LEGAL LINGUISTICS AND GENDER BIAS: STUDY OF LEGAL LANGUAGE MODELS AND DEBIASING STRATEGIES

A Thesis Submitted

in Partial Fulfillment of the Requirements
for the Degree of

MASTER OF TECHNOLOGY
in
ARTIFICIAL INTELLIGENCE

by

TRISHA GHOSH
(2K22/AFI/27)

Under the Supervision of
Prof. Shailender Kumar, Delhi Technological University

Department of Computer Science and Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

May, 2024

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I, **Trisha Ghosh (2K22/AFI/27)** hereby certify that the work which is being presented in the thesis entitled "**Legal Linguistics and Gender Bias: Study of Legal Language Models and Debiasing Strategies"** in partial fulfillment of the requirements for the award of the Degree of Master of Technology submitted in the Department of Computer Science and Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from August 2022 to May 2024 under the supervision of Prof. Shailender Kumar.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Place: Delhi **Trisha Ghosh**

Date: 31.05.2024 **(2K22/AFI/27)**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that Trisha Ghosh (2K22/AFI/27) has carried out their project work presented in this thesis entitled “**Legal Linguistics and Gender Bias: Study of Legal Language Models and Debiasing Strategies”** for the award of **Master of Technology** from the Department of Computer Science and Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Place: Delhi
Date: 31.05.2024

Prof. Shailender Kumar
(Professor)

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

ACKNOWLEDGEMENT

I am extremely grateful to my project guide, Prof. Shailender Kumar, Professor, Department of Computer Science and Engineering, Delhi Technological University, Delhi, for providing valuable guidance and being a constant source of inspiration throughout my research. I will always be indebted to him for the extensive support and encouragement he provided.

I am highly indebted to the panel faculties during all the progress evaluations for their guidance, constant supervision and for motivating me to complete my work. They helped me throughout by giving new ideas, providing necessary information, and pushing me forward to complete the work.

Place: Delhi **Trisha Ghosh**

Date: 31.05.2024 **(2K22/AFI/27)**

ABSTRACT

This thesis delves into the intersection of legal linguistics and gender bias, employing natural language processing (NLP) techniques for legal judgment prediction using transformer-based models. The research is centered on the Indian legal context, focusing on the performance and bias mitigation of various models, including BERT, XLNet, RoBERTa, DeBERTa, ELECTRA, and BigBird, evaluated on the ILDC-single dataset. The analysis requires a complex legal language, characterized by keywords and archetypes, which pose significant challenges to traditional NLP models.

Working in the Common Law system, Indian judiciary is characterized by complex and outdated legal documents, which lead to ambiguities and inconsistencies. These challenges require advanced NLP models to deal with the nuances of legal text. Transfomer-based models, with their self-focusing methods, offer promising solutions for providing comprehensive analysis of legal complexity.

An important aspect of this thesis is to examine gender bias in legal proceedings, an important issue that affects the fairness of judicial outcomes. Gender bias manifests itself in a variety of ways, including discriminatory language contained in legal documents as well as biased judicial decisions. This study examines this bias with Law2Vec embeddings and proposes strategies to overcome and reduce bias. Methods such as projection into gender subspace, k-means clustering are used to measure bias, followed by Hard Debiasing algorithm to reduce it. The effectiveness of bias is evaluated through court decision prediction tasks to ensure semantic retention integrity in embeddings.

The method involves applying six transformer-based models to the ILDC-single dataset, comparing their performance to predict court decisions BigBird-RoBERTa demonstrated superior performance at about 80% accuracy, which drew attention to its ability to process long sequences and extract relevant contexts. The study also includes a detailed analysis of Law2Vec embeddings, which identifies the gender bias and effectively reduces it.

Despite the promising results, the thesis acknowledges the limitations of the current models, in particular the lack of explanation. The non-transparent decision-making processes in these models pose challenges for regulatory professionals who need transparent and interpretable information. Future research will focus on integrating descriptive methods and extending the data sets to increase the robustness and generalizability of the models.

The findings highlight the importance of prior domain-specific training and the need for continued efforts to address biases in legal NLP applications. With transformer architectures advanced with the use of advanced bias methods, this research can help in the development of more ethical, transparent, and effective AI tools for legal services, increasing its efficiency in the process.

In conclusion, this thesis demonstrates the effectiveness of transformer-based models in predicting legal decision making and the critical importance of reducing gender bias in legal language models. The study contributes to future developments in legal NLP and provides a uniform and transparent judgement process.

Keywords: Legal Linguistics, NLP, Transformer Models, Gender Bias, Legal Judgment Prediction, Indian Legal System, BERT, XLNet, RoBERTa, DeBERTa, ELECTRA, BigBird, ILDC Dataset, Law2Vec Embeddings, Hard Debiasing, Explainability in AI, Legal Language Models, Bias Mitigation, Court Judgment Prediction, Ethical AI in Law

TABLE OF CONTENTS

LIST OF TABLE(S)

LIST OF FIGURE(S)

LIST OF ABBREVIATION(S)

AI	Artificial Intelligence
BEC-Cri	Bias Evaluation Corpus for Crime
BERT	Bidirectional Encoder Representations from Transformers
BiGRU	Bidirectional Gated Recurrent Unit
CJP	Court Judgement Prediction
CJPE	Court Judgement Prediction and Explanation
Cri-Bias	Crime Bias
DeBERTa	Decoding-enhanced BERT with Disentangled Attention
ECtHR	European Court of Human Rights
EEC	Equity Evaluation Corpus
ELECTRA	Efficiently Learning an Encoder that Classifies Token Replacements Accurately
GAP	Gender Ambiguous Pronouns
GPD	Gender Preserving Debias
ILDC	Indian Legal Documents Corpus
kNN	k-Nearest Neighbors
LCD	Legal Context Debias
LIME	Local Interpretable Model-agnostic Explanations
LSI	Legal Statute Identification
LSTM	Long Short Term Memory
NLP	Natural Language Processing
PCA	Principal Component Analysis

PLM	Pre-trained Language Models
RBF	Radial Basis Function
RNN	Recurrent Neural Network
RoBERTa	Robustly Optimized BERT Approach
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machines

CHAPTER 1
INTRODUCTION

1.1 Legal Linguistics using NLP

Legal linguistics is a field in which several aspects of law and linguistic cross paths in order to comprehend and examine language used in legal texts. The evolution of NLP has brought about massive changes by enabling the automation and analysis of numerous quantities of legal documents. This convergence between NLP and legal linguistics has seen major advancements made in areas such as contract analysis, automated document drafting, legal judgement prediction, and legal information retrieval.

Complex, systematic and precise legal language is often difficult to understand because it sometimes contains some special terminology, old-fashioned words, different grammar styles (especially archaic). This makes it challenging to apply NLP in the field of legal linguistics. The traditional NLP models which were created for language processing generally do not perform well on legal texts. For example, transformer-based models [1] have outperformed classical NLP techniques in the area of law. Such models employ self-attention mechanisms to comprehend every term, sentence or precedent within any given legal document. As a result, this kind of approach enables intricate analysis of complicated legal narratives. Transformers are different from typical models like RNNs [2] and LSTMs [3], which struggle to process complicated legal texts; instead, they analyze documents continuously rather than sequentially. As a result, legal statutes and precedents are better understood.

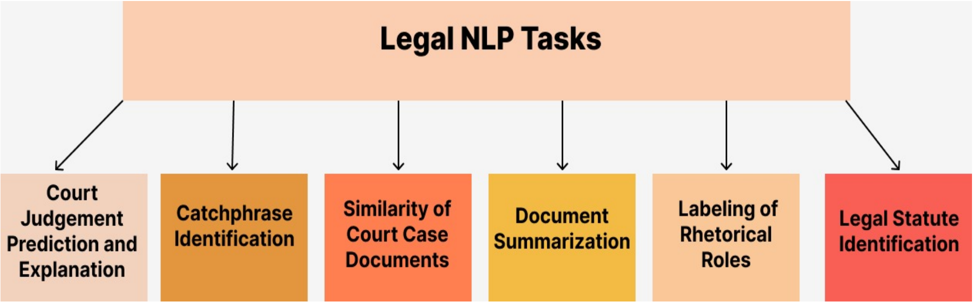

Legal NLP Tasks
Court Judgement Prediction and Explanation
Catchphrase Identification
Similarity of Court Case Documents
Document Summarization
Labeling of Rhetorical Roles
Legal Statute Identification

Fig 1.1 Types of Legal NLP tasks

1.2 Indian Judiciary

The Indian judicial system operates within the framework of the *Common Law system*, relying heavily on the legacy of British colonial rule. This system enables the law to change over time because it mostly relies on earlier court rulings to determine how matters should be decided in the present. The judiciary is organized hierarchically, with the Supreme Court of India at the top, which interprets the Constitution, hears appeals, and decides disputes between the states and the central government. Below the Supreme Court are the High Courts, each serving one or more states, for civil and criminal cases. District courts are the primary courts for most civil and criminal cases in each district. The judicial system is completed by special courts such as family courts and consumer courts, as well as subordinate courts including Magistrate and Sessions Courts for criminal matters.

Indian legal writings are difficult for the general public and lawyers to understand due to their complexity and outdated terminology. Legal system contradictions and inconsistent results come about because several legislations have ambiguous and overlapping provisions. The legal environment gets more complicated when there is a lot of unstructured court filings. It means that some laws continue to exist while being outdated, unable to accommodate changes in either technological advancements or society at large since statutory amendment moves slowly. Multiple case backlogs have led to delays which undermine public confidence and compromise the quick dispensation of justice. This uneven application due to the judge’s cross-country policies has built up perceptions of unfairness and inequality among many people out there. To address these issues, it is essential to reduce the backlog of cases, streamline the legal system, and make major changes in legislation. In order for judiciary systems to enforce law and administer justice effectively, improving readability of legal documents as well as coherence may lead to more successful intervention by courts in India.

