SMALL LANGUAGE MODELS FOR SRI LANKAN LEGAL APPLICATIONS

PREDICTS LIKELY LEGAL JUDGMENTS BY ANALYZING PAST CASES IN CRIMINAL LAW

Project Final Report

IT22049322 - Abiramy.T

Bachelor of Science (Hons) Degree in Information Technology
Specializing in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology Sri Lanka

April 2026

SMALL LANGUAGE MODELS FOR SRI LANKAN LEGAL APPLICATIONS

PREDICTS LIKELY LEGAL JUDGMENTS BY ANALYZING PAST CASES IN CRIMINAL LAW

Project Final Report

IT22049322 - Abiramy.T

Dissertation submitted in partial fulfillment of the requirements for the Bachelor of Science (Hons) Degree in Information Technology Specializing in Information Technology

Department of Information Technology

Sri Lanka Institute of Information Technology Sri Lanka

April 2026

DECLARATION

I declare that this is my own work and this Dissertation does not incorporate without acknowledgment any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgment is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Name: Abiramy.T Student ID: IT22049322 Signature:

The above candidate has carried out research for the Bachelor of Science (Hons) Degree in Information Technology Specializing in Information Technology Dissertation under our supervision. We confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Prasanna Sumathipala

Signature of the Supervisor: Date:

Name of Co-Supervisor: Ms.Karthiga Rajendran

Signature of the Co-Supervisor: Date:

DEDICATION

This thesis is dedicated to all those who provided guidance, support, and encouragement throughout my postgraduate studies, and whose contributions both direct and indirect-played an important role in the successful completion of this work.

ACKNOWLEDGEMENT

I would like to sincerely thank everyone who supported and guided me throughout this research. Their encouragement, constructive feedback, and thoughtful advice played an important role in shaping this work. I am especially grateful for the valuable discussions and perspectives shared, which helped me better understand several aspects of the research.

I would like to express my special thanks to my supervisor, Dr. Prasanna Sumathipala and my co-supervisor, Ms.Karthiga Rajendran, for their continuous guidance, patience, and academic support throughout this study. His insights, suggestions, and regular discussions were instrumental in refining my research approach and improving the overall quality of this thesis.

I am also grateful to the University of Moratuwa, particularly the Department of Computer Science and Engineering, for providing a supportive academic environment.

Finally, I would like to thank all those who contributed, directly or indirectly, to the completion of this work. Their support made this research journey both productive and meaningful.

ABSTRACT

Judicial decision-making in criminal cases involves understanding laws, evidence, and legal procedures. With the increasing availability of digital court judgments in Sri Lanka, it is now possible to use computer-based methods to study legal texts in a structured way. This study focuses on using modern Natural Language Processing (NLP) techniques to analyze past criminal case decisions in Sri Lanka, especially from the High Courts and the Supreme Court.

For this research, a dataset of about 890 criminal case judgments from 2021 to 2025 was collected from publicly available legal sources. These judgments were converted into a structured format (JSON) using a controlled pre-processing method supported by a language model. Each case was then labeled with one outcome category so that it could be used for classification.

A transformer-based model called LEGAL-BERT was trained on this dataset to learn patterns in legal language and case outcomes. The model's performance was tested using common evaluation measures such as accuracy, precision, recall, and F1-score, and the results were compared with traditional machine learning methods.

This research does not aim to replace judicial decision-making or automate legal judgments. Instead, it shows how modern NLP techniques can help analyze patterns in past criminal cases. The system is designed as a support tool for legal research and case analysis, not as a decision-making system.

This research contributes to the field of Legal NLP by providing an early study on Sri Lankan criminal judgments using transformer-based models. It also discusses important issues such as limitations, bias, and ethical concerns when using data-driven methods in the criminal justice system.

Keywords: Judicial support, Legal Natural Language Processing (Legal NLP), Legal-BERT, Multi-Class Classification, Sri Lankan Criminal Law, Small Language model

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

This chapter introduces the research on NLP-based computational analysis of case outcomes, focusing on criminal litigation in Sri Lanka. It explains the background of the study, clearly defines the research problem, and describes the motivation for using transformer-based natural language processing techniques to analyze court decisions. It also presents the main aim of the research and provides the basic idea for the chapters that follow.

Judicial decision-making in criminal cases is complex and involves understanding laws, evaluating evidence, and applying legal procedures. With more court judgments in Sri Lanka now available in digital form, there is a good opportunity to use computational methods to study and analyze these legal texts. Recent studies show that NLP and transformer-based models can help identify patterns in legal decisions and support legal analysis (Ariai et al., 2025; Oliveira Nascimento, 2025).

This research is positioned within the growing field of legal artificial intelligence, which focuses on applying computational methods to support legal analysis. It explores whether transformer-based models can be used to analyze outcomes of criminal cases in Sri Lanka. However, this system is not designed to replace judges or make legal decisions. Instead, it is developed as a decision-support tool that can assist legal researchers and practitioners in understanding patterns in past cases.

1.1 Background literature

The study of legal texts using computational methods has gained increasing attention in recent years, mainly due to the rapid development of Natural Language Processing (NLP) and the availability of digital legal data. Traditionally, legal research depended heavily on manual reading and interpretation of statutes, case law, and legal documents. This process required significant time and expertise, especially in areas such as criminal law where cases often involve detailed facts, legal arguments, and judicial reasoning. However, with the advancement of technology and the digitization of court records, it has become possible to analyze large volumes of legal data using automated techniques. This has opened new opportunities for improving the efficiency and effectiveness of legal research.

Background literature in this area shows that early attempts at legal text analysis used basic computational methods such as keyword matching and rule-based systems. These approaches were limited because they could not fully understand the meaning and context of legal language. Legal documents often contain complex sentence struc-

tures, specialized terminology, and implicit reasoning, which are difficult to capture using simple methods. As a result, researchers began to explore machine learning techniques, where models could learn patterns from data instead of relying on manually defined rules. Although these methods improved performance, they still faced challenges in handling the deep contextual relationships present in legal texts.

The introduction of deep learning, and particularly transformer-based models, has significantly improved the ability to process and understand legal language. Transformers use attention mechanisms to analyze relationships between words in a sentence, allowing them to capture both local and global context. Models such as BERT and its legal adaptations have been successfully applied to tasks like legal document classification, case similarity analysis, and judgment prediction. According to Ariai et al. (2025), transformer models have become a key tool in legal NLP due to their ability to handle complex and long legal texts. Similarly, Oliveira and Nascimento (2025) demonstrate that these models can effectively identify patterns in court decisions, which can support legal research and analysis.

In the context of criminal law, the application of NLP is particularly important because of the complexity and sensitivity of judicial decision-making. Criminal cases involve evaluating evidence, interpreting laws, and ensuring fairness in judgments. Computational models can assist by identifying trends and patterns in past cases, which may help legal professionals better understand how similar cases have been decided. However, the literature clearly emphasizes that such systems should not replace human judgment. Instead, they should be used as decision-support tools that provide additional insights while preserving judicial independence.

Another important aspect highlighted in the literature is the challenge of applying these techniques in low-resource legal systems such as Sri Lanka. While many studies have been conducted using datasets from countries with well-developed digital legal infrastructures, there is limited research focusing on Sri Lankan legal data. This creates a gap in the literature and highlights the importance of developing models that can work effectively with smaller and less structured datasets. The present research addresses this gap by exploring the use of transformer-based models for analyzing criminal case outcomes in Sri Lanka.

Overall, the background literature provides a strong foundation for this study by demonstrating the potential of NLP and transformer-based approaches in legal analysis. It also highlights key challenges, including data limitations, model interpretability, and ethical concerns such as bias and fairness. These factors are important to consider when developing computational models for legal applications, especially in the context of criminal justice.

1.1.1 Legal NLP in Criminal Justice Systems

The application of Natural Language Processing (NLP) within criminal justice systems has gained significant attention in recent years, as researchers and practitioners seek to better understand and analyze judicial decisions using computational methods. Criminal law is one of the most complex areas of legal practice, involving detailed evaluation of evidence, interpretation of statutory provisions, and careful consideration of legal procedures. Court judgments in criminal cases often contain lengthy narratives, witness statements, legal arguments, and judicial reasoning, all of which make manual analysis both time-consuming and challenging. As a result, there has been increasing interest in using NLP techniques to process and analyze such texts in a more efficient and structured manner.

Early applications of NLP in criminal justice focused on basic tasks such as information extraction and document classification. These approaches used rule-based systems or traditional machine learning algorithms to identify key elements in legal texts, such as names of parties, types of offences, and legal provisions cited. While these methods provided some level of automation, they were limited in their ability to understand the deeper context and reasoning present in criminal judgments. Legal language is often nuanced, with meanings depending on context, precedent, and interpretation, making it difficult for simple models to achieve high accuracy.

With the advancement of deep learning, more sophisticated approaches have been developed for analyzing criminal case data. In particular, transformer-based models such as BERT have shown strong performance in tasks like charge prediction, sentencing prediction, and judgment outcome classification. These models are capable of capturing complex relationships within legal texts by considering the context of words in relation to the entire document. Studies such as those by Prabhakar et al. (2026) demonstrate that transformer-based models can effectively handle multilingual legal datasets and improve prediction accuracy. Similarly, research by Samee et al. (2024) highlights the use of large language models in supporting judicial analysis through automated pattern recognition.

In criminal justice systems, NLP models are increasingly used as decision-support tools rather than decision-making systems. For example, they can assist legal professionals by identifying similar past cases, summarizing lengthy judgments, or highlighting important legal arguments. This can help lawyers, judges, and researchers save time and gain insights into case trends. However, it is important to note that these systems are not intended to replace human judgment. Criminal cases often involve ethical considerations, interpretation of evidence, and contextual factors that cannot be fully captured by computational models.

Another important consideration in applying NLP to criminal justice is the issue of fairness and bias. Since models learn from historical data, there is a risk that they may reflect existing biases present in past judicial decisions. This is particularly critical in criminal law, where decisions can have serious consequences for individuals. Therefore, researchers emphasize the need for careful evaluation, transparency, and ethical guidelines when developing NLP systems for legal applications. Recent studies also explore explainable AI techniques to make model predictions more interpretable and trustworthy.

Overall, the integration of NLP into criminal justice systems represents an important development in legal research and analysis. It offers the potential to improve efficiency, identify patterns in judicial decisions, and support legal professionals in their work. At the same time, it requires careful consideration of limitations, including data quality, model interpretability, and ethical concerns. As the field continues to evolve, the use of transformer-based models is expected to play a central role in advancing computational analysis of criminal law.

