AI/ML/RL based techniques for Judicial Analytics and Eficient Court Management

Project-I (CS47005) report submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

by

Atishay Jain

(20CS30008)

Under the supervision of

Professor Partha Pratim Chakrabarti

Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur

Autumn Semester, 2023-24

November 1, 2023

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: November 1, 2023 (Atishay Jain)
Place: Kharagpur (20CS30008)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**AI/ML/RL based techniques for Judicial Analytics and Eficient Court Management**" submitted by **Atishay Jain** (Roll No. 20CS30008) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering is a record of bona fide work carried out by him under my supervision and guidance during Autumn Semester, 2023-24.

Date: November 1, 2023

Place: Kharagpur

Professor Partha Pratim Chakrabarti
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur - 721302, India

Abstract

Name of the student: **Atishay Jain** Roll No: **20CS30008**

Degree for which submitted: **Bachelor of Technology**

Department: **Department of Computer Science and Engineering**

Thesis title: **AI/ML/RL based techniques for Judicial Analytics and Eficient Court Management**

Thesis supervisor: **Professor Partha Pratim Chakrabarti**

Month and year of thesis submission: **November 1, 2023**

Abstract:

In the era of rapid advancements in language technology and artificial intelligence, innovative solutions have been harnessed across diverse domains, including law, medicine, and mental health. Notably, AI-based Language Models like Judgement Prediction have emerged as promising tools within the legal sector. However, these models are not immune to the biases ingrained in their training data, raising concerns about fairness.

This research undertakes an initial exploration of fairness from an Indian context within the legal domain. We endeavor to develop a tool that aids law officials in making unbiased predictions for bail requests based on past cases. Leveraging a substantial dataset comprising around 900,000 Hindi legal cases, we employed Latent Dirichlet Allocation (LDA) for topic labeling and represented documents with a limited set of keywords. Various classifiers utilizing these keywords as parameters were employed for prediction.

Our key findings reveal that, when training models on 70% of the dataset and evaluating on the remaining 30%, the best accuracy achieved across all improved models approximated 75%. However, we recommend broadening the keyword representation within documents to potentially enhance results.

Moreover, given the extensive backlog of pending cases in Indian courts, this tool holds undeniable potential to aid legal practitioners and students alike. Nevertheless, it's important to acknowledge the limitations, as uncertainty in prediction results can significantly fluctuate among documents, occasionally resembling random 'yes' or 'no' decisions.

In a broader context, we address the prevalence of algorithmic biases within AI-based legal tools, with a specific focus on the Indian perspective. We highlight the need for further research and studies on fairness and bias within AI applications in the legal sector, emphasizing the imperative goal of equitable decision-making, free from societal biases.

Acknowledgements

I would like to extend my heartfelt gratitude to my thesis advisor, Prof. Partha Pratim Chakrabarti, for his exceptional guidance and unwavering support. Without his invaluable mentorship, this project would not have realized its full potential. Prof. Chakrabarti consistently encouraged me to take ownership of my work while providing crucial direction whenever needed. His motivation to explore extensive research, delve into numerous papers, and experiment with various ideas has been instrumental in the project's success. I am deeply appreciative of his support and the resources made available.

I also wish to convey my profound appreciation to my parents, whose unwavering support and continuous encouragement have been my pillars throughout my years of study and the extensive research and writing process of this thesis. This achievement would not have been possible without their steadfast belief in me.

Thank You

Atishay Jain

Contents

Declaration i

Certificate ii

Abstract iii

Acknowledgements v

Contents vi

List of Figures viii

List of Tables ix

Abbreviations x

1 Introduction 1
- 1.1 Background . . . 1
- 1.2 Objective and Motivation . . . 2
- 1.3 Scope . . . 2
- 1.4 Methodology . . . 3
- 1.5 Outline of Report Structure . . . 3
- 1.6 Significance . . . 4
- 1.7 Audience . . . 4

2 Literature Background 5
- 2.1 Current Scenario . . . 5
  - 2.1.1 Special Cases . . . 6
- 2.2 Fairness Gap . . . 6
  - 2.2.1 Bail Prediction . . . 7

3 Methods 8
- 3.1 Overview . . . 8
- 3.2 Dataset . . . 9

3.3 Base Model 10
3.3.1 LDA 10
3.3.2 Classifiers 12
3.3.3 Combining the results 14
3.4 Analysis 15
3.4.1 Issues 16
3.4.2 Suggestions 17
3.5 Refined Model 17
3.5.1 Results and Analysis 18