1.3 Gender bias in legal context

Gender discrimination in law refers to discrimination or unfair treatment of people based on their sex, which can take many forms in the legal system acts, decisions by courts, legislative strategies and general legal language is an example of this bias. This fosters bias and can lead to inequitable justice, potentially influencing the outcome of court proceedings and public perceptions of the impartiality of the justice system.

Family laws have been known to be a good example of gender bias which dates back in history. For a long time, mothers were given the custody of their children after separation on the premise that women are naturally better nurturers than men in what was a myth. This not only seemed to be best for women but also helped maintain the notion that men are less likely to provide love and care. On the other hand, women who disclose abuse in domestic violence cases often face doubt and mistrust, thereby leaving them with no enough protection and assistance from the judiciary. But mostly victims will always bear the burden of proof that propagates exploitation as well as silence over and over again.

The criminal justice system suffers from gender bias. Research has shown that women tend to receive lighter sentences than men for comparable offenses; This phenomenon is also known as the "chivalry hypothesis." This tolerance is based on the assumption that women are more flexible and less threatening than men. But because they are seen as inconsistent with social and legal norms, women who commit crimes against gender norms—violent crimes—such as that—will be severely punished.

Gender bias may exist in the legal text itself. The use of masculine pronouns in traditional legal documents and statutes sometimes implies the exclusion of women and nonbinary persons. Although still opposed in many countries, efforts are underway to incorporate gender-neutral language into legal proceedings. This linguistic bias reflects and perpetuates profound gender inequalities. To address this bias, efforts must be constantly made to confront and change all forms of discrimination, so that the legal system is fair and equal for all, regardless of gender.

1.4 Legal Language Models

A legal language model is an advanced AI system designed to understand, interpret, and generate human speech legally. This example uses natural NLP techniques to analyze complex legal issues such as statutes, case law, and contracts. Training with large legal documents allows them to understand the unique characteristics of legal language, including key vocabulary and complex sentence structure This enables them to efficiently search relevant court cases, compile records and extract important information.

The development of this model has been greatly enhanced by transformer-based architectures such as BERT [4] and GPT. Specialized translations such as Law2Vec [5], Legal-BERT [6] and InLegalBERT [7] are tailored to specific legal contexts, increasing the ability to understand legal concepts and arguments. Legal language models offer many advantages provide, such as helping to create law and evaluate documents and time and effort when conducting legal research. Defenders can provide initial pleadings and agreements, reduce errors by equalizing assurances, and also ensure that the documentation complies with legal requirements.

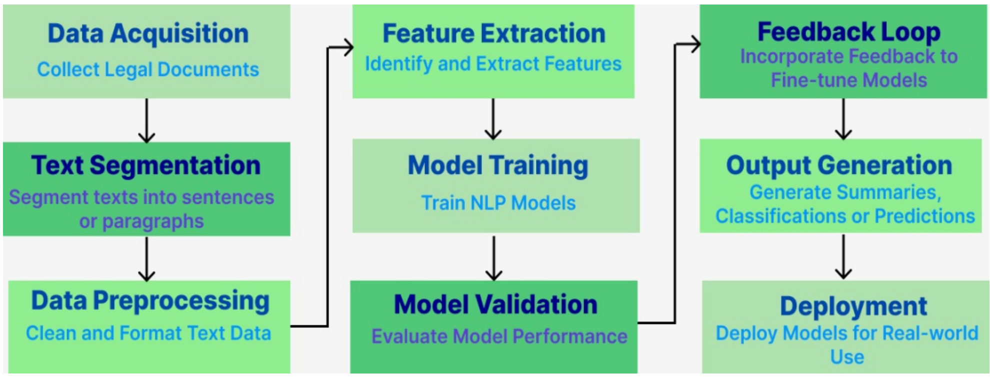

Data Acquisition
Collect Legal Documents
Text Segmentation
Segment texts into sentences or paragraphs
Data Preprocessing
Clean and Format Text Data
Feature Extraction
Identify and Extract Features
Model Training
Train NLP Models
Model Validation
Evaluate Model Performance
Feedback Loop
Incorporate Feedback to Fine-tune Models
Output Generation
Generate Summaries, Classifications or Predictions
Deployment
Deploy Models for Real-world Use

Fig 1.2 Legal Text Analysis Framework

Despite the capabilities of legal language models, there are limitations, especially when it comes to training data quality and bias. Historical biases can be revealed in legal texts and reinforced by examples. To minimize these biases, it is important to ensure that the training data are representative and diverse. These models must also be transparent and reasonable in order to preserve public confidence in the legal system. To address these problems, researchers have attempted to redefine these models. All things considered, legal language models are a major step forward in the application of AI to the legal field, with the potential to revolutionize legal practice and enhance access to justice.

This section briefly introduced the topic. Chapter 2 will provide a review of the existing work that was referred for this work. Chapter 3 will get into the details of our methodology. Chapter 4 will state the results and discuss them. Chapter 5 will point out the limitations of this work. Finally, Chapter 6 concludes this work.

CHAPTER 2
LITERATURE REVIEW

Particularly in a highly populated nation like India, the volume of digital legal documents is growing, which emphasizes the necessity for sophisticated techniques to effectively search and manage this data. This literature review focuses on the relevant work which was taken into consider for our current study. This includes, the task of CJPE [8] considering documents from the Indian Supreme Court, enhancing its performance using jurisdiction-specific pre-trained language models. In addition to these, we have also considered work done in the legal sector to reduce gender bias in judicial documents.

In order to facilitate the development of systems for anticipating and interpreting court judgements, the paper [8] presents the ILDC, a dataset of 35,000 Indian Supreme Court cases annotated with court rulings and expert-provided explanations. The task of CJPE, which involves predicting case outcomes and offering lucid justifications, is put out in the study. After experimenting with several models, the researchers discovered that a hierarchical model that included XLNet [9] and BiGRU [10] worked best, achieving an accuracy of 78% as opposed to 94% for human experts, emphasising the task's difficulty. This study shows how important it is to explain what is behind legal artificial intelligence, presents an outline about how we can account for it through hierarchical occlusion, and observes that there are several key differences between the logic followed by the system under discussion on the one hand, and the one used by people who practice law every day on the other.

In association with three other researchers, the group in [7] has created a new BERT-based model known as CustomInLawBERT. In this model, we focus on the Indian context by incorporating an Indian-specific legal vocabulary into it. Furthermore, it has re-trained LegalBERT [6] and CaseLawBERT [11] using a considerable proportion of nearly 5.4 million Indian legal documents due to its large corpus. In order to evaluate these models, LSI [12], Semantic Segmentation of Judgments [13, 14] and Court Appeal Judgment Prediction [8] were performed. Specifically, inLegalBERT was much better compared with the earlier ones, showing that designing country-oriented legal PLMs can enhance their efficiency. These findings underscored the importance of domain-based prior work as well as the disadvantages involved in applying ordinary language models which for instance are based on BERT in particular case laws.

In the investigation in the [15] study by the writers (Smith et al 2019), it was discovered whether there exists some sort of bias towards gender within legal texts as well as recommendations on how to minimize it. Its gender biases significantly affect fairness in legal decision-making processes as noted by the authors. A variety of word embedding models including Law2Vec were employed for measuring & removing gender bias using training datasets with legal texts across different countries; these included those models such as Law2Vec among others.Methods such as clustering and cosine distance uncovered the level of gender bias, while a new metric, CriBias, was created to measure bias with respect to criminal activities. De-biasing methods were found capable of reducing bias effectively, while preserving the semantic coherence within the embeddings. These findings stress the need for redressing any form of preconceived gender stereotypes within legal NLP systems so as to ensure ethical and fair use of the laws.

In this paper [16], we investigate if transformer-based language models employed by legal products or services are biased towards any particular gender using innovative measurement techniques. To address this issue, one can utilize sophisticated models that are equipped with comprehensive knowledge concerning gender neutrality as among others legal aspects such as Named Entity Recognition (NER), from which the experiment will be conducted. Additionally, these complex models exhibit unique facets of bias compared to previous works that focused on plain static embeddings. The authors propose a new method of assessing bias in BERT-based models which is known as BEC-Cri". This method uses terminology related to crime to assess gender prejudice.

They also suggest a way to fine-tune the debiasing process using an ECtHR corpus, which guarantees that the models' semantic integrity is maintained after debiasing. Their strategy differs from other approaches in that it highlights the necessity of domain-specific debiasing strategies that are tailored for legal scenarios. They also demonstrate notable advancements in bias reduction without sacrificing model performance. Table 2.1 highlights the key findings of our literature review.