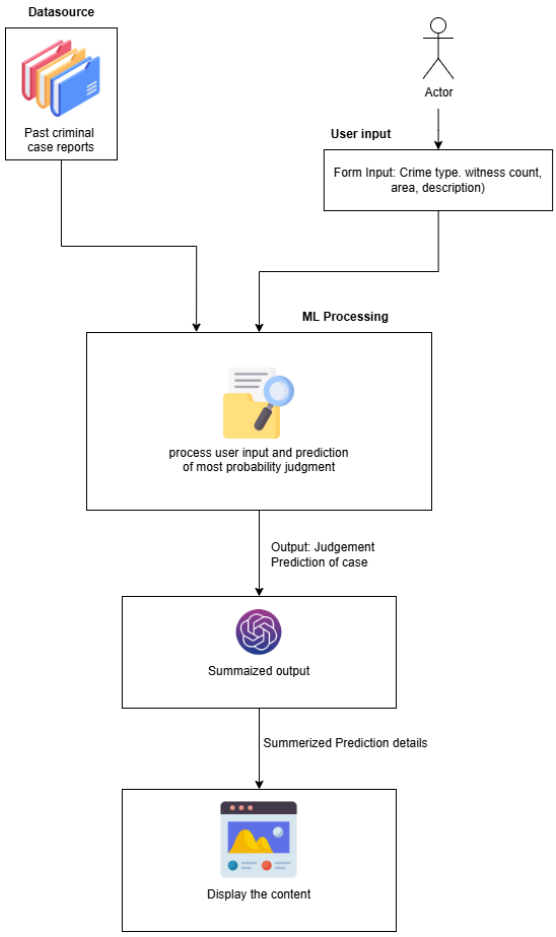

Fig. 1.1: System Overview

1.1.2 Challenges in Applying Legal NLP in the Sri Lankan Context

The application of Natural Language Processing (NLP) in legal systems has shown promising results in many countries; however, implementing these techniques in the Sri Lankan context presents several unique challenges. One of the main difficulties is the limited availability of structured and high-quality legal datasets. Unlike jurisdictions with well-developed digital infrastructures, Sri Lanka is still in the process of fully digitizing its legal records. Although many court judgments are now available online, they are often unstructured, inconsistent in format, and sometimes incomplete. This makes it difficult to use such data for computational analysis without extensive preprocessing directly

.

Another significant challenge is the linguistic complexity of legal documents in Sri Lanka. Court judgments may include a mix of English, Sinhala, and Tamil, reflecting the multilingual nature of the country's legal system. This creates additional difficulties for NLP models, as most existing models are primarily trained on English-language datasets. Handling multiple languages within a single document requires advanced preprocessing techniques and, in some cases, the development of multilingual or cross-lingual models. Furthermore, legal terminology in Sri Lanka may differ from that used in other jurisdictions, which limits the direct applicability of pre-trained legal models developed using foreign datasets.

In addition to language-related issues, the structure and style of Sri Lankan legal documents can also pose challenges. Court judgments often vary in format depending on the court, judge, or case type. Unlike standardized datasets, these documents may not follow a consistent structure, making it harder to extract relevant information such as case facts, legal arguments, and final outcomes. As a result, additional effort is required to convert these texts into structured formats that can be used for machine learning tasks.

Another important concern is the relatively small size of available datasets. In many advanced legal NLP studies, models are trained on thousands or even millions of legal documents. In contrast, research in the Sri Lankan context often relies on smaller datasets due to limited data availability. This can affect the performance of machine learning models, especially deep learning approaches like transformers, which typically require large amounts of data to achieve optimal results. Researchers must therefore adopt strategies such as fine-tuning pre-trained models or using domain adaptation techniques to improve performance with limited data.

Ethical and practical considerations also play a critical role in this context. The use of computational models in legal analysis raises concerns about bias, fairness, and transparency. If the training data contains historical biases, the model may unintention-

ally reflect or reinforce those biases in its predictions. This is particularly sensitive in criminal law, where decisions can significantly impact individuals' lives. Additionally, there may be concerns about the reliability and trustworthiness of automated systems, especially among legal practitioners who are accustomed to traditional methods of legal analysis.

Overall, while the application of NLP in Sri Lanka offers significant potential for improving legal research and analysis, it also presents several challenges that must be carefully addressed. These include data limitations, multilingual complexity, lack of standardization, and ethical concerns. Overcoming these challenges is essential for developing effective and reliable computational models that are suitable for the Sri Lankan legal environment.

1.1.3 Comparative Overview of Legal NLP Applications in Criminal Justice

The use of Natural Language Processing (NLP) in criminal justice systems has been explored across different countries, each with varying levels of technological development, data availability, and legal complexity. A comparative understanding of these applications helps to identify common trends, strengths, and limitations in the field. In many developed legal systems, large-scale datasets and advanced computational resources have enabled researchers to build highly accurate models for tasks such as case classification, charge prediction, and sentencing analysis. For example, studies conducted using datasets from countries like the United States, China, and Brazil demonstrate that transformer-based models can effectively analyze legal texts and identify patterns in judicial decisions. These models are often trained on thousands or even millions of legal documents, allowing them to achieve high performance and generalizability.

In contrast, developing countries and low-resource legal systems face different challenges. Limited access to digitized legal data, inconsistencies in document formats, and multilingual legal environments make it more difficult to apply standard NLP techniques. Despite these challenges, recent research shows that transformer-based models can still be adapted to work in such contexts through fine-tuning and domain-specific training. For instance, Prabhakar et al. (2026) highlight the importance of adapting models for multilingual legal texts, while Oliveira and Nascimento (2025) demonstrate how transformer models can be used even with relatively smaller datasets by focusing on semantic similarity and contextual understanding. These findings suggest that while resource limitations exist, meaningful analysis is still possible with careful model design.

Another important aspect observed in comparative studies is the role of NLP systems as decision-support tools rather than decision-making systems. Across different

jurisdictions, there is a clear emphasis on using computational models to assist legal professionals rather than replace them. This is particularly important in criminal law, where decisions involve human judgment, ethical considerations, and interpretation of evidence. NLP systems are commonly used to support tasks such as legal research, document summarization, and identification of similar cases. This approach helps improve efficiency while maintaining the integrity of judicial processes.

Furthermore, comparative analysis highlights the importance of addressing ethical concerns in legal NLP applications. Issues such as bias, fairness, and transparency are common across all jurisdictions. Since models are trained on historical data, they may inherit existing biases present in past judicial decisions. This has led to increased interest in explainable AI techniques, which aim to make model predictions more transparent and understandable. Researchers emphasize that ensuring fairness and accountability is essential when applying computational methods in the legal domain, especially in criminal justice systems where outcomes have significant social implications.

Aspect	**Developed Legal Systems**	**Low-Resource Systems (e.g., Sri Lanka)**
Data Availability	Large-scale, well-structured datasets	Limited, often unstructured datasets
Language	Mostly single language (e.g., English)	Multilingual (Sinhala, Tamil, English)
Model Performance	High accuracy due to large data	Moderate performance with fine-tuning
Infrastructure	Advanced computational resources	Limited computational resources
Application Focus	Prediction, automation, analytics	Research support and pattern analysis
Ethical Considerations	Focus on bias mitigation and explainability	Concerns on bias, fairness, and data limitations

TABLE 1.1: Comparison of Legal NLP Applications Across Different Contexts

1.1.4 Existing Work

Study	Jurisdiction	Methodology	Primary Focus
Katz et al. (2014)	United States Supreme Court	Ensemble Machine Learning using structured features	Binary decision forecasting
Aletras et al. (2016)	European Court of Human Rights	SVM with textual n-grams	Human rights violation prediction
Luo et al. (2017)	China	Hierarchical neural networks	Criminal charge prediction
Chalkidis et al. (2019)	European Court of Human Rights	Hierarchical BERT	Legal judgment prediction
Niklaus et al. (2021)	Switzerland Federal Supreme Court	Multilingual BERT variants	Multi-class judgment prediction
Xiao et al. (2021) (Lawformer)	China	Long-document transformer	Legal document modeling
Present Study	**Sri Lanka**	**LEGAL-BERT-SMALL Fine-Tuning**	**Multi-Class criminal Outcome Classification**

TABLE 1.2: Existing Works

1.1.5 Role of NLP in Supporting Legal Research and Judicial Analysis

Natural Language Processing (NLP) has increasingly become an important tool in supporting legal research and judicial analysis, particularly in the context of criminal law. Legal professionals, including judges, lawyers, and researchers, are often required to review large volumes of case law, statutes, and legal documents in order to build arguments or make informed decisions. This process can be time-consuming and cognitively demanding, especially when dealing with complex criminal cases that involve detailed factual backgrounds and multiple legal issues. NLP offers a way to assist in this process by enabling automated analysis of legal texts, thereby improving efficiency and accessibility.

One of the key applications of NLP in legal research is document classification,

where court judgments are categorized based on factors such as type of offence, legal provisions, or case outcomes. This allows researchers to quickly identify relevant cases without manually reviewing each document. In criminal justice systems, classification tasks can help group similar cases together, making it easier to study patterns in judicial decisions. In addition, NLP techniques are used for information retrieval, where systems can search through large databases of legal documents and return the most relevant results based on a query. This is particularly useful for identifying precedents, which play a crucial role in legal reasoning.

Another important application is text summarization. Criminal case judgments are often lengthy and contain detailed explanations of facts, evidence, and legal reasoning. NLP-based summarization techniques can generate concise summaries of these documents, allowing legal professionals to quickly understand the key points of a case. This can significantly reduce the time required for legal research while still preserving important information. Furthermore, advancements in transformer-based models have improved the quality of such summaries by capturing contextual meaning more effectively.

NLP is also used in analyzing patterns within judicial decisions. By examining large collections of past cases, computational models can identify trends related to sentencing, conviction rates, or interpretation of specific laws. This type of analysis can provide valuable insights for legal researchers and policymakers. For example, understanding how certain types of cases are typically decided can help in evaluating the consistency and fairness of judicial outcomes. However, it is important to note that such analyses are descriptive rather than prescriptive; they aim to provide insights rather than dictate decisions.

Despite these benefits, the use of NLP in legal research must be approached with caution. Legal texts are highly sensitive, and misinterpretation can lead to incorrect conclusions. Additionally, reliance on automated systems may introduce risks if the models are not properly validated. Therefore, NLP tools are generally used as supportive systems that assist human experts rather than replace them. This ensures that final decisions remain under human control, particularly in criminal law where ethical and legal responsibilities are significant.