4 Conclusion **21**
4.1 Future Work 22

List of Figures

3.1 Structure of Dataset 9
3.2 The flow of the method used 12
3.3 Confidence vs Accuracy for Decision Tree Classifier 13
3.4 Probability vs accuracy for other classifiers 13
3.5 Confidence vs Accuracy for output with highest probability 14
3.6 Confidence vs Accuracy for output with highest accuracy 15
3.7 Confidence vs Accuracy for refined model 19
3.8 Documents having same vector representation and the ratio of their output. X-axis is the % of having same output(100% means that all the documents with this representation have same output). Y-axis represent the no. of documents. 19

List of Tables

3.1 Token Counts for the text in the facts-and-arguments section. Preprocessing steps included removal of Hindi stopwords, punctuation marks, filtering out weblinks and too small documents. 10
3.2 Accuracies of different model 15
3.3 Comparison between base and refined model 20

Abbreviations

LDA	**L**atent **D**irichlet **A**llocation
HLDC	**H**indi **L**egal **D**ocument **C**orpus
AI	**A**rtificial **I**ntelligence
JSON	**J**ava**S**cript **0**bject **N**otation
ML	**M**achine **L**earning

Chapter 1

Introduction

Recent strides in language technology and Natural Language Processing (NLP) have propelled us into a new era, particularly transformative in the fields of law and justice. This leap forward has rekindled our commitment to crafting domain-specific datasets and fine-tuned language models, customized explicitly for the intricate world of legal discourse, giving birth to the burgeoning domain of LegalNLP.

Yet, amid these innovations, challenges persist in dealing with issues of hallucination, where language models generate content that may not align with factual reality, and uncertainty, as these models often struggle to express the degree of doubt associated with certain information. Addressing these complexities is integral to harnessing the full potential of LegalNLP.

1.1 Background

Amid these innovations, challenges persist in dealing with issues of hallucination, where language models generate content that may not align with factual reality, and uncertainty, as these models often struggle to express the degree of doubt associated

with certain information. Here, we will be using a completely different and much simpler approach to adderss the problem in this domain.

1.2 Objective and Motivation

India, having the highest number of pending cases in the world, have around 4.3 crores pending court cases, predominantly in district courts. We contend that not all cases are equal and that bail cases, being relatively straightforward, can be expedited. To achieve this, we've worked to developed a predictive model that offers unbiased predictions on bail outcomes, eliminating any form of identity-based bias, such as religion or caste. Our aim is to accelerate judicial processes, ensure equitable justice, and optimize resource allocation, addressing a critical need within India's legal system.

1.3 Scope

We acknowledge that distinct regions within our country may hold slightly varying norms and perspectives on specific cases. Consequently, the final outcome of bail requests could indeed be influenced by these regional disparities. In our research, we have exclusively utilized data from the District Courts of Uttar Pradesh (UP) to train our model.

However, we strongly advocate expanding the scope of our model's training to encompass datasets from various states across India. This broader approach will ensure that the model is attuned to the diverse legal landscapes and norms that exist throughout the nation, thereby enhancing its applicability and accuracy.

1.4 Methodology

We have sampled our data points from a large dataset containing 900,000 Hindi legal documents. We utilized the concept of topic modeling based on the idea that "a document is best described using a group of topics, and topics are best described using a group of words." Our methodology can be succinctly described as follows: we begin by selecting a few keywords from a given document. These keywords are then augmented with additional relevant terms to create a concise representation of the case document. This representation is standardized to a uniform length.

Subsequently, we employ this set of keywords as input parameters for various binary classifiers. These classifiers are tasked with predicting the outcome of the case—whether bail should be granted or denied. The outcomes of the classifiers are combined to achieve better results. To refine and enhance the accuracy of our model, we have implemented several iterative improvements. These improvements include using more keywords, refining keyword selection and preprocessing. Furthermore, we have conducted a comprehensive analysis to gain deeper insights into the results and to identify areas with potential for further enhancement.

1.5 Outline of Report Structure

Section 2 of this report provides a comprehensive overview of the prior research conducted in domains that directly impact our work. The detailed methodology, including the techniques applied, refinements made, and the subsequent analysis of outcomes, is expounded upon in Section 3. In Section 4, we delineate the future avenues of research and development.

1.6 Significance

Considering the substantial backlog of pending cases, this endeavor represents a concerted effort to expedite the judicial process. It aims to ensure the judicious allocation of resources, directing them toward cases of utmost significance and gravity, ultimately contributing to a more efficient and equitable legal system.