Table 2.1 Review of literature associated with prior works

Year, Ref.	Model/Method Used	Best Performance	Dataset	Pros	Cons
2021, [8]	Classical Models(LR,SVM, RF), Sequence Models(BiGRU+ Attn), transformer models (BERT, DistilBERT, RoBERTa, XLNet), and hierarchical transformer models	F1 Score of 78% using XLNet+BiGRU	ILDC curated from 35K court cases from the Supreme Court of India	The temporal features and explainability of the models employed are the main topics of this study. The aim is to shed light on these models' predicted performance across different time frames by investigating how they manage the passage of time in court	The accuracy of the model's predictions and those of human experts differs significantly. The model performs much worse than human experts, with an accuracy rate of only 78%, while human experts reach an incredible 94% accuracy rate. There is a significant difference between the

				cases. The models are also anonymously designed, so the results are not influenced by any identifiable information. This method not only improves the results' objectivity but also tackles important ethical issues with privacy and justice in the judicial system.	legal experts' explanations and the ones produced by the machine models. This disparity emphasises how difficult it is to match the complex and contextualised interpretations that legal experts offer with machine-generated insights. These discrepancies highlight the need for better explainability models.
2022, [15]	Law2Vec, Law2VecNew, UK2Vec, EU2Vec, AU2Vec, CAN2Vec, JAPAN2Vec, Hard Debiasing applied on Projections, Clustering,	Works best on the projection method as all the bias values are reduced to 10^{-8}	HUDOC Dataset, which is made up of ECtHR judgements. It is used to forecast court case decisions	The study provides a worldwide comparative analysis by introducing the CriBias metric, which uses numerous	The difficulty in obtaining the original Law2Vec corpus limits replication and additional research.

	CriBias, EEC	kNN,		in order to evaluate the semantic utility of the debiased embeddings. It contains cases that were available on the HUDOC website as of September 11, 2017.	measuring techniques to assess bias in legal texts. Through the use of POS Tagging and CJP, it is made sure that the models maintain semantic utility. The results of the study are validated by the comprehensive data collection and unique legal embeddings (Law2Vec New).	The failure of debiasing for JAPAN2Vec can be due to the massive computational resources needed to gather, analyse, and train embedding models from large legal corpora. Although debiasing techniques are helpful in eliminating gender bias, it is unclear whether or not they can be applied to other domains or types of biases. Semantic utility verification is still in its early stages and requires more thorough analyses with a

					larger number of tasks and datasets. Other biases, such as racial or ethnic bias, go unnoticed because of the emphasis on gender bias.
2023, [16]	BERT-based models, LegalBERT-Small, and debiasing methods (LCD(proposed), GAP, GPD)	LCD showed the most effective reduction in gender bias	The ECtHR corpus, BEC-Cri for bias measurement, and LexGLUE benchmark for evaluating debiased models	The paper introduced BEC-Cri, a legal-specific bias evaluation corpus using crime-related words, ensuring relevant bias assessment in the legal context. It developed a novel LCD method that significantly reduced gender bias in BERT-based models	Although the work presents a novel debiasing algorithm and corpus, there may be gaps between experimental conditions and practical applications, making it unclear how effective these methods will be in actual legal settings. Large transformer models like BERT can be difficult to fine-tune

	while maintaining semantic utility. The debiased models were evaluated using the LexGLUE benchmark, confirming that debiasing did not compromise performance on legal NLP tasks. Additionally, the paper proposed an innovative bias-penalized performance metric to emphasize the impact of gender bias on model performance	for debiasing due to their complexity and computational expense, which can be especially difficult for smaller businesses or those with constrained computing capabilities. Furthermore, the evaluation metrics employed in the study are primarily focused on gender bias, which may result in the neglect of other significant biases associated with race, ethnicity, or socioeconomic class that may exist in legal texts.

2023, [7]	BERT, LegalBERT, CaseLawBERT, InLegalBERT and InCaseLawBERT retrained on Indian legal corpus, CustomInLawBERT pretrained on Indian specific legal vocabulary	F1 score-64.58 for InLegalBERT for ILSI dataset and 75.88 using InLegalBERT for ECtHR-B dataset in Legal Statute Identification, F1 score of 68.98 using InLegalBERT for ISS data, 61.54 using InLegalBERT for UKSS dataset for Semantic Segmentation of judgements, F1 score-83.09 forInLegalBERT for IDLC dataset for Court Judgement Prediction	ILSI(65K examples from SCI and 6 High Courts), ECtHR-B(11K examples from the European Court of Human Rights), ISS(50 documents from SCI labeled with 7 rhetorical roles), UKSS(50 documents from U.K. Supreme Court labeled with 7 rhetorical roles), ILDC-multi(35K cases from the SCI)	Domain-specific retraining on Indian dataset makes sure the models are better adapted to the peculiarities of the Indian legal language, training a new model from scratch with a custom vocabulary tailored for Indian legal texts presents an innovative approach to handling domain-specific challenges in NLP	Best performance for semantic segmentation of court judgements on Indian data performs worse than the baseline models even though model retraining was specially done on Indian legal data, customized vocabulary tailored from Indian legal texts should have performed the best but didn't due to fewer iterations(lack of resources)

CHAPTER 3
METHODOLOGY

3.1 Requirement Specifications

Our work, which mostly consists of executing various pre-trained models, was carried out in Google Colab due to lack of computational power in the default system and also to ensure faster execution. Table 3.1 specifies the settings in which the work was carried out.

Table 3.1 Requirement Specifications

Requirements	Details
Hardware	NVIDIA A100 Tensor Core GPU
Software	Google Colab Pro
Programming Language	Python (version 3.x)
Framework	PyTorch
Hardware	NVIDIA A100 Tensor Core GPU

3.2 Dataset

For the purpose of solving the Court Judgement Prediction and Explanation (CJPE) problem, annotated cases from the Indian Supreme Court were used to construct the ILDC dataset. Although this dataset is not publicly accessible, it can be acquired by submitting a Google form with a request to the authors.

There are three sections to the dataset:

- *ILDC-multi*: This collection of 35K Supreme Court of India cases contains those in which the same petitioner filed multiple petitions. Because it demands the prediction of numerous, possibly different outcomes from the same case, this dataset poses a considerable difficulty.

• *ILDC-single*: Only instances with a single petition filed are included in this subset of the ILDC-multi dataset.

• *ILDC-expert*: This set of 56 documents, which is annotated with expert explanations for the judgements made, is used to assess the degree to which judgement prediction algorithms are able to both anticipate outcomes and offer reasons that are consistent with expert thinking.

This task can be treated as a classification problem, where our case descriptions, after being tokenized will be treated as input features and the predicted judgement(1: if petition was accepted, 0: if the petition was rejected) will be the output.

For this work, we have used the ILDC-single dataset. It is a subset of the ILDC-multi dataset, and has details of 7593 cases, out of which 80% is used for training the models and rest 20% for testing.

3.3 Contributions in this work

Our work has majorly been divided into 2 major parts:

- Six Transformer-based models, namely bert-base-uncased [17], xlnet-base-cased [18], roberta-base [19], deberta-large [20], electra-large-discriminator [21] and bigbird-roberta-base [22] were applied on the above dataset and their performances were compared. The first 3 models were already used in a prior work, while the later 3 models were utilised in this work. Best performance accuracy of 80% was achieved by the bigbird-roberta-base model.
- A list of 40 different words related to the Indian crime scene was created (see Section 4.2). Using the pre-trained Law2Vec embeddings, the bias in these words was shown using two methods – projection in gender subspace and k-means clustering algorithm. After this, the *Hard Debiasing* algorithm was

applied to mitigate the bias and effects of this debiasing was checked again using the above two methods. Furthermore, to check that debiasing of word embeddings doesn’t tamper with the model performance, we carried out the task of court judgement prediction, using Law2Vec embeddings and SVM classifier, before and after debiasing, and compared their performances.

3.4 Details of models/embeddings/algorithms used

3.4.1 Hyper-parameters for CJP

Only the last 512 tokens from the case descriptions found in the dataset's "text" column were used to create the embeddings and train the model. This was done because prior research has indicated that court decisions typically appear towards the end of case descriptions, making it more effective to use the final 512 tokens for this purpose. Table 3.2 provides specifics on the hyper-parameters that were utilised to train each model while taking into account resource constraints and the optimal results attained after validating accuracies were checked.

Table 3.2 Hyper-parameters set for training

Model	Learning rate	Number of epochs	Batch-size
BERT	5e-6	3	16
XLNet			16
RoBERTa			16
DeBERTa			8
ELECTRA			16
Big Bird			8

3.4.2 Model Details

A company called Hugging Face offers NLP-related tools and technology. It is most known for its 'Transformers' library, which is used by the AI community as a standard to create sophisticated algorithms for a wide range of NLP tasks. Table 3.3 provides details of the models used for this work.