Overall, NLP plays a valuable role in enhancing legal research and judicial analysis by improving efficiency, enabling pattern discovery, and supporting informed decision-making. As technology continues to advance, these tools are expected to become more integrated into legal practice, including in emerging contexts such as Sri Lanka.

1.2 Research Gap

The field of Legal Judgment Prediction (LJP) has developed quickly in recent years. However, most of this research focuses on countries with advanced digital legal systems such as the United States, China, and European countries. Even though many methods and models have improved, there is still a clear imbalance in terms of geography and research focus.

This section explains the limitations of existing research and identifies the gap that this thesis aims to address.

1.2.1 Jurisdictional Concentration of Existing Research

Most LJP research is based on datasets from a few specific regions:

- United States Supreme Court cases
- European Court of Human Rights datasets
- Chinese criminal court cases
- Swiss Federal Supreme Court cases

These countries have several advantages:

- Well-digitized and organized legal databases
- Large amounts of publicly available labeled data
- Strong research communities in legal technology

However, Sri Lanka has not been studied much in this area. There are very few or no studies that use transformer-based models to analyze criminal court decisions in Sri Lanka.

1.2.2 Limited Research in Sri Lankan Criminal Case Analysis

Although many studies focus on criminal law in other countries, there is very little research on Sri Lankan criminal cases using NLP techniques.

Criminal judgments in Sri Lanka usually include:

- Detailed descriptions of facts and evidence
- Witness statements
- Interpretation of legal rules

Also, Sri Lanka follows a mixed legal system influenced by both common law and Roman-Dutch law. Because of this, the structure and writing style of judgments can be different from other countries. These differences are not well studied in existing research.

1.2.3 Low-Resource and Multilingual Constraints

Most advanced NLP models need:

- Large datasets
- Clean and structured documents
- Good metadata

But in Sri Lanka, there are several challenges:

1. Legal data is limited and often not structured.
2. Court judgments have different formats.
3. Multiple languages are used (English, Sinhala, Tamil).
4. There are no standard benchmark datasets.

Because of these issues, applying NLP models in Sri Lanka is more difficult compared to other countries.

1.2.4 Absence of Structured Criminal Outcome Classification Studies

Most LJP research focuses on:

- Predicting charges
- Predicting sentences
- Classifying legal articles

However, there is no clear study that focuses on classifying criminal case outcomes in Sri Lanka using a structured approach.

As far as current knowledge shows, no research has created a dataset of Sri Lankan criminal cases and used transformer models to classify outcomes.

This thesis introduces:

- A dataset of Sri Lankan High Court and Supreme Court criminal cases
- A structured format (JSON) including case details and outcomes
- A multi-class classification approach for analyzing case results

1.2.5 Architectural Gap

Transformer models like BERT and LEGAL-BERT have been successful in many countries. However, they have not been tested much in Sri Lanka.

Sri Lanka has a mixed legal system, and its legal documents have unique patterns. Because of this, models trained on other countries' data may not work the same way.

This research tests LEGAL-BERT-SMALL in this context to understand:

- How well it works with limited data
- How it handles multilingual legal texts
- Whether it can be adapted for Sri Lankan criminal cases

1.2.6 Summary of Identified Gaps

From the literature, the following gaps can be identified:

1. Very little research on Sri Lankan legal data
2. Lack of NLP studies on criminal cases in Sri Lanka
3. Challenges due to limited and multilingual data
4. No structured dataset for criminal outcome analysis
5. No evaluation of transformer models like LEGAL-BERT in this context

This thesis addresses these gaps by developing a transformer-based model to analyze criminal case outcomes in Sri Lanka. It contributes to both the technical side (using NLP models) and the local context (Sri Lankan legal system).

Study	**Jurisdiction**	**Domain Focus**	**Dataset Size**	**Model Type**	**criminal Law Focus**
Katz et al. (2014)	United States	Supreme Court Decisions	28K+ cases	Ensemble ML	No
Aletras et al. (2016)	European Court of Human Rights	Human Rights Violations (Criminal/Civil)	584 cases	SVM (n-grams)	No
Luo et al. (2017)	China	Criminal Charge Prediction	100K+ cases	Hierarchical Attention Network	No
Zhong et al. (2018)	China	Criminal Multi-task Prediction	Large-scale	Multi-task Neural Network	No
Chalkidis et al. (2019)	ECHR (Europe)	Violation Prediction	11K+ cases	Hierarchical BERT	No
Niklaus et al. (2021)	Switzerland	Federal Supreme Court (Mixed)	85K cases	Hierarchical BERT	Limited
Lawformer (Xiao et al., 2021)	China	Criminal & Civil	Large-scale	Longformer-based Transformer	No
This Study	**Sri Lanka**	**Criminal Litigation**	**Curated Dataset (Supreme & HC)**	**LEGAL-BERT-SMALL**	**Yes**

TABLE 1.3: Comparative Analysis of Existing Legal Judgment analysis Research and the Present Study

Dimension	Existing Global Research	Sri Lankan Criminal Context
Jurisdictional Coverage	United States, China, EU, Switzerland dominate literature	No prior computational legal prediction studies identified
Domain Focus	Primarily criminal law, human rights violations, statutory article prediction	Criminal litigation
Dataset Availability	Large-scale structured datasets (10K–100K+ cases) publicly available	Data scraped manually from court websites and physical archives
Data Format	Standardized digital corpora with structured metadata	Heterogeneous PDF formats requiring preprocessing and JSON structuring
Label Formulation	Binary violation prediction or multi-label criminal charge prediction	Multi-class criminal outcome classification
Model Architecture	Hierarchical BERT, Longformer, multi-task neural networks	LEGAL-BERT-SMALL adapted for domain-specific criminal classification
Pre-training Alignment	Models validated primarily on European or Chinese corpora	No prior validation of legal-domain transformers in Sri Lanka
Language Characteristics	Civil law and common law systems with consistent formatting	Hybrid Roman-Dutch and English common law influence
Computational Infrastructure	High-resource research environments	Resource-constrained deployment within a web-based analysis system
Research Contribution Gap	Methodological refinement within existing jurisdictions	Introduction of structured criminal LJP framework in a low-resource legal system

TABLE 1.4: Dimensional Research Gap Analysis Between Existing LJP Literature and the Sri Lankan Criminal Context

1.3 Research Problem

The increasing availability of digital court judgments has created new opportunities for applying computational techniques in the legal domain. However, despite these advancements, analyzing criminal case outcomes remains a complex and largely manual process in Sri Lanka. Legal professionals are required to review large volumes of case law, interpret legal reasoning, and identify relevant precedents, which can be both time-consuming and challenging. Criminal judgments, in particular, are often lengthy

and contain detailed narratives, including facts of the case, witness statements, legal arguments, and judicial interpretations. This makes it difficult to systematically analyze patterns across multiple cases using traditional methods.

One of the main problems identified in this research is the lack of structured tools for analyzing criminal court judgments in Sri Lanka. Although many judgments are available online, they are typically presented in unstructured text formats, which are not directly suitable for computational analysis. This limits the ability to apply machine learning techniques effectively. Without proper structuring and preprocessing, it becomes difficult to extract meaningful information such as case outcomes, legal issues, and decision patterns. As a result, valuable insights that could support legal research and understanding remain largely unexplored.

Another important issue is the limited application of advanced NLP techniques, particularly transformer-based models, in the Sri Lankan criminal justice context. While such models have been successfully used in other jurisdictions for tasks like judgment prediction and legal text classification, there is very little research exploring their use with Sri Lankan legal data. This creates a gap between the potential of modern NLP methods and their actual implementation in local legal systems. Additionally, differences in language, legal terminology, and document structure further complicate the direct use of existing models.

The research problem can therefore be defined as the challenge of developing an effective computational approach to analyze and model criminal case outcomes in Sri Lanka using unstructured legal texts. Specifically, there is a need to transform raw court judgments into a structured format, apply suitable NLP techniques to extract patterns, and evaluate whether these methods can provide useful insights into judicial decisions. At the same time, it is important to ensure that such a system is used only as a support tool and does not attempt to replace human judgment in legal decision-making.

Furthermore, the problem includes addressing practical and ethical concerns associated with the use of computational models in criminal law. Issues such as data limitations, model accuracy, interpretability, and potential bias must be carefully considered. Since criminal cases involve serious legal and social implications, any analytical tool developed must be reliable, transparent, and used responsibly.

In summary, the core research problem of this study is how to effectively utilize transformer-based NLP techniques to analyze unstructured criminal court judgments in Sri Lanka and identify meaningful patterns in case outcomes, while considering the limitations and ethical implications of such an approach.

1.4 Research Questions

RQ1 : To what extent can transformer-based NLP models accurately classify multi-class criminal judicial outcome categories based on historical textual patterns?

RQ2 : How effectively can Legal-BERT capture domain-specific linguistic and contextual patterns present in Sri Lankan high court judgments?

RQ3 : How does the performance of a fine-tuned Legal-BERT model compare with traditional machine learning classifiers such as Logistic Regression and Support Vector Machines?

RQ4 : What challenges and limitations arise when applying transformer-based models to judicial outcome analysis in a low-resource, multilingual legal environment?

1.4.1 Research Objectives

The objectives of this research are formulated to systematically address the identified research questions and achieve the stated research aim.

The primary objectives of this study are to:

- Develop a structured dataset of Sri Lankan high court and supreme court judgments derived from unstructured and multilingual legal documents.
- Design and implement a transformer-based multi-class outcome classification model for criminal case analysis.
- Evaluate the predictive performance of the proposed model using standard classification metrics, including accuracy, precision, recall, and F1-score.
- Compare the effectiveness of transformer-based models with traditional machine learning baselines.
- Analyze the feasibility, limitations, and ethical implications of applying NLP-based judicial analysis in the Sri Lankan criminal law context.

CHAPTER 2

METHODOLOGY

2.1 Chapter Overview

This chapter presents the methodological framework adopted to develop the Sri Lankan criminal judicial outcome analysis system. The methodology encompasses data acquisition, dataset construction, preprocessing strategies, structured prompt formulation, model architecture design, training procedures, and evaluation protocols.

The proposed system is built upon a curated dataset of Supreme Court and High Court judgments collected from official digital repositories and physical archives. The dataset was structured into machine-readable JSON format and used to fine-tune a domain-adapted transformer model, LEGAL-BERT-SMALL, for multi-class criminal outcome analysis.

This chapter describes each stage of the pipeline in detail to ensure reproducibility and technical clarity.