1.7 Audience

The model is created to assist legal practitioners in obtaining predicted outcomes for bail requests. Moreover, this report can also be a valuable learning resource for students interested in machine learning.

Chapter 2

Literature Background

2.1 Current Scenario

In addition to addressing concerns related to judicial system delays, it is imperative to discuss the issue of bias. The analysis (Elliott Ash (2023)) is based on an extensive dataset comprising 5 million criminal cases spanning over 7000 district and subordinate courts in India, encompassing the years 2010 to 2018. In this context, the focus is not on the legal intricacies of the cases, but rather on the roles played by the identities of the victim, defendant, and presiding judges in influencing case verdicts.

To begin, it is essential to establish a clear understanding of what we mean by the "identity" of an individual. This identity can encompass various aspects, such as membership in social groups defined by religion, caste, gender, or surname. A notable challenge arises from the fact that case filings often do not explicitly mention an individual's religion or caste. Consequently, a predictive model was developed capable of ascertaining these attributes from names alone, achieving a remarkable accuracy score of 97%.

It is important to note that the composition of judges within the judiciary does not mirror the demographic distribution of the population, particularly in terms of factors such as religion, gender, and caste. For the purpose of our analysis, we operate under the assumption that judges are assigned to cases in a random fashion.

2.1.1 Special Cases

A fundamental question arises: Does the likelihood of acquittal increase when a shared identity exists between the judge and the victim? Surprisingly, the answer to this question is a resounding 'no,' or to be more precise, 'almost no.' Our analysis reveals that there are no discernible differences in case outcomes based on factors such as religion or gender. While this observation might raise the possibility of a skewed dataset, the likelihood of a substantial dataset exhibiting such bias is exceedingly low.

However, it is noteworthy that instances of bias become more pronounced when a heightened sense of identity belonging is evident, such as during the month of Ramadan, or when shared identities are very narrowly defined. It is essential to acknowledge that such cases are relatively rare, and as such, we can reasonably assume a lack of bias in general.

2.2 Fairness Gap

In recent times, the rapid progress and implementation of language technology and artificial intelligence have yielded significant achievements in various domains, including law, medicine, and mental health. Notably, the introduction of AI-based Language Models, such as Judgement Prediction, has garnered attention in the legal sector. While these advancements hold immense promise, deploying legal technology

without a meticulous evaluation of bias and fairness carries the risk of generating unjust and biased outcomes, potentially eroding public trust in the legal system.

It is essential to recognize that Natural Language Processing (NLP) systems, trained on extensive corpora sourced from legal archives, are susceptible to assimilating historical social biases ingrained within the data (Yijun Xiao (2021)). Consequently, this poses a risk of perpetuating unfair decision-making in the future. It is well-documented that historical legal data fails to represent all social groups equally or fairly, as it inherently reflects the pervasive human and institutional biases that permeate society (Aniket Deroy and Ghosh (2023)).

2.2.1 Bail Prediction

This research marks the initial phase of our exploration into the concept of fairness in the realm of Indian legal data (Sahil Girhepuje1 and Ravindran (2023)). We delve into the intriguing dynamics of algorithmic biases that become embedded in models trained on Hindi legal documents, particularly when applied to the bail prediction task. The heart of our investigation revolves around the assessment of fairness, with a specific focus on demographic parity.

One of the key findings from our study is striking: when a decision tree model is deployed for the bail prediction task, it exhibits a pronounced fairness disparity. Specifically, this disparity amounts to 0.237 concerning the input features associated with Hindus and Muslims. This discovery sheds light on significant imbalances within the predictive model, igniting discussions about the potential consequences for individuals hailing from diverse demographic backgrounds.

Chapter 3

Methods

3.1 Overview

India, as a nation, boasts substantial diversity across various dimensions, including religion, caste, language, ethnicity, and more. This multifaceted diversity underscores the impracticality of relying on a single, uniform AI model to cater to the entire country's needs. Consequently, the demand for region-specific AI models becomes imperative. However, the intricacies of this diversity also exert a notable influence on the data collected, potentially introducing biases into AI models.