Table 3.3 Specifics of the models used

Model	Layers	Attn - heads	hidden - units	Params	Features
bert-base-uncased	12	12	768	110M	In order to ensure that it can process text consistently regardless of capitalization, this model concentrates on activities that do not require sensitivity to letter case. The model has been trained using two main methods: NSP, which helps the model grasp the link between successive phrases, and MLM, which predicts masked words inside a sentence to assist the model understand context. [4]
xlnet-base-cased	12	12	768	110M	By taking into account various word ordering within a sentence, permutation-based training allows this model to learn from bidirectional contexts. Its capacity to comprehend word associations and context more thoroughly is improved by this

					method. It also uses methods from Transformer-XL, which helps it handle long-range dependencies in text more efficiently. This combination ensures correct text processing and understanding where capitalization counts, making the model especially well-suited for jobs requiring sensitivity to letter case. [9]
roberta-base	12	12	768	125M	This model adjusts the original BERT training process by eliminating the NSP task, which simplifies the training procedure and allows for a more focused approach to understanding sentence-level contexts. Additionally, it trains with larger mini-batches, which enhances the efficiency of the learning process and contributes to better generalization. It also uses a substantially larger data set for training the model, making possible a range of

					varied linguistic patterns and other kinds of language, and thus enhancing both its performance and robustness. [23]
deberta-large	24	16	1024	304M	This model introduces an innovative disentangled attention mechanism, which differentiates between content-based attention and position-based attention. By separating these two types of attention, the model can more effectively capture and utilize the meaning of the words (content) and their positions within the sentence, leading to a more nuanced understanding of the text. Additionally, during the pre-training phase, the model employs an enhanced mask decoder, which improves its ability to predict masked tokens with greater accuracy. This enhancement allows the model to better learn and represent complex linguistic patterns. [24]

electra-large-discriminator	24	16	1024	335M	This model has a special way of teaching where it teaches both a generator and a discriminator at the same time. The generator usually makes ‘fake’ tokens, that are then discriminated by the discriminator who can distinguish them from the ‘real’ ones. This creates competition in learning for the model, so in some way it improves its accuracy in recognizing tokens as well as generating new ones. In addition to that, the reason this method improves the efficiency of learning is that the model uses all input tokens during training as opposed to concentrating only on masked ones, which helps us achieve high accuracy levels by incorporating more useful data in our decision-making processes. Consequently, the entire dataset is understood better

					with respect to language or any other kind of text because this way more extensive information can be derived while drawing upon it as whole, thus providing more efficient results at large since the AI is able to know so much about topics covered in such texts during interpretation. [25]
bigbird-roberta-base	12	12	12	125M	The model involves the rock-solid design of RoBERTa with BigBird by Google Research. By using sparse attention, this technique allows passage of sequences easily past the usual 512-token boundary seen in regular transformers. By combining local, random, and global attention patterns, the model can efficiently handle and analyze long-range dependencies within the text. This innovative approach ensures that the model can capture both fine-grained details and broader contextual

					information, making it exceptionally well-suited for tasks involving lengthy documents and complex contextual relationships. [26]

3.4.3 Law2Vec embeddings

This focused method betters tasks such as document classification, summarization, and legal information retrieval, capturing intricate links and meanings in legal vocabulary. Applying context to discriminate between distinct meanings of terms based on how they are used, Law2Vec boosts performance in different NLP tasks within the legal arena. The training of Law2Vec is the process in which different legal texts are gathered, preprocessed and then vector representations are created based on word co-occurrence patterns using Word2Vec or GloVe algorithms by fine-tuning the embeddings on specific sub-domains, one can improve on exact tasks performance.

Applications of Law2Vec are used include achieving more perfect results in legal research tools, upgraded search outcomes, document classification, extraction of information among others, and even forecasting the legal outcome. The greater research skills, improved information extraction and exact document classification that legal experts have are among what contributes to an enhanced intellectual property field.

3.4.4 Projections to calculate bias

This method of measuring bias draws from the method of [27]; which uses cosine distances as well as projections of word vectors on given gender factors. In that way, gender biases are measured by creating a sub-space for gender within word embeddings. Gender subspaces are usually constructed using pairs of words known to have gender connotations like “man-woman” and “he-she.”

The first thing one must do is to locate pairs of words that have gender perspective. These pairs make sure that a vector can be generated, which is set aside for gender. Once a word vector has been identified, it should be projected in the direction of this «gender» vector, that is situated in «genderized» space. The rest of the vector of whatever word belongs to gender can be assessed pursuant to its projection onto the direction of the gender sub space.

The projection is calculated as the inner product between the vector for gender 'g ' and that of target word ‘w’ where ‘θ’ is the angle between the two.

$$g\,.w = |g||w|\cos\theta \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..\text{(i)}$$

To adapt this method to the legal domain, a list of 40 words (Section 4.2) was taken to reflect terms commonly found in Indian legal texts. This adaptation ensures that the evaluation of bias is relevant to the context in which these embeddings will be applied. By projecting legal-specific word vectors onto the gender subspace, the authors can quantify the gender bias present in legal corpora.

3.4.5 Clustering to calculate bias

This is a version of the projection method described above. The same set of words—which are believed to be gender-neutral but still have significant bias projections—is utilised, and the k-means technique is employed to do unsupervised clustering on the data. The degree to which the resulting clusters adhere to gendered expectations is examined through analysis. Without identifying words as male or female, the clustering method enables researchers to examine established gender biases. Through the method's observation of word clusters, socially marked gender biases can be identified. For example, words with comparable gender biases tend to be adjacent to each other in the embedding space. The approach can measure gender bias even when it is not evident in the clusters directly.

3.4.6 *Hard Debiasing algorithm*

By identifying and eliminating the gender subspace within the embeddings, the Hard Debiasing described in [28] seeks to eradicate or drastically reduce gender bias in word embeddings. There are two primary steps in this process: Neutralise and Equalize.

First, the gender subspace is found by using principal component analysis (PCA) to the gender pair words' difference vectors. For gender pair words like "he-she" and "man-woman," the difference vectors are computed. After calculating the principal components of these vectors, the gender subspace is represented by the first principal component, which primarily captures the gender information.

The component of each word vector w which lies in the gender subspace is calculated and then subtracted from the original word vector during the *Neutralize* step; thus this for every word in the model's vocabulary, the embeddings corresponding to gender neutral words will have no component along this line with one to three words; This operation is achieved by projecting we onto the space spanned by a set of basis vectors$\{b_1, b_2, \dots, b_M\}$ which forms the gender subspace:

$$w_\beta \ = \sum_{i=1}^{M} \langle w, b_i \rangle \, b_i \text{……………………………(ii)}$$

where $\langle .,. \rangle$ denotes the inner product. The neutralized word vector w' is then obtained by subtracting w_β from w and normalizing it:

$$\mathrm{w}' = \frac{\mathrm{w}-\mathrm{w}_\beta}{||\mathrm{w}-\mathrm{w}_\beta||} \text{…………………………………..(iii)}$$

During the *Equalize* step, a set containing equality pairs such as “man-woman” and “king-queen” will be used. We calculate the average Euclidean distance per pair for every pair of words from each other pair before adjusting all pairs so that they will be equidistant from this average distance ensuring that they lie within the zero subspace maintained at all times.

3.4.7 SVM in CJP

A versatile machine learning algorithm known as the SVM is widely used for court judgement prediction. It does this by finding a decision boundary which best separates data points in outcomes of different legal options. By maximizing the margin between classes, it makes it possible for the model to generalize well on new instances. In this context, support vectors are important since they determine where and how the hyperplane should be located with respect to the decision boundary.

SVM finds out which one is appropriate by solving a convex optimization problem during training. When data is not linearly separable in its original feature space, SVM employs the kernel trick to map data into a higher-dimensional space where it becomes linearly separable. Common kernels include linear, polynomial, and RBF, each providing flexibility to handle various types of legal data. This capability is particularly effective in high-dimensional spaces, making SVM suitable for court judgment prediction where features like legal arguments, case facts, and precedents are numerous.

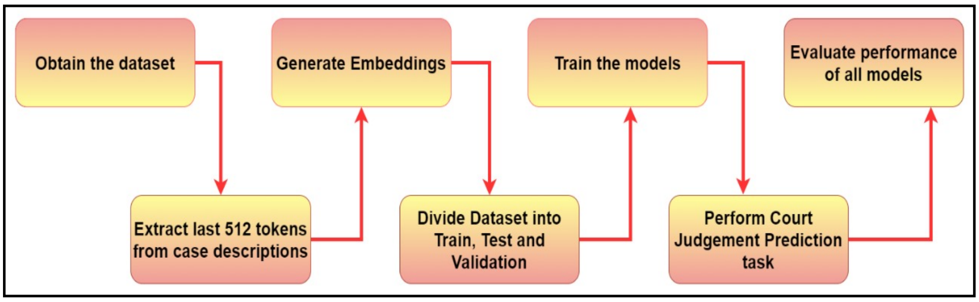

Obtain the dataset
Extract last 512 tokens from case descriptions
Generate Embeddings
Divide Dataset into Train, Test and Validation
Train the models
Perform Court Judgement Prediction task
Evaluate performance of all models

Fig 3.1 Steps to predicting court rulings

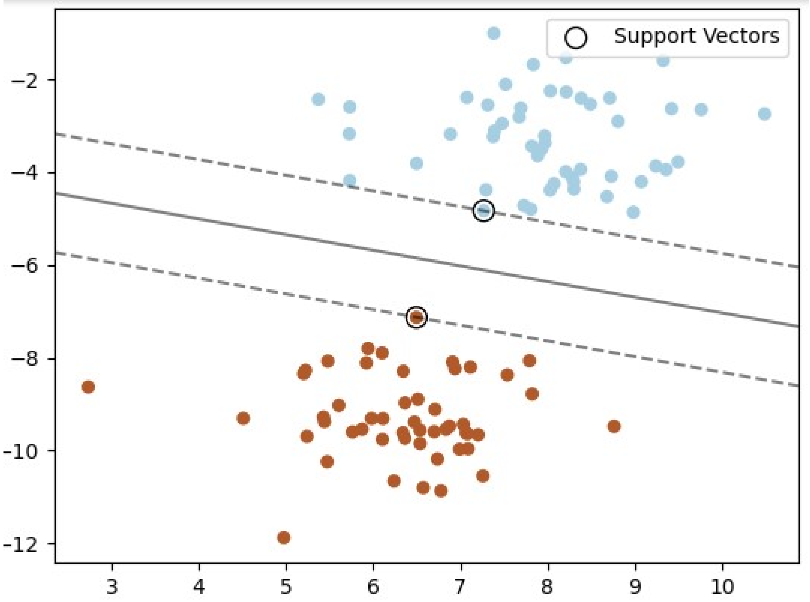


Fig 3.2 Visualizing SVM

CHAPTER 4
RESULTS AND ANALYSIS

4.1 Prediction of Court Judgements

The prior studies employed the Transformer models BERT, XLNet, and RoBERTa; on the ILDC-single dataset, the combination of XLNet and BiGRU produced the best results (the publication reported an accuracy of 76%). In this work, we compared the performance of the newer models—DeBERTa, ELECTRA, and Big Bird—with that of the older versions. The AdamW optimizer was utilised to train our models, and binary cross-entropy was employed to calculate the loss. Metrics including macro-F1 (mF1), accuracy, macro-precision (mP), and macro-recall (mR) were used to track the performance of the model (Table 4.1).