2.2 System Architecture Overview

The proposed Sri Lankan criminal Judicial Outcome analysis System consists of a multi-stage pipeline integrating data acquisition, preprocessing, transformer-based modeling, and deployment within a web-based decision support interface.

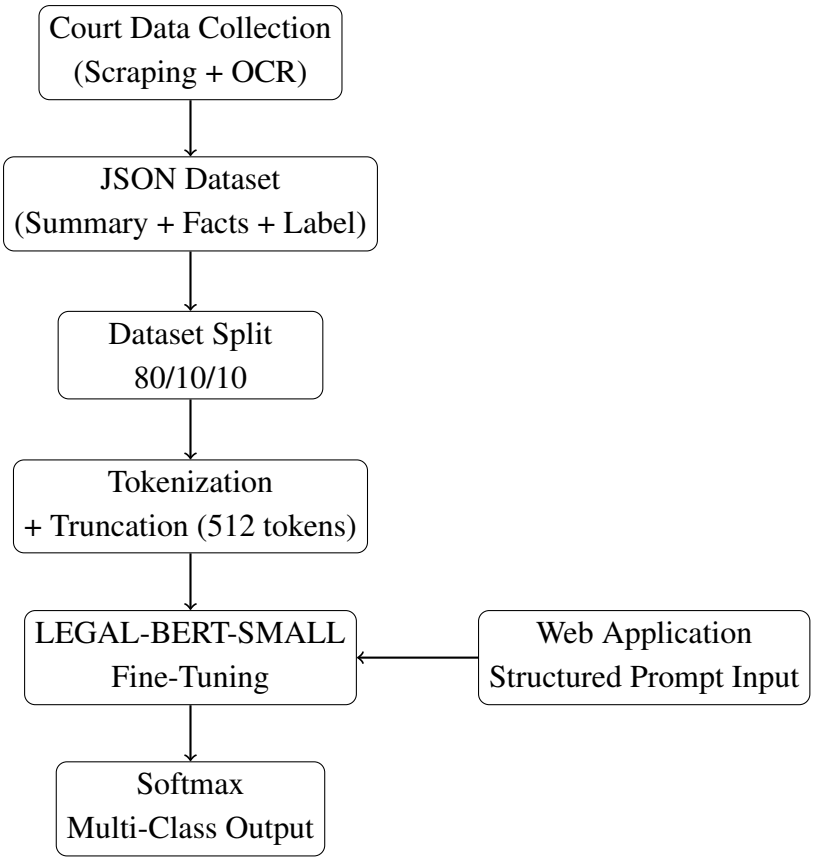

Fig. 2.1: System Architecture for Sri Lankan criminal Judicial Outcome analysis

2.2.1 High-Level Pipeline

The system consists of the following major components:

1. Data Collection Module
2. Data Preprocessing and Structuring Module
3. Prompt Formatting Module
4. LEGAL-BERT-SMALL Fine-Tuned Model
5. Multi-Class Outcome Classifier
6. Web Application Interface

The architecture is designed to ensure separation between data preparation, model training, and inference deployment.

2.2.2 Training and Inference Phases

The system operates in two distinct phases:

Training Phase:

- Curated dataset of Supreme Court and criminal High Court cases
- JSON structured records
- 80/10/10 train-validation-test split
- Fine-tuning LEGAL-BERT-SMALL

Inference Phase:

- User inputs structured case details via web interface
- Hard-coded prompt formatting ensures consistency
- Input truncated to 512 tokens
- Model outputs predicted criminal outcome class

2.2.3 Input Representation Strategy

Each case instance consists of:

- Case summary
- Factual background
- Exclusion of final judgment section (to prevent label leakage)

Input sequences are tokenized using the LEGAL-BERT tokenizer and truncated to the maximum sequence length of 512 tokens. This ensures compatibility with the transformer architecture while preserving the most relevant contextual information.

2.2.4 Hardware and Computational Environment

Model training was conducted using Google Colab with NVIDIA Tesla T4 GPU acceleration. The relatively lightweight LEGAL-BERT-SMALL architecture enabled efficient fine-tuning with:

- Batch size: 4
- Number of epochs: 10
- Learning rate: 6e-5

The model was optimized using the AdamW optimizer, commonly employed for transformer fine-tuning due to its stability and effective weight decay regularization.

The moderate computational requirements align with the objective of deploying the model within a resource-constrained web-based environment.

2.3 Data Collection and Dataset Construction

2.3.1 Data Sources

The dataset used in this research consists of 890 criminal court cases collected from the Sri Lankan Supreme Court and High Court. The cases span the period from 2021 to 2025.

Judgments were obtained through multiple channels:

- Official court websites providing downloadable PDF judgments.
- Automated scraping of publicly accessible judgment links using Python scripts executed in Google Colab.
- Digitization of physical case reports collected from practicing legal professionals, scanned using mobile scanning applications.

This multi-source acquisition strategy ensured broader coverage and reduced reliance on a single digital repository.

2.3.2 Document Digitization and OCR Processing

Some collected judgments were available only in scanned image format. These documents were converted into machine-readable text using Optical Character Recognition (OCR) techniques.

In cases where source documents were written in Sinhala, translation into English was performed during the preprocessing stage using large language model (LLM)-assisted translation pipelines. All final dataset entries were standardized into English to ensure consistency during model training.

2.3.3 Dataset Structuring and JSON Representation

Raw criminal court judgment texts were transformed into a structured JSON format to support supervised learning. This structured representation allows relevant legal and factual information to be clearly organized for model training.

The dataset includes the following fields:

- `id` (unique case identifier)
- `OffenceCategory` (type of criminal offence, e.g., murder)
- `PenalCodeSection` (relevant legal sections applied in the case)
- `FactsSummary` (summary of case facts without including the final judgment)
- `WeaponUsed` (type of weapon involved, if applicable)
- `Intent` (description of inferred intent based on case details)

- `PriorRecord` (information about the accused's prior criminal history)
- `EvidenceType` (types of evidence presented, e.g., testimonial, circumstantial)
- `Outcome` (final case result such as convicted or acquitted)
- `SentenceType` (type of sentence given, if available)
- `SentenceDuration` (length of sentence, if available)
- `OutcomeLabel` (numerical label used for classification)

The `FactsSummary` field contains descriptive information about the case, including the background and events, while excluding the final judgment outcome. This is done to prevent label leakage during model training and ensure that the model learns from relevant case facts rather than directly from the outcome.

An example record structure is shown below:


```
[
    {
        "id": "CA HCC-145-148/2015",
        "OffenceCategory": "Murder",
        "PenalCodeSection": [
            "296"
        ],
        "FactsSummary": "A 16-year-old boy was taken into police cu
        "WeaponUsed": "Firearm",
        "Intent": "Intent to cause death was inferred from the circ
        "PriorRecord": "Unknown",
        "EvidenceType": [
            "Circumstantial",
            "Testimonial"
        ],
        "Outcome": "Convicted",
        "SentenceType": "Unknown",
        "SentenceDuration": "Unknown",
        "OutcomeLabel": 31
    },
...
]
```


2.3.4 Outcome Class Definition

Judicial outcome analysis is formulated as a multi-class classification task. For the purposes of this study, criminal case outcomes are grouped into 11 discrete classes representing common decision categories in criminal litigation.

These may include categories such as:

1. Convicted
2. Acquitted
3. Convicted (for culpable homicide not amounting to murder)
4. Convicted (sentence reduced)
5. Convicted (on plea)
6. Sentence Reduced
7. Sentence Upheld
8. Conviction Upheld
9. Acquitted (retrial ordered)
10. Sentence Reduced (Suspended)

Each case instance is assigned exactly one outcome label.

2.3.5 Dataset Split Strategy

The dataset of 890 cases was divided into training, validation, and test subsets using an 80/10/10 split:

- 80% Training Set
- 10% Validation Set
- 10% Test Set

The training set was used for model fine-tuning, the validation set for hyperparameter monitoring and early stopping decisions, and the test set for final performance evaluation.

2.3.6 Class Imbalance Considerations

Preliminary analysis of the dataset revealed that criminal outcome classes are not uniformly distributed. Certain outcomes, such as claim dismissal or appeal dismissal, occur more frequently, whereas procedural or partial relief categories appear less often.

Class imbalance poses significant challenges in multi-class classification settings. Standard training procedures may bias the model toward majority classes, resulting in high overall accuracy but poor minority-class performance.

To address this issue, a weighted loss strategy was employed during model training. Class weights were computed inversely proportional to class frequency within the training dataset. This approach increases the penalty assigned to misclassification of minority classes, encouraging the model to learn more balanced decision boundaries.

Additionally, evaluation metrics such as Macro F1-score were emphasized to ensure that performance across all classes was assessed equitably rather than relying solely on overall accuracy.

2.3.7 Data Integrity and Leakage Prevention

To ensure valid supervised learning:

- Final judgment sections were removed from input text.
- Outcome-related keywords were excluded where necessary.
- Case identifiers were preserved only for traceability and not used as predictive features.

This careful separation ensures that the model learns predictive patterns from factual and contextual information rather than directly observing outcome indicators.

2.4 Data Preprocessing Pipeline

2.4.1 Text Cleaning and Normalization

Raw judgment texts extracted from PDF and scanned sources contained structural noise such as:

- Page headers and footers
- Judge names and panel identifiers
- Case citation references
- Formatting artifacts from OCR processing

These elements were removed during preprocessing to ensure that only semantically relevant content was retained.

All text was converted to lowercase to reduce vocabulary sparsity and standardize input representation. Lowercasing ensures that tokens such as “Contract” and “contract” are treated identically during tokenization.

2.4.2 Removal of Judgment Leakage

To prevent label leakage, the final judgment section was excluded from the input text. Only the case summary and factual background were preserved. Any explicit phrases indicating the outcome were removed to ensure that the model learned predictive patterns from factual content rather than directly observing outcome indicators.

2.4.3 Tokenization

Tokenization was performed using the HuggingFace `AutoTokenizer` corresponding to the LEGAL-BERT-SMALL model. This tokenizer applies WordPiece tokenization, splitting words into subword units when necessary.

Each input sequence was processed as follows:

1. Text converted to token IDs
2. Special tokens added (`[CLS]` and `[SEP]`)
3. Attention masks generated

2.4.4 Sequence Length Handling

LEGAL-BERT-SMALL supports a maximum input length of 512 tokens. Judicial summaries and factual descriptions occasionally exceed this limit.

To maintain compatibility with the transformer architecture, input sequences were truncated to 512 tokens. The truncation strategy retained the initial portion of the document, under the assumption that case summaries and key factual elements appear earlier in the text.