This report is primarily focused on the bail prediction aspect within the legal context and endeavors to achieve unbiased and accurate predictions. Our core objective revolves around the development of a model tailored to the nuances of the Indian legal landscape. This model is designed to predict whether bail should be granted to an individual based on prior cases that share similarities. The overarching goal is to ensure fairness and accuracy in the bail prediction process while accommodating the rich diversity of the Indian populace.

```
{
    '{serial_no.}':
    {
        'decision': granted/dismissed,
        'segments':
        {
            'judge-opinion':
            'facts-and-arguments':
        },
        'district':,
        'case_number':
    }
}
```

FIGURE 3.1: Structure of Dataset

3.2 Dataset

For this investigation, we will leverage the Hindi Legal Documents Corpus (HLDC), a meticulously curated collection introduced by Kapoor et al. The HLDC encompasses over 900,000 legal documents in Hindi, thoughtfully prepared, Named Entity Removal executed and structured to facilitate the development of downstream applications, including the crucial domain of bail predictions.

The dataset is available in JSON format, with each entry containing a wealth of case-specific information. This includes essential details such as the decision (whether bail was granted or dismissed), the district in which the bail was sought, case numbers, and two pivotal sections: "judge-opinion" and "facts-and-arguments." The "facts-and-arguments" section provides insight into the case's details and the rationale for bail, while "judge-opinion" encapsulates the judge's viewpoint on the matter. The structure of the json file is given in Figure 3.1.

A noteworthy point to highlight is that, in an automated system, we will not have access to the judge's opinion.

Token Count	Original test	Cleaned text
Average	298.4	164.9
Min	1	20
Max	1377	765

TABLE 3.1: Token Counts for the text in the facts-and-arguments section. Preprocessing steps included removal of Hindi stopwords, punctuation marks, filtering out weblinks and too small documents.

Given the substantial size of the dataset, which comprises 900,000 entries, we will employ a subset of this dataset for various methods and approaches. Within this subset, 70% of the data entries will be utilized for model training, with the remaining 30% earmarked for testing and evaluation. The initial phase involves the formatting and serialization of approximately 100,000 case documents, paving the way for a robust and insightful analysis. The statistics of the dataset is given below in Table 3.1.

3.3 Base Model

We will use the model mentioned in the Fail Paper[.] as the base model. To predict the decision of a bail request, we will first represent the document by some keywords and then use a classifier to get the result.

3.3.1 LDA

In our analysis, we leverage a dataset comprising 116,962 cases. To enhance interpretability in subsequent stages, we employ a straightforward feature representation strategy. Our goal is to represent each case using pertinent keywords, drawn from the case text. Additionally, we consider the category or type of crime to gauge its significance in determining an applicant's eligibility for bail.

To extract these keywords from the case documents, we employ the concept of topic modeling, with a specific focus on Latent Dirichlet Allocation (LDA). LDA is a topic modeling technique that unveils latent topics within a document by examining a given text corpus. It operates on the foundational notion that "a document is best described using a group of topics, and topics are best described using a group of words." LDA is an unsupervised learning approach, requiring the user to specify the number of topics to be extracted from the corpus.

In our application of LDA topic modeling to the HLDC corpus, we meticulously analyzed the resulting topic clusters. After careful consideration, we determined that extracting 85 topics yields a respectable perplexity score of approximately -19. During the training of the LDA model, we exclusively utilized the "facts-and-arguments" section of the case documents. Furthermore, we undertook a text cleaning process, removing all links and stop words. Subsequently, we extracted the top 10 keywords for each topic.

LDA employs a soft-clustering methodology. Building on this characteristic, we extracted the top two topics to which each case belongs. In essence, each case is assigned a primary and a secondary dominant topic. These topic tells us about the category of the crime in a borad sense. This "soft clustering" concept is instrumental in providing a more comprehensive understanding of a case, as it considers the influence of two topics rather than a single one. Consequently, we extracted three keywords from the document associated with the primary dominant topic and two keywords related to the secondary-dominant topic. This selection process involves verifying the presence of the top keywords we previously extracted for each topic within the case document. In cases where fewer than three words out of the ten top keywords are present in the document, we employ a placeholder word, such as 'None.' As a result, each document is effectively represented as a vector comprising seven keywords, enhancing the interpretability and utility of our feature representation strategy.

