating an informative index from specific parts progresses. A comparatively better result is obtained by combining a decision tree with cosine similarity as cosine similarity shows semantic similarity score and decision tree is an excellent classifier.

In [5], researchers have employed data mining to forecast criminology and the causes of criminal activity. SNAP Gowalla, Data SF till Feb'15 was used as a dataset for this study. A hybrid of Random Forest along with SVM and Linear regression has been used which again causes model dissemblance. As a part of novelty added from our side, Random Forest has been hybridized with Support Vector Machine (SVM) and jacquard similarity with forms a two-pass classifier consisting of one very powerful classifier while another one being a binary linear classifier along with semantic similarity measure secures much higher precision, recall, etc.

In [6], a hybrid of Naive Bayes with SVM has been taken into consideration. Dataset was curated from Websites, blogs, RSS feeds. With Decision Tree in the picture instead of SVM, the assessment measures witness a slight increase because of Decision Tree being a very powerful classifier. The combination of the Decision tree along with Naive Bayes stabilizes the model to yield better results.

The goal of this comparative study is to look at the data mining techniques that have been utilized to predict crime from legal documents. The addition of a few powerful classifiers and algorithms to the existing work has resulted in the yield of better results. All the models were debugged in this experiment, and the ideal evaluation scores of each model were recorded. The data was recorded to two decimal places. The percentage unit is used, and the precise data is displayed in the table below. As a result of all of this, the result's relevancy is fairly high.

6 Methodology

This study presents a unique technique for predicting court case judgments based on previous similar instances to clients with requirements that are possibly relevant to their queries and needs. This research was carried out using a self-curated dataset as an input feature for classification and prediction. To make the final prediction, multi-fold relevance computation and semantic measures-based algorithms are used. The data preparation step includes PoS Tagging and Named Entity Recognition (NER). To generate a dataset, legal case materials were scraped from different jurisprudence websites of the Indian Supreme Court and other Indian High Courts, and then semantic matching and feature selection were performed. The inclusion of Logistic Regression allows for the classification of case judgments and final prediction.

Data Preparation Phase: When the dataset is encountered, it is first outfitted with the suggested system architecture, which includes a variety of Natural Language Processing (NLP) approaches. The proposed framework starts with the extraction of the required schema from the legal documents. The schema is represented in Figure 1. For extracting the schema Lex Predict is used as the dictionary for the keywords. The whole document corpus is traversed and based on specific keywords; data is extracted and is then appended into a CSV file having the same schema. The metadata is extracted and stored for later processing.

The data collected comprises judgment records that include unstructured language, cluttered documentation style, and so forth. As a result, this raw textual data is initially tokenized for preliminary phase pre-processing, in which each phrase is divided into smaller parts known as tokens. By studying the word sequences, these smaller pieces aid in context understanding and interpretation. Parts of Speech (PoS) Tagging comes after this pre-processing step which is the process of transforming a phrase into forms, such as a list of words or a list of tuples (each of which has a form (word, tag)). In this case, the tag is a part-of-speech tag, which indicates whether the word is a noun, adjective, verb, or other types of word. Named Entity Recognition (NER) is used to complete the data processing step (NER). Pre-defined characteristics are used to find and classify named items. Because the algorithm would be extracting a large number of names that are necessary for processing, we concentrated on named entity recognition and making it as exact as possible. The names of the presiding judges, the defendant(s), the prosecutor(s), and other relevant information. After that, a few preparation techniques are conducted to enhance the captured text's language and linguistics. The data pre-processing is done in such a way that well-structured data is obtained with the least amount of data loss feasible.

Feature Selection Phase: Hereupon, the actual task is initialized, which involves doing multiple-phase processing on the input data in order to extract suitable categorization and feature extraction from the dataset. We have considered three different categories for classification in this study, namely, win, lose, settlement. Once, the CSV has been formed in the previous phases, the feature extraction phase is initiated. The first step of this phase is marked by dependency parsing which examines the relationships between the words of a sentence to determine its grammatical structure. The mechanism is based on the idea that every linguistic element in a phrase has a direct relationship. Dependencies are the names given to these connections. After this, Lemmatization is carried out on obtained dependencies. Lemmatization typically refers to doing things correctly using a vocabulary and morphological study of words, with the goal of removing only inflectional ends and returning the base or dictionary form of a word, known as the lemma. With this process, non-identical inflections of a word are brought together so that they can be analyzed as a distinct entity.