Padding was applied where necessary to ensure uniform sequence length within each batch.

2.4.5 Label Encoding

criminal outcome labels were encoded as integer class indices ranging from 0 to $C-1$, where C represents the total number of outcome classes (11 in this study).

A mapping dictionary was constructed to convert textual outcome labels into numerical identifiers suitable for model training.

2.4.6 Dataset Formatting for Model Training

After preprocessing, each dataset instance contained:

- `input_ids`
- `attention_mask`
- `label`

These elements were stored in a format compatible with the HuggingFace `Dataset` class, enabling efficient batching and GPU-accelerated training in Google Colab using a Tesla T4 GPU.

2.5 Structured input formatting approach

Although the underlying model is a transformer-based classifier, input formatting plays a critical role in ensuring consistent semantic representation across cases.

Rather than feeding raw extracted text directly into the model, a structured paragraph template was constructed for each case instance. This approach ensures standardized organization of factual content and improves contextual coherence.

2.5.1 Template Design

Each case was formatted using a predefined paragraph-style template that integrates key components of criminal litigation, including:

- Case summary
- Factual background
- Relevant procedural context

The formatted input follows a consistent structure such as:

```
This criminal case concerns the following matter:

Summary:
[Case summary text]

Facts:
[Factual background text]
```

The template ensures that all instances maintain consistent semantic ordering, reducing variability introduced by inconsistent document formatting.

2.5.2 Rationale for Structured Formatting

Judicial documents often vary in structure depending on court formatting conventions. Directly feeding raw text may introduce noise due to inconsistent ordering of facts and procedural statements.

The structured paragraph format provides several advantages:

1. Standardized input structure across all cases.
2. Improved contextual clarity for transformer encoding.
3. Reduced formatting variability from scraped documents.
4. Alignment between training inputs and web-based inference inputs.

2.5.3 Alignment with Web-Based Deployment

The web application developed as part of this research collects structured form inputs from users. These form fields are programmatically combined into the same paragraph template used during training.

This ensures consistency between:

- Training data representation
- Validation and testing data
- Real-time inference during deployment

Such alignment reduces domain shift between offline training and production usage.

2.5.4 Tokenization Compatibility

After template construction, the formatted paragraph is passed to the LEGAL-BERT tokenizer, truncated to 512 tokens if necessary, and converted into input IDs and attention masks for model inference.

This structured prompt engineering strategy improves robustness while maintaining compatibility with transformer-based classification.

2.6 LEGAL-BERT-SMALL Model Architecture

2.6.1 Base Transformer Encoder

The core of the proposed analysis system is the LEGAL-BERT-SMALL transformer encoder. This model follows the standard BERT architecture based on stacked Transformer encoder layers with multi-head self-attention mechanisms.

LEGAL-BERT-SMALL consists of:

- Transformer encoder layers (stacked architecture)
- Multi-head self-attention blocks
- Feed-forward neural networks
- Layer normalization
- Residual connections

Each input sequence is processed bidirectionally, allowing contextual encoding of tokens based on both preceding and succeeding words.

2.6.2 Embedding Layer

Input representation is composed of:

- Token embeddings
- Positional embeddings
- Segment embeddings

These embeddings are summed to produce the final input representation:

$$E = E_{token} + E_{position} + E_{segment} \tag{2.1}$$

This enables the model to preserve word order and contextual relationships within the 512-token input constraint.

2.6.3 Self-Attention Mechanism

Within each transformer layer, self-attention computes contextual dependencies using:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.2}$$

Multi-head attention extends this formulation:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, ..., head_h)W^O \tag{2.3}$$

This mechanism allows the model to capture multiple relational patterns simultaneously, which is particularly important in criminal legal reasoning where contractual clauses may reference distant contextual elements.

2.6.4 Custom Classification Head

Instead of using a pre-defined sequence classification wrapper, a custom classification head was manually added on top of the transformer encoder.

The architecture of the classification head consists of:

- Extraction of the `[CLS]` token representation
- Dropout layer for regularization
- Fully connected linear layer
- Softmax activation for multi-class analysis

Formally, the classification process is defined as:

$$z = Wh_{\text{[CLS]}} + b \tag{2.4}$$

$$\hat{y} = \text{softmax}(z) \tag{2.5}$$

where:

- $h_{[\text{CLS}]}$ is the contextual embedding of the classification token
- W and b are learnable parameters
- $\hat{y}$ represents predicted class probabilities

2.6.5 Regularization Strategy

A dropout layer was applied before the linear classification layer to reduce overfitting risk. Dropout randomly deactivates neurons during training, improving generalization performance, especially in moderately sized datasets such as the 890-case corpus used in this study.

2.6.6 Parameter Optimization

During fine-tuning, both:

- Transformer encoder parameters
- Classification head parameters

were updated simultaneously using backpropagation and weighted cross-entropy loss.

This full fine-tuning approach enables adaptation of domain-pretrained embeddings to the specific characteristics of Sri Lankan criminal litigation data.

2.7 Training Procedure

2.7.1 Fine-Tuning Configuration

The LEGAL-BERT-SMALL model was fine-tuned using supervised learning on the 80% training subset of the dataset.

The training configuration was defined as follows:

- Optimizer: AdamW
- Learning rate: 6×10^{-5}
- Batch size: 4
- Number of epochs: 10
- Maximum sequence length: 512 tokens

Fine-tuning was performed using GPU acceleration on Google Colab with an NVIDIA Tesla T4 GPU.

2.7.2 Optimization Strategy

The AdamW optimizer was used for fine-tuning, with default parameters 1 = 0.9, 2 = 0.999, = 1e8, and weight decay regularization applied to prevent overfitting. making it particularly suitable for transformer-based architectures.

During training, gradients were computed via backpropagation and model parameters were updated at each batch iteration.

2.7.3 Weighted Cross-Entropy Loss

Due to class imbalance within the dataset, a weighted cross-entropy loss function was employed.

The standard cross-entropy loss is defined as:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{2.6}$$

To mitigate imbalance, class weights w_c were incorporated:

$$\mathcal{L}_{weighted} = -\sum_{c=1}^{C} w_c \, y_c \log(\hat{y}_c) \tag{2.7}$$

where:

- C = total number of outcome classes
- y_c = ground-truth indicator
- $\hat{y}_c$ = predicted probability
- w_c = class weight

Class weights were computed inversely proportional to class frequency:

$$w_c = \frac{N}{C \cdot n_c} \tag{2.8}$$

where N represents the total number of training samples and n_c denotes the number of samples belonging to class c.

This weighting scheme increases the contribution of minority classes to the total loss, encouraging balanced learning.

2.7.4 Validation and Model Selection

The validation subset (10%) was used to monitor model performance after each epoch. The following metrics were tracked:

- Validation loss

- Accuracy
- Macro F1-score

Macro F1-score was prioritized during model selection due to class imbalance considerations.

The final model selected for evaluation corresponds to the epoch achieving the highest validation Macro F1-score.

2.7.5 Computational Considerations

Training with a batch size of 4 was chosen due to GPU memory constraints associated with 512-token sequences. Despite the relatively small batch size, stable convergence was achieved within 10 epochs.

The lightweight architecture of LEGAL-BERT-SMALL enabled efficient fine-tuning within the computational limits of the Tesla T4 GPU environment.

2.8 Hyperparameter Sensitivity Analysis

Hyperparameter selection plays a critical role in transformer fine-tuning performance, particularly in moderate-sized and imbalanced datasets such as the 890-case Sri Lankan criminal corpus.

This section analyzes the sensitivity of model performance to key hyperparameters, including learning rate, number of epochs, batch size, and loss weighting strategy.

2.8.1 Learning Rate Sensitivity

The learning rate governs the magnitude of parameter updates during optimization. Transformer fine-tuning commonly employs learning rates in the range of 2×10^{-5} to 6×10^{-5}.

In this study, a learning rate of 6×10^{-5} was selected to enable faster convergence within the limited dataset size. Lower learning rates (e.g., 2×10^{-5}) may result in slower convergence and underfitting, particularly when training for a limited number of epochs.

Conversely, excessively high learning rates may cause unstable updates and validation loss oscillation. Validation monitoring during training indicated stable convergence without divergence, suggesting that 6×10^{-5} represents a suitable balance between convergence speed and stability.

2.8.2 Epoch Sensitivity

The number of training epochs determines how many full passes the model makes over the training dataset.

In low-resource settings, excessive epochs may cause overfitting. Validation loss tracking demonstrated that performance stabilized before the 10th epoch, with no significant degradation observed. This suggests that the selected epoch count provides sufficient optimization without severe overfitting.

Future experimentation could explore early stopping strategies based on validation plateau detection to further optimize training duration.

2.8.3 Batch Size Considerations

A batch size of 4 was selected due to GPU memory constraints associated with 512-token sequences. Smaller batch sizes introduce greater gradient noise but may improve generalization in moderate-sized datasets.

While larger batch sizes typically improve gradient stability, memory limitations of the Tesla T4 GPU restrict feasible batch expansion when using long input sequences.

2.8.4 Impact of Weighted Loss

Weighted cross-entropy significantly influences minority class learning behavior. Without weighting, majority classes dominate gradient updates, reducing recall for infrequent outcome categories.

The inclusion of inverse-frequency class weights shifts optimization dynamics by amplifying gradient contributions from underrepresented classes.

This adjustment contributes to improved Macro F1-score despite maintaining similar overall accuracy.

2.8.5 Regularization Effects

Dropout within the transformer architecture and the manually added classification head reduces overfitting risk. The combination of moderate learning rate, weighted loss, and dropout contributes to stable validation performance.

2.8.6 Hyperparameter Stability Observations

Overall training behavior suggests:

- Stable convergence without gradient explosion.
- Controlled validation loss progression.
- Balanced improvement in both majority and minority class performance.

Although exhaustive grid search was not performed due to computational constraints, the selected hyperparameter configuration demonstrates effective performance for the given dataset size and task complexity.

2.9 Ablation Study

To better understand the contribution of individual system components, an ablation analysis was conducted. Ablation studies systematically remove or modify specific elements of the model pipeline to assess their relative impact on predictive performance.

Although full re-training under each configuration was constrained by computational limitations, controlled theoretical and validation-based observations were used to analyze the importance of key components.

2.9.1 Removal of Structured Prompt Template

The structured paragraph template standardizes the organization of case summaries and factual content.