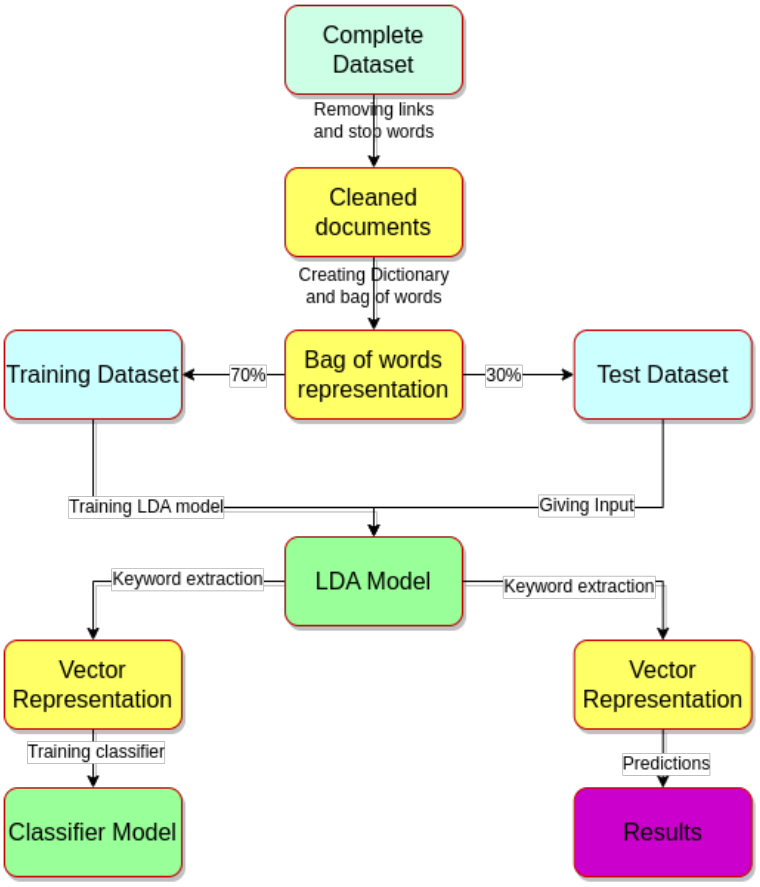


FIGURE 3.2: The flow of the method used

3.3.2 Classifiers

We employ a Decision Tree model for the bail prediction task, utilizing the seven aforementioned features. Decision Trees are chosen due to their superior interpretability. The model undergoes training on 70% of the entire dataset, achieving an accuracy score of 0.742 when evaluated on the remaining 30%. The flow chart of the model can be seen in Figure 3.2.

Notably, the probability/ uncertainty/ confidence of the model's output varies considerably among the test cases, ranging from as low as 0.5 to a full 1. Interestingly, even when the model assigns a probability of 1 to its decision, it does not consistently yield accurate predictions. The relationship between accuracy and probability/confidence is depicted in Figure 3.3.

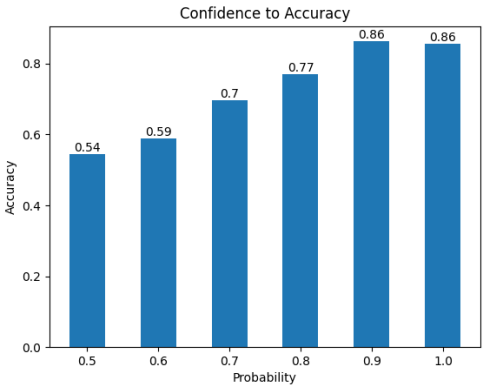
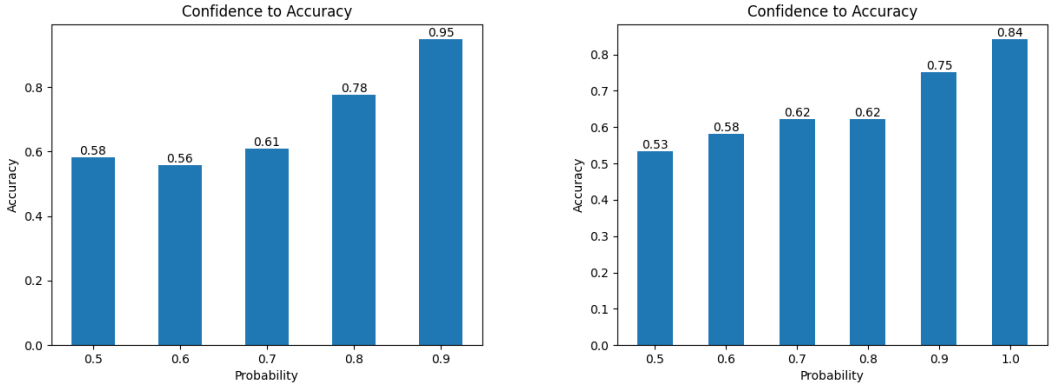


FIGURE 3.3: Confidence vs Accuracy for Decision Tree Classifier

In addition to the Decision Tree model, we have also employed Support Vector Classifier and Naive Bayes Classifier for prediction, despite their relatively lower interpretability. The accuracies achieved for these classifiers are 0.753 and 0.734, respectively. The plots illustrating the relationship between accuracy and probability/confidence can also be observed in Figure 3.4. Surprisingly, the probability exhibits similarity across all three classifiers on a larger scale.