The lemmatized text corpus is given in for n-gram extraction which is an n-item continuous sequence from a given text sample. In the form of a (n 1)–order, an n-gram model is a sort of probabilistic language model for predicting the next item in a sequence. Different variations of N-grams have been implemented on the data under study and it was substantiated that Trigram and 4-gram suit the proposed model better than the other contemporaries. Tri-grams and 4-grams are generated for each of the case judgment outcomes. All the term sets are stored and a corresponding dictionary is created to store their count. These dictionaries are converted to data frames. The top 50 % of the segregated instances are taken into account based on their frequency for each of the categories: win, lose and settlement.

Once n-grams are extracted, verb-class clustering is exercised on the instances obtained. The classes and verbs for lexical-semantic verb-class clustering was chosen at random. The sub categorization frames (SCFs) and related frequencies in corpus data, which record the grammatical context in which the verbs appear in the text, are the characteristics of each verb. SCFs were taken from the VALEX vocabulary, which is freely accessible (Korhonen et al., 23). VALEX was obtained automatically using RASP (Briscoe & Carroll et al., 24), a domain-independent statistical parsing toolbox, and a classifier that recognizes verbal SCFs. Following that, each verb's feature vector was normalized to unit length so that the frequency of the verb had no effect on its representation. With a considerably clean version of the lexicon, we got an appreciable result. The number of clusters to be found was set equal to the number of classes in the dataset to get this result. At the end of this phase, a feature set is devised after a rigorous treatment under a multi-fold semantic matching score calculation technique, which constitutes the most refined set produced.

Features are then selected from the acquired feature set based on which we categorize the documents. The chosen characteristics are described in a schema, and they are utilized for classification after being encoded into the model. The texts are clustered here under a number of well-chosen features, making them particularly acceptable to the realm of jurisprudence. The categorical aspect of the target variable is heavily leveraged in the proposed technique; hence Logistic Regression is used to do the classification task. The dependent variable is modeled using a logistic function. In Logistic Regression, a sigmoid function is utilized to map the predicted values. Our model is particularly efficient due to the structure of this function, which has a non-negative derivative for each point and exactly one inflection point. Equation 1 depicts the logistic regression cost function.

$$Cost\left(h0(i), J(actual)\right) = -log\left(h0(i)\right), if y = 1$$

$$log\left(1 - h0(i)\right), if y = 0$$

The negative sign denotes the requirement to maximize probability by minimizing the loss function. Once, the classification results are obtained, the prediction of case judgment is given out to the user based on their query.

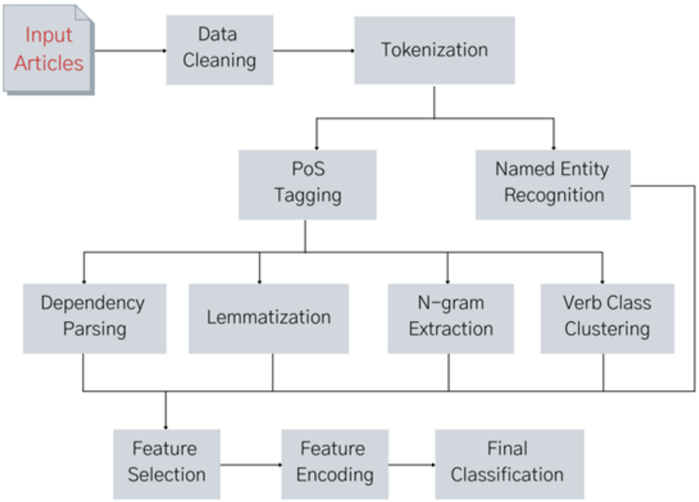

Fig 1: Proposed Architecture

Figure 1 depicts the whole system architecture of the suggested paradigm. The system begins with data entry, which is then subjected to pre-processing procedures. Following pre-processing, multi-phase semantic matching begins with a semantic similarity-based feature extraction approach, followed by the creation of a classified truth set with improved data. When the data reaches this stage, all redundant data has been removed. The data is now sufficiently qualified to be used for training, and the final prediction is then acquired.