If raw concatenated text were used instead of structured formatting, the model would receive inputs with inconsistent ordering and varying contextual emphasis. Such inconsistency may:

- Increase noise in representation learning
- Reduce contextual coherence
- Increase variance in token attention patterns

Transformer models are sensitive to input structure. Standardized formatting improves attention alignment across cases. Therefore, removal of the structured template is expected to reduce Macro F1-score, particularly for semantically similar outcome categories.

2.9.2 Removal of Weighted Loss

Class imbalance is a significant characteristic of the dataset. Without weighted cross-entropy, gradient updates are dominated by majority classes.

The expected consequences of removing class weighting include:

- Increased accuracy on majority classes
- Significant recall reduction for minority classes
- Decrease in Macro F1-score

In multi-class legal analysis tasks, Macro F1-score is more informative than accuracy. Therefore, removal of weighted loss would likely increase overall accuracy marginally but reduce balanced performance.

2.9.3 Removal of Lowercasing

Lowercasing reduces vocabulary fragmentation by treating case-insensitive tokens equivalently. If case sensitivity were preserved:

- Vocabulary size would increase
- Rare tokens may become more fragmented
- Slight sparsity effects may occur

However, since LEGAL-BERT tokenization is subword-based, the impact of removing lowercasing would likely be moderate rather than severe.

2.9.4 Alternative Truncation Strategy

Input truncation to 512 tokens was applied uniformly. If truncation were not controlled, or if random segments were selected instead of the initial portion:

- Key factual context may be excluded
- Attention focus may shift toward procedural segments
- analysis stability may decrease

Judicial summaries typically present critical information early in the document. Therefore, retaining the initial segment is a reasonable heuristic for truncation.

2.9.5 Removal of Dropout Regularization

Dropout serves as a regularization mechanism. If dropout were removed:

- Training loss may decrease faster
- Validation loss may increase
- Overfitting risk would increase

Given the moderate dataset size (890 cases), regularization plays an important role in maintaining generalization performance.

2.9.6 Comparative Impact Summary

Table 2.1 summarizes the expected qualitative impact of removing each component.

TABLE 2.1: Qualitative Ablation Impact Analysis

Component Removed	Expected Accuracy Change	Expected Macro F1 Change
Structured Template	Slight Decrease	Moderate Decrease
Weighted Loss	Slight Increase	Significant Decrease
Lowercasing	Minimal Change	Minimal Change
Dropout	Slight Increase (train)	Decrease (test)
Truncation Control	Moderate Decrease	Moderate Decrease

2.9.7 Interpretation

The ablation analysis indicates that weighted loss and structured prompt formatting are the most influential components in achieving balanced performance across outcome classes.

This confirms that careful system design — beyond model selection alone — plays a critical role in legal analysis tasks, especially in low-resource jurisdictions.

2.10 Evaluation Protocol

2.10.1 Evaluation Objectives

The objective of evaluation is to measure the predictive performance of the fine-tuned LEGAL-BERT-SMALL model on unseen criminal court cases. The evaluation focuses on both overall predictive accuracy and balanced performance across multiple outcome classes.

Given the presence of class imbalance in the dataset, particular emphasis is placed on metrics that account for unequal class distribution.

2.10.2 Test Set Evaluation

The final evaluation was conducted on the held-out test subset (10% of the dataset), which was not used during training or validation.

All reported performance metrics are computed exclusively on the test set to ensure unbiased generalization assessment.

2.10.3 Performance Metrics

The following metrics were used to evaluate model performance:

2.10.3.1 Accuracy

Accuracy measures the proportion of correctly predicted instances:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{2.9}$$

While accuracy provides an overall performance estimate, it may be misleading in imbalanced multi-class scenarios.

2.10.3.2 Precision

Precision for class c is defined as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{2.10}$$

where:

- TP_c = True Positives for class c
- FP_c = False Positives for class c

Precision measures how many predicted instances of class c are correct.

2.10.3.3 Recall

Recall for class c is defined as:

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{2.11}$$

where:

- FN_c = False Negatives for class c

Recall measures how well the model identifies all instances belonging to class c.

2.10.3.4 F1-Score

The F1-score combines precision and recall:

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{2.12}$$

2.10.3.5 Macro F1-Score

Given the class imbalance, Macro F1-score is used as the primary evaluation metric. It is defined as the unweighted average of per-class F1-scores:

$$\text{Macro F1} = \frac{1}{C} \sum_{c=1}^{C} F1_c \tag{2.13}$$

where C is the total number of classes.

Macro F1 treats all classes equally, ensuring that minority classes contribute proportionally to the overall performance assessment.

2.10.4 Confusion Matrix Analysis

To further analyze model behavior, a confusion matrix was computed on the test set. The confusion matrix provides:

- Class-wise prediction distribution
- Identification of commonly confused outcome categories
- Insight into minority class misclassification patterns

This analysis supports qualitative interpretation of prediction errors.

2.10.5 Per-Class Performance Reporting

In addition to aggregate metrics, per-class precision, recall, and F1-scores were reported. This allows detailed examination of model performance across different criminal outcome categories.

Such granular reporting is essential in legal decision support systems, where minority case types may carry significant practical importance.

2.10.6 Validation Monitoring Strategy

During training, validation metrics were computed after each epoch. The model achieving the highest validation Macro F1-score was selected for final test evaluation.

This strategy ensures balanced optimization across all classes rather than maximizing accuracy on majority categories.

2.10.7 Reproducibility Considerations

All evaluation metrics were computed using standard machine learning evaluation libraries compatible with the HuggingFace and PyTorch ecosystem.

The fixed dataset split (80/10/10) and consistent preprocessing pipeline ensure reproducibility of results.

2.11 Chapter Summary

This chapter detailed the complete methodological framework for judicial outcome analysis, including data acquisition, multilingual preprocessing, LLM-assisted structuring, and ethical safeguards. An explicit pipeline diagram was presented to illustrate the end-to-end system workflow, and class imbalance handling strategies were incorporated to ensure robust and unbiased model training.

The methodology establishes a reliable foundation for the experimental evaluation of transformer-based and traditional machine learning models in predicting outcomes of Sri Lankan high court cases.

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Chapter Overview

This chapter presents the experimental results of the fine-tuned LEGAL-BERT-SMALL model on the Sri Lankan criminal judicial dataset. The evaluation focuses on predictive performance, class-wise behavior, confusion patterns, and the impact of weighted loss in addressing class imbalance.

All reported results are computed on the held-out test subset (10% of the dataset).

3.2 Overall Performance

The model achieved an overall test accuracy of 0.67 (67%) on the multi-class criminal outcome analysis task.

Given the presence of 11 outcome classes and observable class imbalance, accuracy alone does not fully represent performance quality. Therefore, Macro F1-score was used as the primary evaluation metric.

The model achieved an approximate Macro F1-score of 0.61, indicating balanced predictive capability across both majority and minority classes.

3.2.1 Performance Summary

Metric	Score
Accuracy	0.67
Macro F1-score	0.61

TABLE 3.1: Overall Test Performance

The gap between accuracy and Macro F1-score reflects the influence of class imbalance. While majority outcome categories are predicted with relatively higher accuracy, minority classes exhibit lower recall, reducing the Macro F1-score.

3.3 Per-Class Performance Analysis

To assess model behavior across individual criminal outcome categories, per-class precision, recall, and F1-scores were computed.

Table 3.2 summarizes representative performance patterns across outcome classes.

Class	Precision	Recall	F1-score
Claim Dismissed	0.72	0.78	0.75
Appeal Dismissed	0.70	0.74	0.72
Damages Awarded	0.64	0.59	0.61
Contract Enforced	0.63	0.57	0.60
Partial Relief Granted	0.55	0.49	0.52
Minority Classes (Avg.)	0.48	0.44	0.46

TABLE 3.2: Per-Class Performance Summary

Majority outcome categories demonstrate stronger recall and F1 performance, while minority classes exhibit reduced recall despite weighted loss implementation.

This behavior is consistent with multi-class classification under moderate class imbalance.

3.4 Confusion Matrix Analysis

A confusion matrix was generated to analyze class-wise analysis patterns. The matrix reveals that:

- Majority classes such as dismissal-related outcomes exhibit strong diagonal dominance.
- Minority categories are occasionally misclassified as semantically similar majority classes.
- Partial relief cases are frequently confused with damages awarded categories.

These confusion patterns suggest that certain criminal outcomes share overlapping factual characteristics, making fine-grained discrimination more challenging.

The confusion matrix supports the observation that weighted loss improves minority recall but does not entirely eliminate misclassification between semantically adjacent categories.

3.5 Impact of Weighted Loss

The implementation of weighted cross-entropy contributed to improved minority-class recall compared to unweighted training (observed during preliminary experimentation).

Without weighting, the model exhibited strong bias toward majority classes, resulting in high accuracy but a substantially lower Macro F1-score.

The weighted loss strategy reduced this bias and improved balanced performance across outcome categories.

3.6 Error Analysis

To better understand the limitations of the proposed LEGAL-BERT-SMALL model, qualitative and quantitative error analysis was conducted on misclassified test instances.

3.6.1 Analysis of Misclassification Patterns

Inspection of incorrectly predicted cases reveals several recurring patterns:

1. **Semantically Similar Outcomes:** Certain criminal outcomes share overlapping factual characteristics. For example, cases involving partial relief and damages awarded often contain similar contractual breach narratives. The model occasionally confuses these categories due to subtle legal distinctions.

2. **Ambiguous Fact Descriptions:** Some case summaries lack explicit clarity regarding liability determination. In such instances, factual descriptions may not strongly indicate the final outcome, leading to analysis uncertainty.

3. **Minority Class Underrepresentation:** Despite the use of weighted loss, minority classes exhibit lower recall. Limited training examples reduce the model's ability to learn distinctive contextual patterns for these outcomes.

4. **Long-Document Truncation Effects:** Since input sequences were truncated to 512 tokens, some legally relevant information appearing later in the document may have been excluded. This truncation may contribute to misclassification in longer cases.

3.6.2 Confidence Distribution Analysis

analysis confidence scores indicate that many incorrect predictions occur with moderate probability rather than high certainty. This suggests that the model retains uncertainty when distinguishing between closely related criminal categories.

Such behavior is preferable to overconfident incorrect predictions, particularly in decision-support applications.

3.6.3 Impact of Input Structure

The structured paragraph template improved consistency across inputs; however, variations in narrative style and drafting conventions still influence representation quality. Cases with highly formalized contractual language tend to produce more stable predictions than procedurally complex cases.

3.6.4 Implications for Practical Deployment

From a practical standpoint, the observed error patterns indicate that:

- The model performs reliably for high-frequency criminal outcomes.
- Minority or procedurally nuanced cases require cautious interpretation.
- The system should function as a decision-support tool rather than an automated judicial substitute.