(a) Support Vector Classifier

(b) Naive Bayes Classifier

FIGURE 3.4: Probability vs accuracy for other classifiers

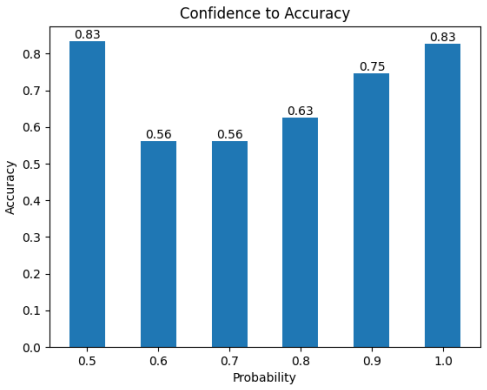


FIGURE 3.5: Confidence vs Accuracy for output with highest probability

3.3.3 Combining the results

Given that the average probability for all individual classifiers stands at 0.74, we embarked on a novel approach by simultaneously running all the models on the input and amalgamating their results in various ways, as elaborated below. This endeavor held the promise of yielding enhanced results.

1. We opted for the output of the classifier that yielded the highest probability for the desired outcome. With this technique, the accuracy improved marginally to 0.746. See Figure 3.5 for probability vs accuracy plot.

2. In another strategy, we chose the output of the classifier that achieved the highest accuracy within a predefined range of discrete probabilities. This range was established to ensure the meaningful aggregation of probabilities, as dealing with a continuous range would render each probability unique and render accuracy comparisons impractical. This approach led to an accuracy score of 0.749. Probability vs accuracy graph can be seen in Figure 3.6.

3. The third approach involved selecting the outcome produced by two or more classifiers. This ensemble strategy yielded an accuracy score of 0.752.

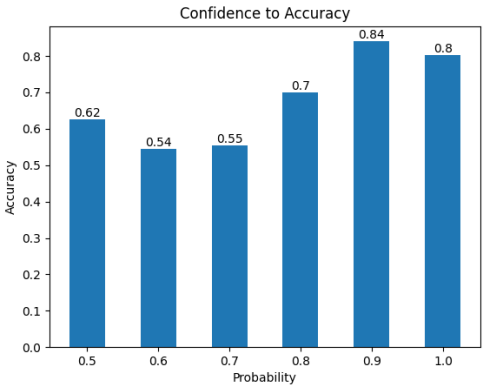


FIGURE 3.6: Confidence vs Accuracy for output with highest accuracy

Classifier	**Accuracy**
Decision Tree	0.742
Support Vector	0.753
Naive Bayes	0.734
Max. Probability output	0.746
Max. Accuracy output	0.749
Dominant Output	0.752

TABLE 3.2: Accuracies of different model

Despite our initial optimism, the accuracy still hovers in the range of 0.74-0.75. All the results are compiled in Table 3.2. Consequently, we deduced that the room for improvement lies not within the classifiers themselves but within the preceding stage of keyword extraction.

3.4 Analysis

Before implementing any changes, a comprehensive analysis of the results is imperative to identify areas of weakness. When examining the plots depicting the relationship between confidence and accuracy, a notable observation is made: for

lower confidence ranges, the output accuracy hovers around 0.5, which is tantamount to random output. To gain deeper insights, a statistical examination of the test data is warranted, aimed at uncovering potential causative factors.

The test dataset comprises 32,400 data points. It is notable that the average disparity in the probabilities assigned to outcomes by different classifiers is 0.203. This disparity can be elucidated as follows: for a considerable number of documents, all the classifiers converge on assigning low probabilities, resulting in a correspondingly low accuracy. Conversely, for specific documents, the discrepancy in assigned probabilities is substantial. A rather concerning observation emerges as well – for approximately 6,000 data points, the outcome predicted by all classifiers is incorrect.