7 Experimental Results and Evaluation

The majority of the Indian populace is unfamiliar with the legal system, and anytime a legal crisis develops, they find themselves in the middle of the crossroad. Though the proposed approach is unlikely to be able to take the role of legal counsel. It can, however, be the first line of defence in guiding the path forward. Our research focuses on not just evaluating existing approaches, but also on designing our own architecture for legal document classification and case judgment prediction. The architecture put forth can grasp the high-level issue using AI and analytics, and it can estimate the indicative case judgment and winning probability based on the previously solved instances. Logistic Regression with feature encoding was used to extract potential features for various client source documents. These parameters demonstrated the highest level of training accuracy. The data for the model training was gathered from around 1000 documents on diverse instances. The training samples are comparatively greater in size at the highest level, and the results obtained advocate that for text classification tasks, Logistic Regression is a significantly considerable approach that produces reliable results.

Experiment Preparation: The data used to carry out this experiment were collected from various Indian Jurisprudence websites, for example, official Supreme Court and high court websites. The data was cleaned and pre-processed to ensure the sampling and quality of the trained data. The total number of samples maintained is nearly 10,000, with the training, verification, and test sets accounting for 90%, 5%, and 5%, respectively. The complete experimentation has been carried out on Google Collaboration.

Evaluation Metrics: Certain assessment criteria, including Precision, Recall, Accuracy, and F-score, have been explored to evaluate the performance of the suggested technique. As illustrated in Equation, precision may be defined as a metric that estimates the number of positive examples that are true. As shown in Equation, recall is the percentage of all relevant occurrences that have been retrieved successfully. As shown in Equation, accuracy is calculated as the ratio of valid predictions made to the total number of samples supplied as input. The F-score is a method of combining the model's accuracy and recall, and it's described as the harmonic mean of the model's precision and recall, as shown in Equation.

Table 2. Metrics for evaluating performance

Search Technique	**Average Precision %**	**Average Recall %**	**Average Accuracy %**	**Average F-Score**
JPILKNN [7] P11	81.42	76.69	79.97	0.78
AMLJPBMF [8] P6	82.27	75.21	78.28	0.78
Proposed Methodology	**87.83**	**80.71**	**82.72**	**0.84**

Table 2 shows that the proposed framework achieves an average precision of 87.83 percent, an average recall of 80.71 percent, average accuracy of 82.72 percent and an average F-score of 0.84. For all of the assessment indicators taken into account, the results provided by the technique are considered noteworthy.

The proposed methodology's performance is greatly aided by the NLP-based feature selection and extraction approach. To create the final feature vector, multiple NLP techniques were used, integrating a variety of dependable algorithms and methodologies. Another key for the proposed algorithm's high percentage performance is data material created from several Indian Jurisprudence websites, which ensures that both recent and prior legal judgment results are incorporated in the strategy. The use of Logistic Regression, a very effective classifier, ensures that the classification of judicial judgments is done with extreme precision. Another reason for the proposed algorithm's better performance is the inclusion of feature weighting techniques, which aid in the elimination of misleading data while simultaneously lowering the computational cost of the model.

Implementation of Named Entity Recognition along with PoS tagging on the tokenized data helps in extracting the major entities in a text which aids in the organization of unstructured data and the detection of relevant information, which is essential when working with huge datasets. The addition of dependency parsing to the model architecture adds three advantages: First, dependency linkages are near to the semantic relationships that are required for the next phase of the interpretation process. Second, instead of the mid-level nodes seen in constituent trees, the dependency tree has one node per word, making processing easier and even allowing for pure corpus-based techniques. Finally, dependency parsing lends itself to operation on a word-by-word basis, i.e., parsing may be done by accepting and attaching words rather than whole sentences. Employment of n-gram extraction assists in making the model significantly stable. N-gram models add simplicity and scalability to the architecture. With bigger n, a model becomes capable of holding more contexts with a well-understood space-time tradeoff, allowing modest experiments to scale up effectively, the result of which can be witnessed in our obtained metric scores. The selection of features based on the semantic measures derived from the previously described techniques incorporated in the model amplifies the model efficiency. Feature encoding has been implemented to amplify the relevance of the results.