Table 4.1 Performances of different models on CJP

Model	mP (%)	mR (%)	mF1 (%)	Accuracy (%)
BERT	66.78	66.74	66.76	67.94
XLNet	68.77	66.85	67.16	69.65
RoBERTa	73.04	69.17	69.62	72.61
DeBERTa	77	75.80	76.22	77.49
ELECTRA	77.56	77.16	77.33	78.08
Big Bird	**82.40**	**79.15**	**79.15**	**80.97**

The aforementioned table indicates that bigbird-roberta-base exhibits the maximum accuracy of 80.97%, which is roughly 4% higher than the results reported in the original work. The model produced a mF1 score of 79.15%, which is roughly 3% higher than the 76.55% score reported in the original work.

Out of all the models used in this work, bigbird-roberta was able to provide the greatest results in terms of accuracy, precision, recall, and F-1 score. However, only its base version was utilised, and it was only ran for a batch-size of 8, which is half that used for the predecessors. Additionally, this model performed better than its equivalents, the DeBERTa and ELECTRA models, which were based on the 'large' category and hence included a significantly higher number of parameters.

This result shows how well the bigbird-roberta architecture naturally extracts relevant details from the data. Bigbird-Roberta's proficiency at processing lengthy sequences (beyond the standard 512 tokens in other models) may account for some of this outperformance by enabling a thorough contextual analysis. Furthermore, combining BigBird with RoBERTa's pretraining approach may enhance representational learning capabilities. Lastly, there's a chance that the smaller batch and model sizes had an unintentional regularising effect that prevented overfitting and improved generalisation. It is important to investigate these options more in order to fully comprehend why such a result was obtained.

4.2 Bias Calculation and Reduction in Law2Vec

4.2.1 Using projections

The projections provided in the table below measure the bias associated with each word, with negative values indicating 'bias towards males' and positive values indicating 'bias towards females'. After debiasing, the projections should ideally be close to zero, indicating reduced bias.

Table 4.2 Outcomes before and after debiasing using projections

Word in Indian CriList	Projection before debiasing	Projection after debiasing
crime	-0.1497	$1.2\text{x}10^{-8}$
offense	-0.0008	$1.3\text{x}10^{-8}$
accused	0.0155	$4.8\text{x}10^{-8}$
victim	-0.6096	$-2.6\text{x}10^{-8}$
case	0.2478	$3.8\text{x}10^{-8}$
police	0.1780	$4.3\text{x}10^{-8}$

evidence	0.1791	-4.4x10^{-8}
arrest	-0.3599	1.8x10^{-8}
suspect	-0.3204	1.8x10^{-8}
witness	0.0024	4.9x10^{-8}
court	0.3171	0.5x10^{-8}
trial	0.3016	-0.4x10^{-8}
sentence	0.2909	3.2x10^{-8}
conviction	0.0789	1.5x10^{-8}
bail	0.0049	1.0x10^{-8}
charge	0.0733	-0.4x10^{-8}
report	0.1869	1.9x10^{-8}
testimony	-0.0001	1.0x10^{-8}
lawyer	-0.2493	-6.4x10^{-8}
judge	0.3453	1.8x10^{-8}
jury	0.1435	2.4x10^{-8}
detention	-0.3256	-0.2x10^{-8}
prosecution	-0.0162	2.2x10^{-8}
defense	0.4121	-13.5x10^{-8}
law	-0.2115	0.9x10^{-8}
order	0.1693	4.6x10^{-8}
statement	0.3332	-0.01x10^{-8}
forensic	-0.1714	-1.4x10^{-8}
inquiry	0.1925	1.4x10^{-8}
raid	-0.4983	-0.7x10^{-8}
custody	-0.5281	1.9x10^{-8}
hearing	0.3384	1.4x10^{-8}
allegation	-0.1531	2.7x10^{-8}
confession	-0.1481	-2.2x10^{-8}
documentation	-0.1339	0.5x10^{-8}

interrogation	-0.0758	1.8x10^{-8}
Search	-0.1809	-0.25x10^{-8}
Surveillance	-0.1207	0.3x10^{-8}
Remand	-0.0428	1.5x10^{-8}
Seizure	0.1078	2.2x10^{-8}

We can see that before debiasing, words like ‘victim’ (-0.6096), ‘arrest’ (-0.3599), ‘suspect’ (-0.3204), ‘detention’ (-0.3256), and ‘custody’ (-0.5281) have significant negative projections, indicating a strong bias towards males while words like ‘court’ (0.3171), ‘judge’ (0.3453), ‘defense’ (0.4121), ‘statement’ (0.3332), and ‘hearing’ (0.3384) have notable positive projections, indicating bias towards females.

The number line below also shows that the word ‘victim’ has the most ‘maleness’ whereas ‘defense’ has the most ‘femaleness’.

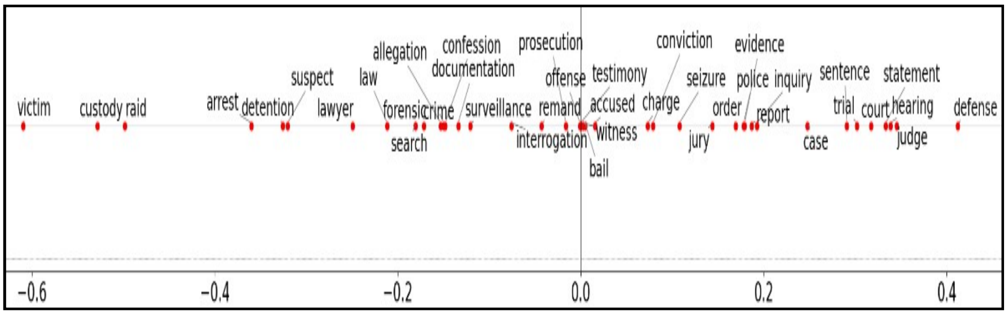


Fig 4.1 Biases of words plotted in number line

After debiasing, the majority of words show projections very close to zero, such as "crime" (1.2x10^-8), "offense" (1.3x10^-8), "accused" (4.8x10^-8), "case" (3.8x10^-8), and "police" (4.3x10^-8), indicating that the debiasing process was effective in reducing bias. Additionally, the consistency observed across words like "court" (0.5x10^-8), "trial" (-0.4x10^-8), "sentence" (3.2x10^-8), and "prosecution" (2.2x10^-8) further supports the effectiveness of the debiasing method.

The initial projections reveal inherent biases in the Law2Vec embeddings, with words associated with crime and legal processes showing varying degrees of bias, reflecting potential underlying prejudices in the training dataset. For example, "victim" had a strong negative projection, indicating possible bias against victims, while "defense" had a strong positive projection, suggesting bias in favor of the defense. Nevertheless, the reduction in most words' projection towards zero after debiasing suggests that these biases were successfully neutralized. This suggests that the applied debiasing techniques effectively mitigated the biases in the Law2Vec embeddings, leading to a more balanced representation of terms related to the Indian criminal justice system.

4.2.2 Using clustering

The initial bias in the Law2Vec embeddings was calculated using clustering techniques and visualized with 2D plots, which highlighted the noticeable biases in the grouping of gendered words. This clustering analysis revealed distinct clusters of words typically associated with male and female stereotypes, indicating significant bias in the word representations. This can be seen in Figure 4.1.

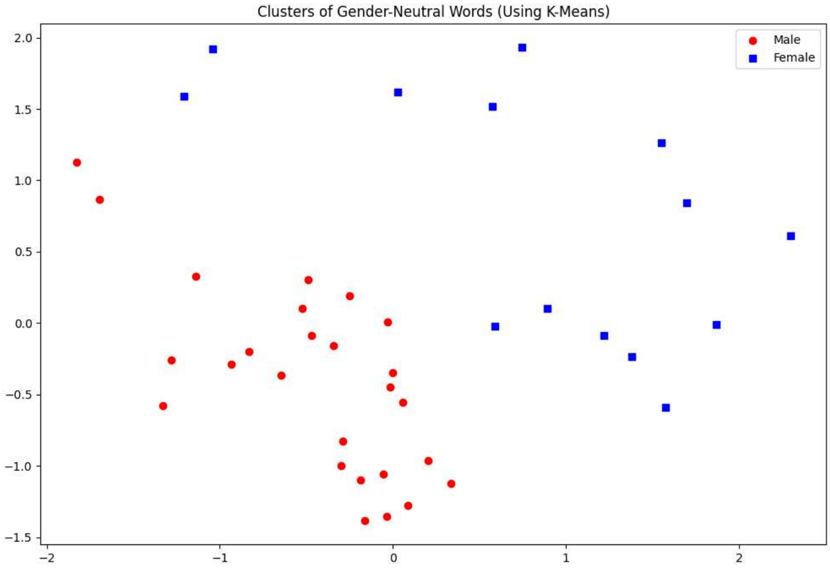


Fig 4.2 Clusters of gender-neutral words before debiasing

From Figure 4.2, it can be observed that after applying hard debiasing, the clustering of gendered words was significantly less pronounced, indicating a reduction in bias. However, some semantic distortions were observed, where certain words lost some of their contextual meanings.