3.4.1 Issues

The primary issue appears to reside within the keyword extraction process. Supporting this contention is the examination of the number of terms in the dictionary, which amounts to 524,682 – signifying the total count of distinct words present in all the documents. Remarkably, the count of unique words in the vector representation of all the documents amounts to a mere 406. This represents less than 0.1% of the total words. This phenomenon leads to the creation of only 12,998 distinct vectors for 10,800 documents, implying that different documents are mapped to the same vector.

What's more, a staggering 90% of these vectors encompass documents that give both bail granted and dismissed decisions. This predicament casts doubt on the trustworthiness of classifier training, as the same input may yield varying outputs for different occurrences.

3.4.2 Suggestions

1. Before tokenizing the document, use stemming and lemmatization of hindi words for more precise bag-of-words creation as different forms of a word are getting treated differently now.

2. Use more precise way to find the top keywords of a document instead of just checking whether the top 10 keywords for a topic are present in the document.

3. Use more than 7 keywords to represent a document. This will surely increase the no. of unique vectors generated for the dataset.

4. Include judge's opinion also in the training set as it will provide more detail about the case.

3.5 Refined Model

In response to the issues and recommendations outlined in the previous section, we have diligently crafted an improved iteration of our model, with the expectation of achieving enhanced results. The following text outlines the notable differences and improvements we have implemented.

We began by utilizing a subset of the previous dataset, comprising a total of 19,123 documents. This dataset was subsequently partitioned into training and test data, adhering to a ratio of 70% for training and 30% for testing. Within the training data, we thoughtfully included text from both sections of the documents, namely "facts-and-arguments" and "judge-opinion." To refine the text preprocessing, we executed lemmatization to transforms words into their base or normal form. Furthermore, we opted to consider only unique words within each document. In this modified bag-of-words model, each term is assigned a frequency of 1.

Following the data preparation, we proceeded to train an LDA model on this refined dataset, incorporating both the judge's opinion and the unique words trait. For the classifier training, we selected 12 keywords from each document. This selection includes 3 keywords pertaining to the primary, secondary and tertiary theme (crime), along with 4, 3, and 2 keywords, respectively, corresponding to distinct themes. In this iteration, we used a Decision Tree Classifier only as our earlier observation showed that different classifiers produced nearly identical results.

Upon completing the classifier training, we proceeded to extract the 12 keywords for the test data. In line with the training dataset, we performed lemmatization on the documents in the test dataset and constructed a bag-of-words representation, where each term has a frequency of 1.

With these comprehensive enhancements and refinements in place, we are now prepared to make predictions based on these features.

3.5.1 Results and Analysis

The training dataset consists of 13,386 documents and the test set comprises 5,737 documents. However, the final accuracy score achieved on the test dataset remains at 0.716, showing no significant improvement contrary to our hopes. The plot depicting the relationship between confidence and accuracy is presented in Figure 3.7. Notably, when the probability of the outcome is lower, the accuracy hovers around 0.5. Furthermore, even when the probability reaches its maximum (1), the accuracy remains notably low.

Let's delve into the statistics of the training data. In this iteration, each document is represented as a set of 12 keywords. However, out of the 13,386 documents, only 8,879 unique vectors were generated. This means that different documents share the same vector representation. Compounding this issue, documents with identical

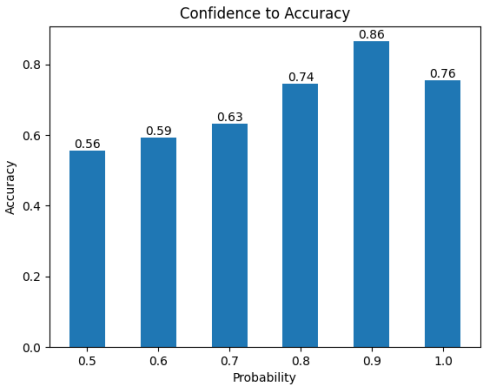


FIGURE 3.7: Confidence vs Accuracy for refined model

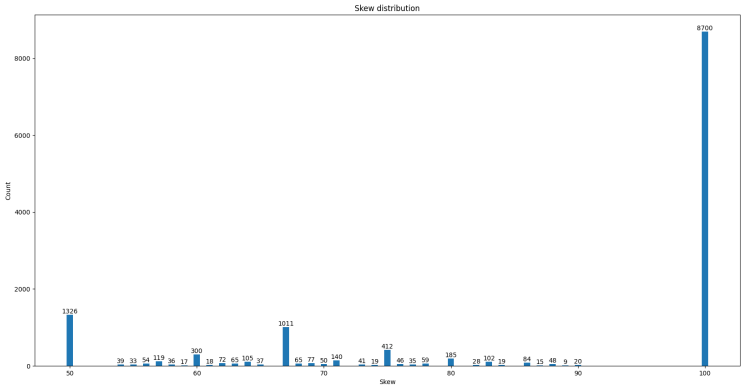


FIGURE 3.8: Documents having same vector representation and the ratio of their output. X-axis is the % of having same output(100% means that all the documents with this representation have same output). Y-axis represent the no. of documents.

vector representations yield different output decisions. The skewed distribution of the training set is visually depicted in Figure 3.8.