Two more baseline models were taken into consideration for the comparative analysis. In AMLJPBMF [8] the results of the experiments suggest that the proposed model can be used for crime prediction and associated legal prediction tasks well, however, it does not precisely optimize the model due to a lack of computational resources. Furthermore, good models such as DHCP and HAN have no model fusion. Multiple classification models have been experimented with in this work, however, none of them stands out as a reliable system for legal judgment prediction.

In JPILKNN [8], researchers look for ways to openly inject legal knowledge into legal judgment prediction. The suggested paradigm encodes declarative legal knowledge as a collection of first-order logic rules, which are then integrated into an end-to-end co-attention network model. The application of logic rules improves the model's interpretability by providing neural networks with direct logical reasoning skills. Furthermore, the inductive bias induced by legal knowledge alleviates deep neural networks' data-hunger. However, this approach makes the model too unstable to yield reliable and relevant results. The absence of semantic knowledge measures highly affects the performance of the model.

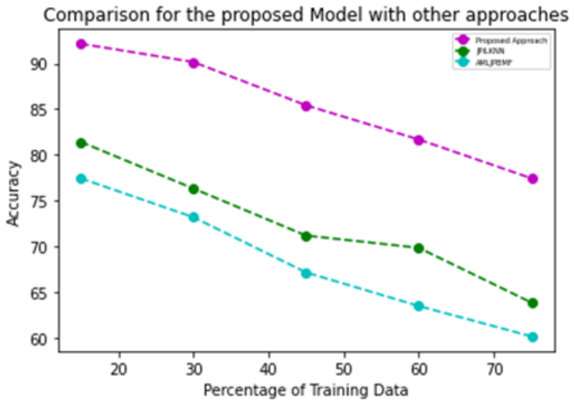

Fig 2: Comparative analysis of various approaches with built method

In terms of accuracy, Fig. 2 depicts a comparison of all three models. Because Accuracy is the percentage of correct predictions made by our model, the higher the Accuracy, the more efficient the model is. The proposed system achieves substantially greater accuracy than JPILKNN [7] or AMLJPBMF [8], as seen in Fig. 2. The suggested algorithm's efficiency is improved by the use of semantic measures-based feature selection. Using feature encoding and then classifying using a classifier based on Logistic Regression improves performance even further. The graphical representation of the Accuracy attained by all three models under examination in relation to the training data utilized demonstrates the competence of the system architecture put forth. For each batch of training data analyzed, the suggested model tends to get the greatest F-measure score. The approach proposed, out of all the data evaluated, appears to have the highest efficiency.

As a result of all of this, the result's relevancy is fairly high. The results of the experiments suggest that the proposed model can be used for crime prediction and associated legal prediction tasks well. Taking into account all of the data and model results, it can be determined that the technique proposed is more efficient than alternative frameworks now in use. As a result, the methodology put forth can be identified as an efficient model for the prediction of legal judgments.

8 Conclusions and Future Work

Researchers have been paying more attention to the topic of forecasting court outcomes using machine learning methodologies in recent years, as technology developments in machine learning and natural language processing are now capable of delivering on this promise. We provide a set of experiments evaluating six machine learning models for predicting judicial judgments using just textual information derived from relevant documents in this study. These models, which include k-NN, logistic regression, bagging, random forests, and SVM, include a variety of classifier options, such as parameterized vs. non-parameterized, high variance algorithms, and feature improvement approaches for small data sets.

This research also proposes a novel and brilliant technique for the prediction of legal judgments based on historical case data. A logistic regression-based classification model has been put forth wherein multiple NLP techniques have been used for feature selection based on semantic measures. The experiments also show that selecting a feature space based on the semantic content of the text matching to fact patterns found in the relevant case documents has a considerable impact on predictive model performance. The feature vectors calculated from the case's factual background, which are thought to be the deciding elements in court case outcomes, consistently exhibit a better correlation with the forecasting findings. As a result, combining semantic information with high-level characteristics is crucial for boosting the classification performance of machine learning systems for judicial judgments.

One of the most challenging issues in NLP is extracting semantic information (or knowledge) from natural language, and one of the prospective ways is word embedding, which is widely used in deep learning for NLP currently. In the future, we'll look into how to use such an approach to construct high-level features. Furthermore, predictive models developed using high-level semantic characteristics are more interpretive for humans, which is critical for legal practitioners, the majority of whom have little or no machine learning experience, to be ready to embrace such an approach.