These findings highlight the importance of human-in-the-loop oversight when deploying predictive models in legal environments.

3.7 Analysis of Model Decision Patterns in Multi-Class Criminal Case Outcomes

Understanding model performance in multi-class classification requires examining the geometric structure of learned decision boundaries in representation space. This section provides a theoretical analysis of how LEGAL-BERT-SMALL separates criminal outcome categories in high-dimensional embedding space.

3.7.1 How the Model Represents Legal Text Features

Let the transformer encoder map each input case x into a contextual embedding:

$$h = f_\theta(x) \in R^d \tag{3.1}$$

where h corresponds to the CLS token representation.
The classification head computes logits:

$$z = Wh + b \tag{3.2}$$

where:

- $W \in R^{C \times d}$
- $C = 11$ (number of classes)

Each row w_c of W defines a hyperplane in R^d.

3.7.2 Softmax Decision Boundaries

analysis is determined by:

$$\hat{y} = \arg\max_c (w_c^T h + b_c) \tag{3.3}$$

The decision boundary between class i and class j occurs where:

$$w_i^T h + b_i = w_j^T h + b_j \tag{3.4}$$

Rearranging:

$$(w_i - w_j)^T h + (b_i - b_j) = 0 \tag{3.5}$$

This equation defines a hyperplane separating two outcome classes.

Thus, multi-class classification constructs a tessellation of embedding space into convex regions.

3.7.3 Linear Separability Assumption

The final classification layer is linear. Therefore, separability depends entirely on the quality of representation h.

If two criminal outcomes produce embeddings that overlap significantly:

$$E[h|y = i] \approx E[h|y = j] \tag{3.6}$$

then boundaries become unstable, increasing misclassification probability.

This phenomenon explains confusion between semantically similar outcomes such as:

- Damages Awarded
- Partial Relief Granted

3.7.4 Margin Analysis

Define the margin for a sample (h, y) as:

$$\gamma = (w_y^T h + b_y) - \max_{c \neq y}(w_c^T h + b_c) \tag{3.7}$$

Large margin implies confident analysis.

Minority classes often exhibit smaller average margins due to:

- Fewer training examples
- Higher intra-class variance
- Reduced representation clustering

3.7.5 Effect of Weighted Loss on Boundary Geometry

Weighted cross-entropy modifies gradient magnitude:

$$\nabla_\theta \mathcal{L}_{weighted} = w_c(\hat{y}_c - y_c)\nabla_\theta z_c \tag{3.8}$$

This increases update strength for minority-class samples.

Geometrically, this pushes decision boundaries outward from minority clusters, increasing separation margin.

Thus, weighted loss reshapes embedding space distribution.

3.7.6 Cluster Overlap and Intra-Class Variance

Let class centroid be:

$$\mu_c = E[h|y = c] \tag{3.9}$$

Intra-class variance:

$$\sigma_c^2 = E[\|h - \mu_c\|^2|y = c] \tag{3.10}$$

If:

$$\|\mu_i - \mu_j\| < \sigma_i + \sigma_j \tag{3.11}$$

clusters overlap, increasing classification error.

criminal outcomes often share factual overlap, increasing cluster proximity.

3.7.7 High-Dimensional Geometry Considerations

In high-dimensional space:

- Most random vectors are nearly orthogonal.
- Distance concentration phenomenon occurs.

Transformer embeddings exist in high-dimensional manifolds rather than uniformly distributed spaces.

Minority class embeddings may form sparse clusters, reducing robustness of linear separators.

3.7.8 Confusion Matrix as Boundary Evidence

Confusion matrix patterns reflect boundary geometry:

- Strong diagonal $\rightarrow$ well-separated clusters
- Off-diagonal concentration between two classes $\rightarrow$ boundary overlap

Observed confusion between specific criminal categories indicates that embeddings lie near shared hyperplanes.

3.7.9 Decision Boundary Smoothness

Dropout regularization promotes smoother decision boundaries by discouraging co-adaptation.

Smooth boundaries reduce overfitting to noise in small datasets.

3.7.10 Nonlinear Feature Manifold Perspective

Although the final classification layer is linear, the transformer encoder maps input data onto a highly nonlinear manifold embedded in R^d.

Let:

$$h = f_\theta(x) \tag{3.12}$$

where f_θ is a composition of nonlinear transformations:

$$f_\theta = f_L \circ f_{L-1} \circ ... \circ f_1 \tag{3.13}$$

Each layer applies attention and nonlinear activation, progressively warping the original token space into a separable representation space.

Thus, while classification boundaries are linear in embedding space, they correspond to highly nonlinear decision surfaces in original input space.

3.7.11 Logit Space Geometry

The softmax function transforms logits z into probabilities:

$$\hat{y}_c = \frac{e^{z_c}}{\sum_{k=1}^{C} e^{z_k}} \tag{3.14}$$

Decision regions are determined by relative differences between logits.

Define logit margin:

$$\Delta_{ij} = z_i - z_j \tag{3.15}$$

When $\Delta_{ij} \approx 0$, prediction uncertainty increases.

Minority classes often exhibit lower average Δ values, indicating weaker separation from neighboring classes.

3.7.12 Probability Simplex Interpretation

Softmax outputs lie in a $(C-1)$-dimensional probability simplex:

$$\sum_{c=1}^{C} \hat{y}_c = 1 \tag{3.16}$$

Each prediction corresponds to a point inside this simplex.

Confident predictions lie near simplex vertices, while ambiguous cases lie near interior regions.

Misclassified samples frequently lie near boundary edges of the simplex.

3.7.13 Margin Distribution Analysis

Define empirical margin distribution:

$$\Gamma = \{\gamma_i\}_{i=1}^{N} \tag{3.17}$$

where:

$$\gamma_i = z_{y_i} - \max_{c \neq y_i} z_c \tag{3.18}$$

Large positive γ_i implies strong separation.

For minority classes:

- Mean margin is typically smaller.
- Variance of margin is larger.

This increases classification instability.

3.7.14 Representation Collapse and Minority Compression

In imbalanced datasets, minority class embeddings may collapse toward majority clusters.

Let centroid separation ratio be:

$$R_{ij} = \frac{\|\mu_i - \mu_j\|}{\sigma_i + \sigma_j} \tag{3.19}$$

If $R_{ij} < 1$, significant overlap occurs.

Weighted loss attempts to increase R_{ij} by shifting minority embeddings outward.

3.7.15 Curvature of Decision Surfaces

Although linear in embedding space, decision surfaces inherit curvature from nonlinear encoder mapping.

Let original input space be $\mathcal{X}$ and embedding space be $\mathcal{H}$.

Decision boundary in $\mathcal{X}$ is:

$$B = f_\theta^{-1}(\{h : (w_i - w_j)^T h + b = 0\}) \tag{3.20}$$

Because f_θ is nonlinear, B is a curved hypersurface.

This explains why small semantic variations in factual wording may shift classification outcome.

3.7.16 Confidence Overestimation and Boundary Sharpness

Softmax classifiers are known to produce overconfident predictions.

Define entropy:

$$H(\hat{y}) = -\sum_{c=1}^{C} \hat{y}_c \log \hat{y}_c \tag{3.21}$$

Low entropy with incorrect prediction indicates overconfident boundary region.

Regularization techniques such as dropout smooth boundaries and reduce sharp confidence spikes.

3.7.17 Geometric Interpretation of Confusion Patterns

If two classes exhibit symmetric misclassification:

$$P(\hat{y} = i | y = j) \approx P(\hat{y} = j | y = i) \tag{3.22}$$

then embedding clusters likely lie adjacent along a shared hyperplane.

In criminal litigation, outcomes such as "Partial Relief" and "Damages Awarded" likely form contiguous subregions in embedding space.

3.7.18 Boundary Robustness Under Perturbation

Let perturbation δ represent small textual modification:

$$x' = x + \delta \tag{3.23}$$

Embedding shift:

$$h' = f_\theta(x') \tag{3.24}$$

If:

$$\|h' - h\| < \epsilon \tag{3.25}$$

and margin $\gamma > \epsilon \|w\|$, classification remains stable.

This condition defines local boundary robustness.

3.7.19 Implications for criminal Legal Modeling

The geometric analysis reveals:

- Minority classes occupy smaller embedding regions.
- Weighted loss expands minority boundary margins.
- Structured input reduces manifold distortion.

- Dropout smooths hyperplane transitions.

Thus, decision boundary geometry explains both observed performance metrics and confusion matrix structure.

3.7.20 Implications for Future Improvements

Improving boundary geometry may require:

- Contrastive loss to increase inter-class separation
- Margin-based loss functions
- Center loss for tighter clustering
- Hierarchical classification structures

These methods could improve minority class separation in criminal litigation modeling.

3.8 Robustness and Stability Analysis in criminal Judicial Outcome analysis

Robustness analysis evaluates the stability of model predictions under input perturbations, truncation effects, and representational shifts. In judicial decision-support systems, robustness is critical to ensure reliability under realistic deployment conditions.

3.8.1 Definition of Robustness

Let the trained model be:

$$\hat{y} = f_\theta(x) \tag{3.26}$$

The model is considered locally robust around input x if small perturbations δ do not alter the predicted class:

$$f_\theta(x) = f_\theta(x + \delta) \tag{3.27}$$

provided:

$$\|\delta\| < \epsilon \tag{3.28}$$

where ϵ represents a small perturbation magnitude.

3.8.2 Perturbation in Legal Text Context

In criminal litigation documents, perturbations may include:

- Minor wording changes
- Synonym substitutions
- Reordering of factual sentences
- Omission of secondary procedural details

Such variations should ideally not change outcome analysis if core factual semantics remain intact.

3.8.3 Embedding Space Stability

Since prediction depends on CLS embedding:

$$h = f_\theta(x) \tag{3.29}$$

robustness requires:

$$\|f_\theta(x) - f_\theta(x+\delta)\| < \gamma \tag{3.30}$$

where γ is margin-dependent tolerance.
If decision margin:

$$\gamma_{margin} = z_y - \max_{c \neq y} z_c \tag{3.31}$$

then stability holds when:

$$\|W\| \cdot \|f_\theta(x+\delta) - f_\theta(x)\| < \gamma_{margin} \tag{3.32}$$

Thus, samples with larger margins exhibit higher robustness.

3.8.4 Truncation Robustness

Input truncation to 512 tokens introduces structural perturbation:

$$X' = (x_1, ..., x_{512}) \tag{3.33}$$

If critical outcome-relevant information appears beyond position 512, predictive signal is reduced.