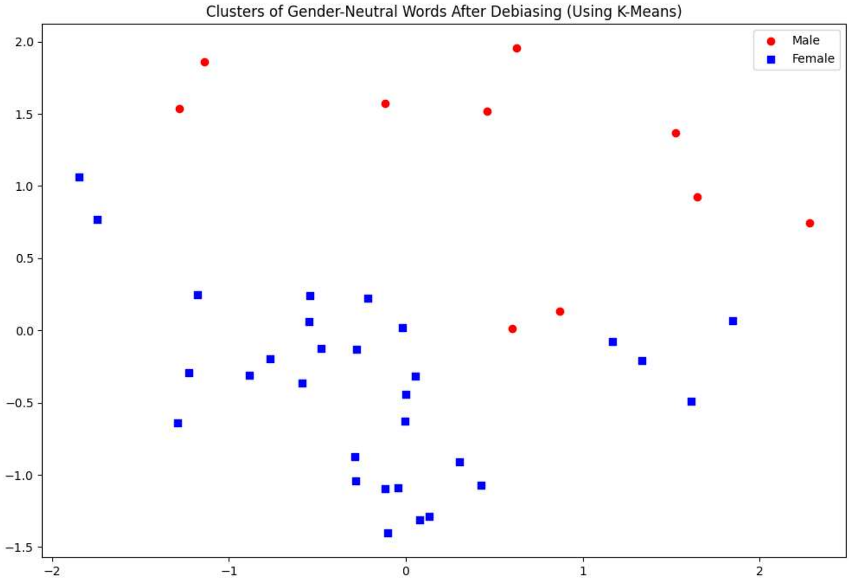


Fig 4.3 Clusters of gender-neutral words after debiasing

4.2.3 Confirming effectiveness of debiasing with CJP

The debiasing technique applied to Law2Vec embeddings seems to have a slight positive impact on the accuracy of the classifier, suggesting a potential improvement in its overall performance.

However, the other metrics (macro-precision, macro-recall, and macro-F1) remain unchanged, indicating that the improvement in accuracy does not translate into better precision, recall, or F1 score. The values are noted in the table below.

Table 4.3 Outcome on CJP before and after debiasing

Law2Vec embeddings with SVM as classifier	**Accuracy**	**Macro-Precision**	**Macro-Recall**	**Macro-F1**
With bias	0.63	0.62	0.62	0.62
Without bias	0.64	0.62	0.62	0.62

The results show that the debiasing process does not significantly alter the usefulness of legal word embeddings for high-level NLP tasks, such as CJP. The accuracy scores obtained using debiased embeddings are generally very close to, or the same as, those obtained using biased embeddings. Reducing gender bias without sacrificing the embeddings' semantic value is the main goal of debiasing. The conclusion that the semantic structure is well-preserved post-debiasing is supported by consistent performance on tasks such as CJP. Debiasing is done to make embeddings ethically sound without sacrificing their usefulness in real-world applications. The goal is to ensure that debiased embeddings function on par with original embeddings while minimising bias.

CHAPTER 5
LIMITATIONS AND FUTURE SCOPE

Although we worked on a small scale here, there is room for growth. This thesis part identifies some of the shortcomings in our findings and makes recommendations for further investigation.

i) ***Lack of explainability***: Despite having decent prediction accuracy, the models used in this work lack inherent explainability techniques, which cause opacity in their decision-making. The outputs of the models may become less trustworthy for legal professionals who require precise explanations and reasoning in important situations due to this lack of transparency. Subsequent research ought to concentrate on incorporating explainability methods like attention visualisation, feature importance scoring, and natural language explanations straight into the model architecture. It will be essential to create explainability-specific models employing techniques like rule-based systems, LIME [29], and SHAP []. Further refining of the models can be guided by performing user-centric assessments with legal professionals to evaluate the effectiveness and clarity of the explanations. This will help to ensure that the models satisfy the practical demands of legal practitioners.

ii) ***Enhancing Dataset and Model Capabilities****:* For the present study, ILDC-single was used, which is a shorter version of the main ILDC-multi dataset. The generalizability and robustness of the models will be improved by increasing the dataset's inclusion to include additional instances from different Indian courts. By making the dataset available to the public, more research and validation activities will be possible, improving the reproducibility of results. Performance could be further enhanced by investigating sophisticated transformer layouts and hybrid models that incorporate transformer and non-transformer technologies. To improve the performance of the model, extensive hyperparameter optimisation and fine-tuning—possibly with more computer resources—will also be essential. The original work [8] has also incorporated additional layers of BiGRU and attention, which we would also like to address in the future.

iii) ***Dealing with other kinds of biases****:* The primary focus of this work was gender bias in legal language models; however, since other types of bias, such racial, ethnic, and socioeconomic prejudices, can also have an impact on the fairness and dependability of AI models in legal contexts, it is essential to expand the scope of debiasing efforts. The creation of techniques to recognise and lessen these extra biases would guarantee more just results and improve the moral use of AI in the legal sector. Furthermore, thorough analyses and improvements to these debiasing methods are required to preserve the models' semantic integrity while reducing bias in all its manifestations.

iv) ***Use of contextualized embeddings****:* This work has used Law2Vec, which is static and the bias calculations methods are also distance based. Implementing contextualized embeddings, such as those derived from models like BERT, and their legal-specific counterparts (LegalBERT, InLegalBERT etc) can significantly enhance the understanding of complex legal texts. These embeddings capture the nuanced meanings of words based on their context within a document, providing more accurate representations compared to static embeddings. Due to this, we have to implement different methods for bias calculation and reduction. Such techniques have already been dealt with in [16], but not in the Indian context, and we plan to take this into consideration as well. More details about these methods are mentioned in the Appendix.

CHAPTER 6
CONCLUSION

This thesis explored the intersection of legal linguistics and gender bias within NLP models tailored for legal judgment prediction, leveraging transformer-based models like BERT, XLNet, RoBERTa, DeBERTa, ELECTRA, and BigBird, evaluated on the ILDC-single dataset. BigBird-RoBERTa demonstrated superior performance in accuracy, precision, recall, and F1 score due to its capacity to process longer sequences and extract relevant contextual information. Additionally, we examined gender bias in legal language models, particularly focusing on Law2Vec embeddings. Using projection and clustering techniques, significant biases were identified and effectively mitigated with the Hard Debiasing algorithm, achieving near-zero projections post-debiasing while maintaining semantic integrity. This debiasing process did not compromise model performance in court judgment prediction tasks. Therefore, legal professionals will have a problem, in the future, which they will solve by combining interpretability into their models because such models do not currently provide any. Expanding the dataset and incorporating advanced transformer architectures could further improve robustness and generalizability. To ensure that fairness and reliability are upheld in legal AI models, it is important to deal with other biases such as racial ones. In other words, introducing contextualized embeddings and developing comprehensive debiasing methods tailored to the Indian legal system will advance this field. To sum it up, this thesis shows how effective transformer-based models can be in predicting legal judgments; and the importance of mitigating gender bias in legal language models for more ethical, transparent and efficient AI tools applicable to legal industry. On the other hand, there can be no explanation of how these systems work by lawyers who must incorporate them within their own practices. Therefore, future incorporation of explainability techniques is warranted given that these simple opaque models cannot be easily explained by law practitioners. It may also necessitate expanding the dataset or including more robust architectures like transformer in order to make those systems more reliable.

Legal AI should also shun all kinds of societal biases (e.g., racial) so as to guarantee that fairness as well as dependability is maintained within it. Additionally, implementing contextualized embeddings and developing comprehensive debiasing methods designed specifically for India's legal system would be an advancement of the field itself. This definitely demonstrates the efficiency of transformer-based model in predicting judicial decisions legally resulting from mitigation of gender prejudice within language modelling meant for law thereby giving rise to fairer and improved artificial intelligence tools used in law practice today.

REFERENCES

[1] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017)

[2] Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." nature 323, no. 6088 (1986): 533-536

[3] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9, no. 8 (1997): 1735-1780

[4] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[5] Chalkidis, I. and Kampas, D., 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artificial Intelligence and Law, 27(2), pp.171-198.

[6] Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. "LEGAL-BERT: The muppets straight out of law school." arXiv preprint arXiv:2010.02559 (2020)

[7] Paul, Shounak, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. "Pre-trained language models for the legal domain: a case study on Indian law." In Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, pp. 187-196. 2023

[8] Malik, Vijit, Rishabh Sanjay, Shubham Kumar Nigam, Kripa Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. "ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation." arXiv preprint arXiv:2105.13562 (2021)

[9] Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. "Xlnet: Generalized autoregressive pretraining for language understanding." Advances in neural information processing systems 32 (2019).

[10] Schuster, M. and Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE transactions on Signal Processing, 45(11), pp.2673-2681.

[11] Zheng, Lucia, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. "When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings." In Proceedings of the eighteenth international conference on artificial intelligence and law, pp. 159-168. 2021

[12] Bhattacharya, Paheli, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. "Hier-spcnet: a legal statute hierarchy-based heterogeneous network for computing legal case document similarity." In Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, pp. 1657-1660. 2020

[13] Ghosh, Saptarshi, and Adam Wyner. "Identification of rhetorical roles of sentences in indian legal judgments." Legal knowledge and information systems: JURIX (2019): 3

[14] Bhattacharya, Paheli, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. "DeepRhole: deep learning for rhetorical role labeling of sentences in legal case documents." Artificial Intelligence and Law (2023): 1-38

[15] Sevim, N., Şahinuç, F. and Koç, A., 2023. Gender bias in legal corpora and debiasing it. Natural Language Engineering, 29(2), pp.449-482.