References:

[1] Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro,D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. PeerJ Computer Science, 2, e93. https://doi.org/10.7717/peerj-cs.93

[2] D. Huang and W. Lin, "A Model for Legal Judgment Prediction Based on Multi-model Fusion,"*2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE)*, 2019, pp. 892-895, Doi: 10.1109/EITCE47263.2019.9094946.

[3] Evans, O., Stuhlmüller, A., Cundy, C., Carey, R., Kenton, Z., McGrath, T., & Schreiber, A. (2018). Predicting Human Deliberative judgments with Machine Learning. Technical report, University of Oxford.

[4] G. Boella, L. D. Caro, and L. Humphreys, ''Using classification to support legal knowledge engineers in the Eunomos legal document management system,'' in Proc. 5th Int. Workshop Juris-Inform., 2011, pp. 1–12.

[5] http://wenshu.court.gov.cn/

[6] http://www.pkulaw.com/

[7] http://cail.cipsc.org.cn/index.html

[8] J. Bala, M. Kellar and F. Ramberg, "Predictive analytics for litigation case management,"*2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 3826-3830, Doi: 10.1109/BigData.2017.8258384.

[9] Khan, Mohiuddin Ali, Sateesh Kumar Pradhan, and Huda Fatima. "Applying data mining techniques in cybercrimes." 2017 2nd International Conference on Anti-Cyber Crimes (ICACC). IEEE, 2017.

[10] Mugdha Sharma, "Z-Crime: A Data Mining Tool for the Detection of Suspicious Criminal Activities based on the Decision Tree", International Conference on Data Mining and Intelligent Computing, pp. 1-6, 2014.

[11] 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 Comput. Sci., vol. 2, p. E93, Oct. 2016.

[12] O. Sulea, M. Zampieri, S. Malmasi, M. Vela, L. P. Dinu, and J. van Genabith, "Exploring the use of text classification in the legal domain," CoRR, vol. abs/1710.09306, 2017. [Online]. Available: http://arxiv.org/abs/1710.09306

[13] O. Şulea, M. Zampieri, M. Vela, and J. van Genabith, ''Predicting the law area and decisions of French supreme court cases,'' in Proc. RANLP, Varna, Bulgaria, 2017, pp. 716–722.

[14] R. Nallapati and C. D. Manning, "Legal docket-entry classification: Where machine learning stumbles," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, ser. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 438–446. [Online]. Available: http://dl.acm.org/citation.cfm? id=1613715.1613771

[15] Stefanie Brüninghaus and Kevin D. Ashley, "Combining Case-Based and Model-Based Reasoning for Predicting the Outcome of Legal Cases" International Conference on Case-Based Reasoning, ICCBR 2003.

[16] Shiju Sathyadevan, M.S. Devan and S. Surya Gangadharan, "Crime Analysis and Prediction using Data Mining", Proceedings of IEEE 1st

International Conference on Networks and Soft Computing, pp. 406-412, 2014.

[17] Tarek Mahfouz and Amr Kandil, “Litigation Outcome Prediction of Differing Site Condition Disputes through Machine Learning Models”, Journal of Computing in Civil Engineering. Volume 26 Issue 3 - May 2012.

[18] W. Lin, T. Kuo, T. Chang, C. Yen, C. Chen, and S. Lin, ‘‘Exploiting machine learning models for Chinese legal documents labeling, case classification, and sentencing prediction,’’ IJCLCLP, vol. 17, no. 4, pp. 49–68, 2012.

[19] Yu-Yueh Huang, Cheng-Te Li and Shyh-Kang Jeng, “Mining Location-based Social Networks for Criminal Activity Prediction”, Proceedings of 24th IEEE International Conference on Wireless and Optical Communication, pp. 185-190, 2015.

[20] Y. Liu and Y. Chen, ‘‘A two-phase sentiment analysis approach for judgment prediction,’’ J. Inf. Sci., vol. 44, no. 5, pp. 594–607, 2018.

[21] Zhong, H., Zhipeng, G., Tu, C., Xiao, C., Liu, Z., & Sun, M. (2018). Legal Judgment Prediction via Topological Learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3540-3549).