An essential consideration is the dictionary size is a total of 179,966 unique words across all documents. However, in both the training and test datasets, a mere 526 unique words are employed for document representation. As the dataset is in Hindi, lemmatization is also not much accurate. The no. of words that have changed from their original form to base form is only around 7% of the dictionary size. While we have made efforts to address the concerns raised in the previous section, it is

/	**Base**	**Refined**
Training data	75600	13386
Test Data	32400	5737
Dictionary	524,682	179,966
Words in vectors (%)	406 (0.07%)	526 (0.3%)
Unique vectors (%)	12,998 (12%)	12,963 (68%)
Accuracy	0.753	0.716

TABLE 3.3: Comparison between base and refined model

apparent that these challenges persist, albeit with reduced intensity. A detailed comparison is available in Table 3.3 for reference.

It is evident that while progress has been made, the issues surrounding classifier training, representation overlap, and dictionary size have not been entirely utilised.

Chapter 4

Conclusion

In conclusion, this report is a comprehensive exploration of the challenges and efforts involved in developing an unbiased and accurate bail prediction model tailored to the Indian legal context. The endeavor to create a model that respects the rich diversity of the Indian populace while ensuring accuracy in legal decisions is complex, and this report has laid bare the intricate journey of achieving this objective. We highlighted the potential for biases to be introduced into AI models due to the influence of this diversity on the data. Our study primarily focused on bail prediction within the legal domain, with the overarching aim of creating an accurate and unbiased model. To do so, we harnessed the Hindi Legal Documents Corpus (HLDC), an extensive dataset of over 900,000 legal documents in Hindi. This corpus served as the foundation for our analysis. Our base model employed Latent Dirichlet Allocation (LDA) for topic modeling, enhancing the interpretability of our feature representation. We leveraged a Decision Tree model for classification, with the support of additional classifiers such as Support Vector Classifier and Naive Bayes. While the base model achieved a reasonable accuracy of approximately 0.742, the accuracy was found to be inconsistent across probabilty/ confidence levels. We explored ensemble techniques, combining results from different classifiers, but the accuracy improvements were marginal. This led to the realization that the primary challenge lay in the

keyword extraction process, which introduced inconsistencies and hindered classifier training. To address these issues, we refined our model, using a smaller dataset and implementing lemmatization, using more keywords, enhanced keyword selection and including judge's opinion in the training set. Despite these efforts, the final accuracy remained at approximately 0.716.

4.1 Future Work

Considerable room for improvement remains evident, with our analysis pinpointing a significant challenge: the generation of identical vector representations for different documents. This issue is primarily rooted in the keyword selection process. Notably, a mere 0.1% of the words in the dictionary are actively contributing to all vectors.

Several suggestions that look promising include the following:

1. Utilizing a word-to-vector model for Hindi words to represent each word as a vector in the vector representation. This approach enables the training of classifiers on continuous parameters, improving the understanding of word similarity and proving more effective than simple lemmatization.

2. Adopting an alternative method for extracting the top keywords from a document, where each word is assigned a weight that factors into the keyword selection process.

3. If working with Hindi data proves challenging, a viable alternative is translating the data to English and then applying various optimization and cleaning techniques to facilitate the analysis.

Bibliography

Aniket Deroy, K. G. and Ghosh, S. (2023). How ready are pre-trained abstractive models and llms for legal case judgement summarization?

Elliott Ash, Sam Asher, A. B. S. B. D. C. T. D. C. G. P. N. B. S. (2023). In-group bias in the indian judiciary.

Sahil Girhepuje1, Anmol Goel, G. S. K. S. G. S. P. P. K. and Ravindran, B. (2023). *Are Models Trained on Indian Legal Data Fair?* PhD thesis, IIT Madras, IIIT Hyderabad, American Express.

Yijun Xiao, W. Y. W. (2021). *On Hallucination and Predictive Uncertainty in Conditional Language Generation.* PhD thesis, University of California, Santa Barbara.