[16] Bozdag, M., Sevim, N. and Koç, A., 2024. Measuring and mitigating gender bias in legal contextualized language models. ACM Transactions on Knowledge Discovery from Data, 18(4), pp.1-26.

[17] https://huggingface.co/google-bert/bert-base-uncased

[18] https://huggingface.co/xlnet/xlnet-base-cased

[19] https://huggingface.co/FacebookAI/roberta-base

[20] https://huggingface.co/microsoft/deberta-large

[21] https://huggingface.co/google/electra-large-discriminator

[22] https://huggingface.co/google/bigbird-roberta-base

[23] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[24] He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. "Deberta: Decoding-

enhanced bert with disentangled attention." arXiv preprint arXiv:2006.03654 (2020).

[25] Clark, Kevin, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. "Electra: Pre-training text encoders as discriminators rather than generators." arXiv preprint arXiv:2003.10555 (2020).

[26] Zaheer, Manzil, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham et al. "Big bird: Transformers for longer sequences." Advances in neural information processing systems 33 (2020): 17283-17297.

[27]Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V. and Kalai, A.T., 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29.

[28] Gonen, H. and Goldberg, Y., 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv preprint arXiv:1903.03862.

[29] Garreau, D. and Luxburg, U., 2020, June. Explaining the explainer: A first theoretical analysis of LIME. In International conference on artificial intelligence and statistics (pp. 1287-1296). PMLR.

[30] Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[31] Webster, K., Recasens, M., Axelrod, V. and Baldridge, J., 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. Transactions of the Association for Computational Linguistics, 6, pp.605-617.

[32] Kaneko, M. and Bollegala, D., 2019. Gender-preserving debiasing for pre-trained word embeddings. arXiv preprint arXiv:1906.00742.

APPENDIX

A. Evaluating and Reducing Discrimination Based on Gender in Legal Contextualised Language Models

Transformer-based contextual language models have revolutionized NLP work, including work in law. Despite the improvements, these models often have human-like biases in training data, causing serious ethical concerns especially in law. Previous studies have shown gender bias in law static embeddings such as Law2Vec and contextualized NLP models (LegalBERT). These biases can affect the fairness and accuracy of applications, making it crucial to develop methods to measure and mitigate such biases. In the legal domain, where decisions can have profound impacts on individuals' lives, eliminating these biases is particularly important. Thus it is important to focus on evaluating and reducing gender bias in BERT-based models applied to legal text processing. Our present research has not considered this but it is a potential research gap for us, where would like to address with respect to the Indian judiciary.

Bias Evaluation

The bias measurement process is mainly based on the MLM properties of BERT-based models, which allow estimation of probability values of words covered in a sentence. This exercise requires the construction of a new bias assessment corpus drawn by BEC-Cri focuses on legal applications and includes template judgments with criminal information from the FBI database. Example sentences from the BEC-Cri included gender-related target words (e.g., 'she', 'he') and crime-related words (e.g., 'murder', 'embezzlement'). The method uses BERT's MLM feature to quantify gender bias, estimating the probability of a gender-specific pronoun or nouns occurring in a particular context to identify a model dependent on gender. This covers target words in a sentence to count how gender pronouns occur in that context.

The process begins with preparing a sentence with a target and an attribute. The target word is then masked, and the model predicts the probability that the target word will appear in that location. Then, the words and the target word are masked, and the model predicts the probability of the target word occurring in this new context. An association score is determined by comparing these probabilities, indicating how much the presence of the attribute affects the probability of the target term.

To ensure a thorough analysis, the average associaiton score starts at zero and is updated as each sentence in the corpus is analyzed. For each sentence, the likelihood is calculated that the target word appears first in the masked sentence. Then, the probability is recalculated by overlapping the target and attribute words. The difference between these possibilities yields an association score for the sentence. This score is summed across all sentences to update the association score. The final mean association score is calculated by averaging the accumulated scores and it shows how much the language model was biased towards associating gendered words with particular crime-related terms. Hence, this procedure helps to show how a language model can have gender bias in legal texts through counting how often it associates between gendered terms and specific law attributes

Bias Calculation Techniques

- **GAP**

  The GAP debiasing method aims to bring gender information in the model into a state of equilibrium by means of fine-tuning. The GAP corpus, developed by [31], comprises 8,908 human-annotated gender-ambiguous pronoun-name pairs taken from Wikipedia. This dataset is used in a coreference resolution task where the model acquires the ability to correctly associate gender-ambiguous pronouns with their respective names. To tackle bias on account of gender, this method tries to expose the model to an equal number of male and female targets. Again, further balancing the dataset involves performing counterfactual data substitution (CDS) on it such as swapping some words that are gender-specific in nature. The fine-tuning process employs this balanced corpus for adjusting the parameters of the model aimed at reducing its inclination towards associating certain genders with specific contexts. This enables encoding more balanced gender information in the model leading to decreased gender bias in predictions made by models.

- **GPD**

The GDP method, advanced by [32] is a way of biasing untrained contextualized models while keeping gender information that is necessary. The present approach employs the News-Commentary-v15 corpus and extracts numerous sentences with feminine, masculine and gender-stereotypical words. These words are grouped into categories of attributes and targets, thereby creating different sets of sentences for each category. The GPD method minimizes the dependency of the embedding for target words with respect to gender attributes using a similarity based loss term. This is done by deriving the non-contextualized word embedding for each attribute and target word and employing them to adjust the target word embeddings. In order to preserve original information in the embeddings so as not to have great information loss during debiasing, a regularizing loss is included in this method. However, despite these interventions; there have been mixed results on GPD which exhibits some decrease in bias while in other instances its bias scores remain high or even increase further.

- **LCD**

Following a particular technique, the LCD method [16] was created in order to combat gender bias across domain-specific legal contextualized language models like LegalBERT-Small. This implies that biases may be deeply embedded within context-dependent embeddings generated by BERT-like transformer models. In this respect, the LCD method involves training the model through a fine-tuning process using a curated gender-neutral legal corpus. Specifically, the fine-tuning dataset consists of court cases from the ECtHR where equal number of male and female applicants have been manipulated into it. The aim of fine-tuning is to ensure that the model can accurately identify cases that are characterized with gendered contexts without being influenced by gender bias itself. By utilizing this balanced dataset, LCD method effectively mitigates any subconscious preference for specific genders in the model's predictions hence resulting into more unbiased and fair decisions made by it. This way, it makes it possible to keep most of what was learned about core semantic knowledge preserved while greatly reducing its gender-related aspect which is necessary for maintaining performance in legal language understanding tasks at hand.

LIST OF PUBLICATIONS

1. Trisha Ghosh, Shailender Kumar, "A Survey of Legal Text Analysis Techniques for Indian Legal Documents". Accepted at the International Conference on Circuit, Systems and Communication (ICCSC 2024).

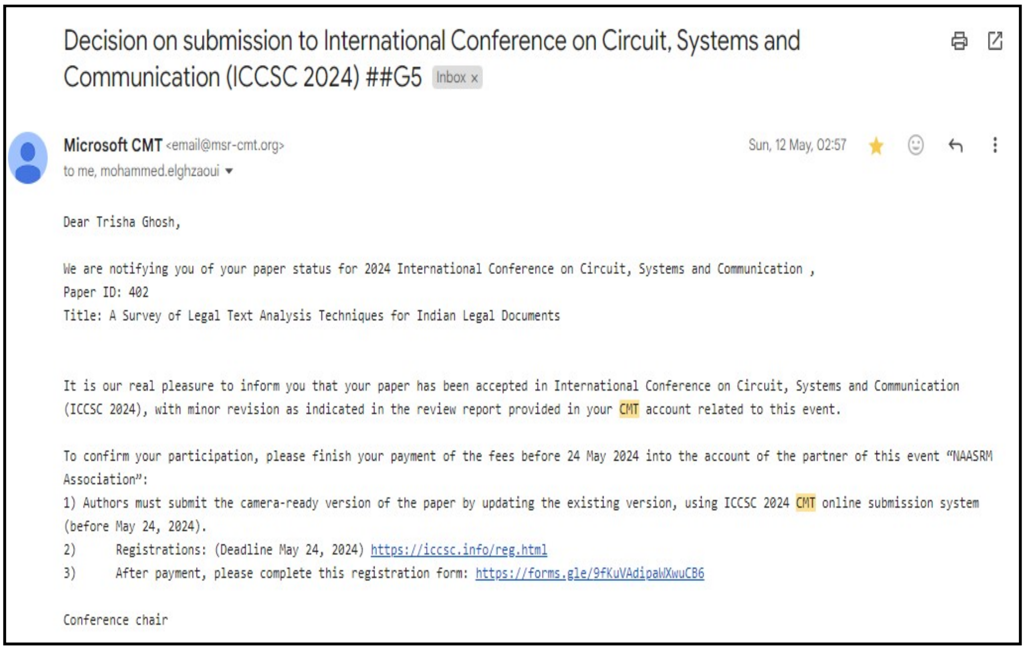
   Paper ID: 402, Indexed by Scopus.

The 2024 IEEE International Conference on Circuit, Systems and Communication (ICCSC 2024)

June 2024 ,28-29, Fez, Morocco

Indexed by :

Scopus



Decision on submission to International Conference on Circuit, Systems and Communication (ICCSC 2024) ##G5

Microsoft CMT <email@msr-cmt.org>

to me, mohammed.elghzaoui

Sun, 12 May, 02:57

Dear Trisha Ghosh,

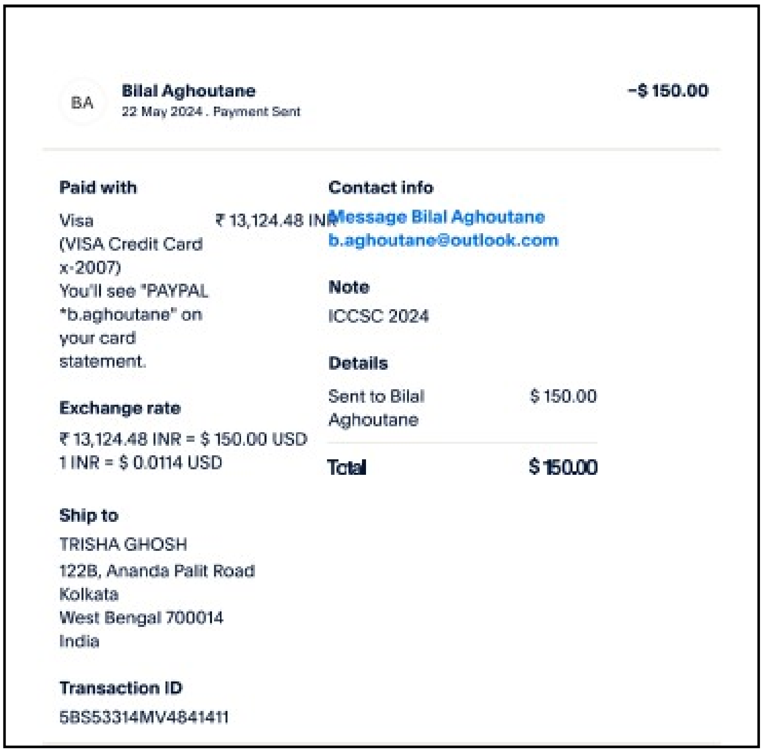
We are notifying you of your paper status for 2024 International Conference on Circuit, Systems and Communication ,
Paper ID: 402
Title: A Survey of Legal Text Analysis Techniques for Indian Legal Documents

It is our real pleasure to inform you that your paper has been accepted in International Conference on Circuit, Systems and Communication (ICCSC 2024), with minor revision as indicated in the review report provided in your CMT account related to this event.

To confirm your participation, please finish your payment of the fees before 24 May 2024 into the account of the partner of this event "NAASRM Association":

1) Authors must submit the camera-ready version of the paper by updating the existing version, using ICCSC 2024 CMT online submission system (before May 24, 2024).
2) Registrations: (Deadline May 24, 2024) https://iccsc.info/reg.html
3) After payment, please complete this registration form: https://forms.gle/9fKuVAdipaWXwuCB6

Conference chair



BA **Bilal Aghoutane** -$ 150.00

22 May 2024 . Payment Sent

Paid with

Visa ₹ 13,124.48 INR
(VISA Credit Card
x-2007)
You'll see "PAYPAL
*b.aghoutane" on
your card
statement.

Exchange rate

₹ 13,124.48 INR = $ 150.00 USD
1 INR = $ 0.0114 USD

Ship to

TRISHA GHOSH
122B, Ananda Palit Road
Kolkata
West Bengal 700014
India

Transaction ID

5BS53314MV4841411

Contact info

Message Bilal Aghoutane
b.aghoutane@outlook.com

Note

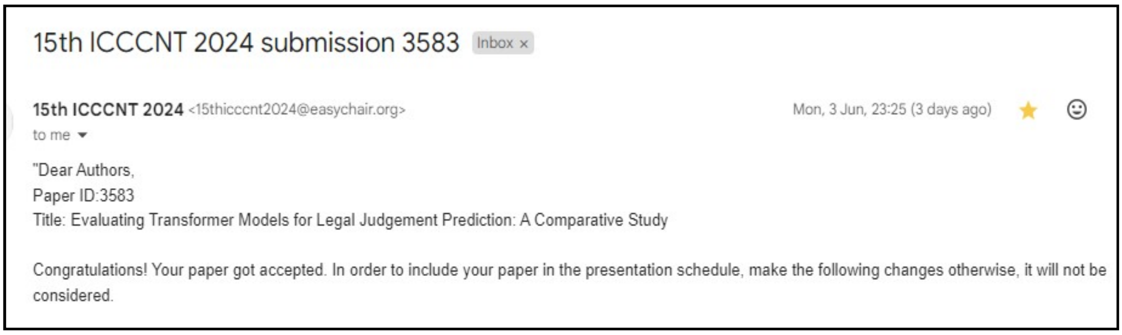
ICCSC 2024

Details

Sent to Bilal Aghoutane	$ 150.00
Total	**$ 150.00**

2. Trisha Ghosh, Shailender Kumar, “Evaluating Transformer Models for Legal Judgement Prediction: A Comparative Study”. Accepted at the 15th International IEEE Conference on Computing, Communication and Networking Technologies (ICCCNT 2024).

   Paper ID: 3583, Indexed by Scopus.



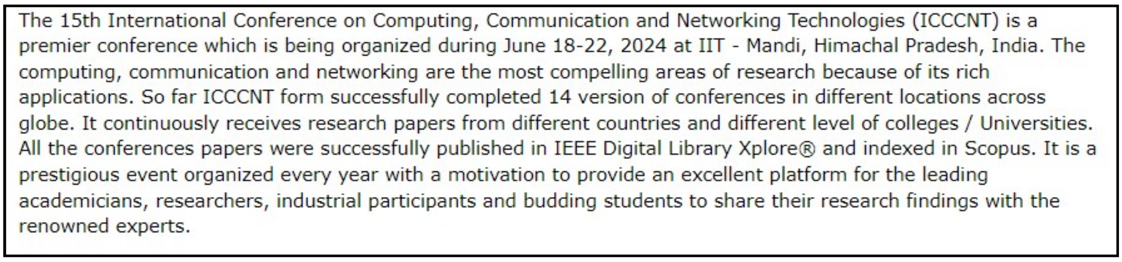
15th ICCCNT 2024 submission 3583 Inbox ×

15th ICCCNT 2024 <15thicccnt2024@easychair.org>
to me

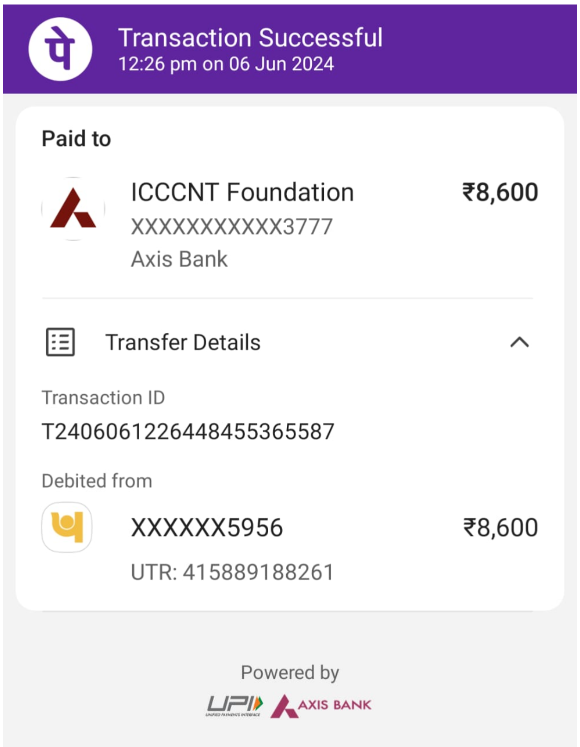
Mon, 3 Jun, 23:25 (3 days ago)

"Dear Authors,
Paper ID:3583
Title: Evaluating Transformer Models for Legal Judgement Prediction: A Comparative Study

Congratulations! Your paper got accepted. In order to include your paper in the presentation schedule, make the following changes otherwise, it will not be considered.



The 15th International Conference on Computing, Communication and Networking Technologies (ICCCNT) is a premier conference which is being organized during June 18-22, 2024 at IIT - Mandi, Himachal Pradesh, India. The computing, communication and networking are the most compelling areas of research because of its rich applications. So far ICCCNT form successfully completed 14 version of conferences in different locations across globe. It continuously receives research papers from different countries and different level of colleges / Universities. All the conferences papers were successfully published in IEEE Digital Library Xplore® and indexed in Scopus. It is a prestigious event organized every year with a motivation to provide an excellent platform for the leading academicians, researchers, industrial participants and budding students to share their research findings with the renowned experts.



Transaction Successful
12:26 pm on 06 Jun 2024

Paid to

ICCCNT Foundation ₹8,600
XXXXXXXXXXX3777
Axis Bank

Transfer Details

Transaction ID
T2406061226448455365587

Debited from
XXXXXX5956 ₹8,600
UTR: 415889188261

Powered by

PAPER NAME

thesis_finaldraft.pdf

WORD COUNT	CHARACTER COUNT
10703 Words	60110 Characters
PAGE COUNT	FILE SIZE
58 Pages	3.1MB
SUBMISSION DATE	REPORT DATE
May 30, 2024 2:28 PM GMT+5:30	May 30, 2024 2:29 PM GMT+5:30

● 7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 6% Internet database
- Crossref database
- 6% Submitted Works database
- 3% Publications database
- Crossref Posted Content database

● Excluded from Similarity Report

- Bibliographic material
- Cited material
- Quoted material
- Small Matches (Less then 8 words)