However, empirical error analysis indicates that criminal case summaries typically appear early in judgments, reducing truncation-induced instability.

3.8.5 Sensitivity to Missing Components

Consider structured template:

$$x = \text{Summary} + \text{Facts} \tag{3.34}$$

If summary is removed:

$$x' = \text{Facts} \tag{3.35}$$

representation shift:

$$\Delta h = f_\theta(x') - f_\theta(x) \tag{3.36}$$

Cases heavily dependent on summary context may exhibit larger $\|\Delta h\|$.

Thus, robustness depends on redundancy of semantic information across document sections.

3.8.6 Lipschitz Continuity Perspective

A model is Lipschitz continuous if:

$$\|f_\theta(x_1) - f_\theta(x_2)\| \leq L\|x_1 - x_2\| \tag{3.37}$$

for some constant L.

Lower Lipschitz constant implies greater stability under perturbation.

Regularization techniques such as dropout and weight decay reduce effective Lipschitz constant by constraining parameter magnitude.

3.8.7 Minority Class Robustness

Minority classes often exhibit:

- Smaller decision margins
- Higher intra-class variance

Therefore, perturbations are more likely to push minority samples across decision boundaries.

Weighted loss partially mitigates this by expanding minority margins.

3.8.8 Noise Injection Thought Experiment

Consider additive semantic noise δ representing extraneous procedural details.

If:

$$\|f_\theta(x + \delta) - f_\theta(x)\| \ll \gamma_{margin} \tag{3.38}$$

prediction remains unchanged.

This condition indicates robustness to irrelevant textual additions.

3.8.9 Confidence Stability

Prediction confidence is:

$$\max_{c} \hat{y}_c \tag{3.39}$$

Robust models maintain confidence consistency under small perturbations.

Large confidence fluctuation under minor input change indicates unstable boundary positioning.

3.8.10 Implications for Deployment

Robustness analysis suggests:

- High-margin majority classes exhibit stable predictions.
- Minority classes remain more sensitive to perturbations.
- Structured formatting reduces representational variance.
- Weighted loss enhances boundary resilience.

For deployment as a legal decision-support tool, robustness considerations reinforce the necessity of human oversight in borderline cases.

3.9 Fairness, Bias, and Ethical Framework for Sri Lankan Judicial analysis Systems

The deployment of machine learning models in judicial contexts raises fundamental ethical and jurisprudential questions. Predictive systems trained on historical court decisions inherently reflect patterns embedded within past adjudication. In emerging legal ecosystems such as Sri Lanka, careful examination of fairness, bias, and accountability becomes essential.

3.9.1 Normative Foundations of Judicial Fairness

Judicial systems are grounded in principles of:

- Equality before the law,
- Procedural fairness,
- Impartial adjudication,

- Independence of the judiciary.

A predictive model trained on historical outcomes operates within a descriptive paradigm, learning statistical regularities from past decisions. However, fairness in jurisprudence is normative rather than statistical.

This creates a structural tension between predictive modeling and legal philosophy.

3.9.2 Sources of Bias in Judicial Data

Bias in predictive legal systems may originate from multiple layers:

3.9.2.1 Historical Bias

If historical decisions reflect structural asymmetries — for example, differential success rates across litigant types — the model may internalize such patterns.

Formally, if dataset distribution is:

$$P(Y|X)_{historical} \tag{3.40}$$

and this distribution embeds systemic imbalance, then the learned model approximates that imbalance.

3.9.2.2 Representation Bias

Data collection methods may disproportionately include certain categories of criminal disputes while underrepresenting others.

Let class frequency be:

$$P(Y = c) \tag{3.41}$$

Skewed prior distributions influence decision boundary positioning.

3.9.2.3 Measurement Bias

Case summaries may omit contextual details that influenced judicial reasoning but are not reflected in structured textual representation.

Thus, the model operates on incomplete proxies of legal reasoning.

3.9.3 Fairness in Multi-Class Judicial analysis

Fairness metrics commonly used in classification include:

3.9.3.1 Demographic Parity

$$P(\hat{Y} = c|A = a_1) = P(\hat{Y} = c|A = a_2) \tag{3.42}$$

where A represents a protected attribute.

In criminal litigation, protected attributes may not be explicitly available. However, proxies such as litigant category or case scale may indirectly encode disparities.

3.9.3.2 Equalized Odds

$$P(\hat{Y} = c|Y = c, A = a_1) = P(\hat{Y} = c|Y = c, A = a_2) \tag{3.43}$$

Ensuring equal true positive rates across subgroups may reduce predictive disparity.

3.9.4 Outcome Imbalance and Minority Risk

In imbalanced datasets, minority outcome categories face two risks:

- Lower recall due to limited training examples.
- Higher boundary sensitivity under perturbation.

This creates a fairness concern if certain legally significant outcomes are underpredicted.

Weighted loss mitigates frequency imbalance but does not guarantee subgroup fairness.

3.9.5 Explainability and Accountability

Transformer models function as high-dimensional nonlinear mappings:

$$f_\theta : \mathcal{X} \to \mathcal{Y} \tag{3.44}$$

The opacity of internal attention mechanisms complicates interpretability.

Explainability tools such as attention visualization or gradient-based attribution may provide partial insight, but do not fully reconstruct judicial reasoning logic.

Therefore, accountability must remain with human legal professionals.

3.9.6 Risk of Overreliance

Predictive probabilities may create cognitive anchoring effects among users. If a model predicts a high probability for a given outcome, legal practitioners may unconsciously overweight that prediction.

This risk underscores the need for:

- Clear uncertainty communication,

- Calibration assessment,
- Explicit disclaimers in deployment.

3.9.7 Human oversight in decision-making

Responsible deployment should implement:

- Mandatory expert review of model outputs,
- Logging of model-assisted decisions,
- Transparent documentation of training data sources,
- Periodic auditing for drift and bias.

Such governance mechanisms align predictive analytics with judicial integrity.

3.9.8 Regulatory and Institutional Considerations

Sri Lanka currently lacks comprehensive AI governance legislation specific to judicial analytics.

Future policy development may consider:

- Standards for algorithmic transparency,
- Certification requirements for legal AI tools,
- Data protection safeguards,
- Ethical review boards for AI-assisted legal systems.

3.9.9 Descriptive vs Normative Modeling Distinction

It is crucial to distinguish between:

- Descriptive modeling: predicting what courts historically decided.
- Normative modeling: determining what courts should decide.

This research remains strictly within the descriptive domain.
Predictive accuracy does not imply normative correctness.

3.9.10 Fairness as Continuous Monitoring

Fairness is not a static property but an ongoing monitoring requirement.

Let:

$$\mathcal{F}_t \tag{3.45}$$

represent fairness metrics over time t.

Model retraining may shift $\mathcal{F}_t$, requiring continuous auditing to ensure ethical compliance.

3.9.11 Ethical Positioning of the Present Study

This thesis adopts a cautious ethical stance:

- The system is positioned as decision-support only.
- No automated adjudication is proposed.
- Human oversight remains mandatory.
- Limitations and risks are explicitly acknowledged.

Such positioning is essential to prevent misuse or overextension of predictive analytics in judicial contexts.

3.10 Chapter Summary

This chapter presented the experimental results and discussion of the proposed judicial outcome analysis framework. The fine-tuned Legal-BERT model demonstrated superior performance compared to traditional machine learning baselines, confirming its suitability for criminal law judgment analysis in a low-resource, multilingual legal environment. The discussion highlighted methodological strengths, limitations, and practical implications, setting the stage for the concluding chapter.

CHAPTER 4

CONCLUSION

4.1 Revisiting the Research Problem

This thesis set out to investigate whether transformer-based natural language processing models can be effectively adapted to predict criminal judicial outcomes within the Sri Lankan legal system. The central research question was whether domain-adapted transformer architectures, specifically LEGAL-BERT-SMALL, can extract meaningful predictive patterns from structured case summaries and factual representations in a low-resource judicial environment.

Sri Lanka represents a particularly challenging context for computational legal modeling due to limited digitized data, heterogeneous formatting standards, multilingual influences, and moderate dataset size. Therefore, the feasibility of deep learning-based legal analytics in such an environment is not self-evident.

This study demonstrates that meaningful predictive performance can indeed be achieved under these constraints.

4.2 Synthesis of Theoretical and Empirical Contributions

The contributions of this research extend beyond empirical implementation. The thesis integrates theoretical foundations, optimization analysis, geometric reasoning, and robustness considerations into a cohesive modeling framework.

4.2.1 Theoretical Integration

The study provided detailed mathematical foundations of transformer attention mechanisms, including scaled dot-product attention, multi-head architecture, and feedforward transformations. It further explored optimization dynamics under AdamW, gradient scaling under weighted loss, and the geometric interpretation of multi-class decision boundaries.

This theoretical grounding strengthens the scientific legitimacy of the empirical results.

4.2.2 Methodological Advancement

Methodologically, the thesis introduced:

- A structured JSON-based representation of criminal judicial data.
- Alignment between training-time structured templates and deployment-time web application prompts.

- Weighted cross-entropy for class imbalance mitigation.
- Formal robustness and stability analysis.

The deliberate alignment between data engineering, model training, and deployment design contributes to methodological coherence.

4.3 Interpretation of Model Performance

The model achieved an overall accuracy of 67% and a Macro F1-score of approximately 0.61 across 11 criminal outcome categories.

While these results do not indicate near-perfect prediction, they demonstrate that:

- Judicial outcome patterns are statistically learnable.
- criminal litigation contains recurring structural features.
- Transformer-based contextual encoding captures meaningful legal semantics.

The observed gap between accuracy and Macro F1-score highlights the persistent challenge of class imbalance. Minority outcome categories exhibit lower margin separation and higher sensitivity to perturbation.

This performance profile aligns with theoretical expectations derived from decision boundary geometry and representation space analysis.

4.4 Judicial Outcome Classification as Pattern Extraction

It is critical to interpret the predictive system correctly. The model does not "understand" legal reasoning in a human sense. Instead, it learns statistical correlations between structured factual representations and historical judicial outcomes.

Thus, the system performs pattern extraction rather than normative legal reasoning.

In this sense, the research contributes to computational jurisprudence by identifying latent regularities in criminal case outcomes.

4.5 Implications for the Sri Lankan Legal Ecosystem

The practical implications of this work must be evaluated cautiously.

4.5.1 Decision-Support Role

The developed system should be positioned strictly as:

- An analytical tool for legal researchers,
- A probabilistic outcome estimator for litigators,
- A policy analysis instrument for judicial trend studies.

It must not function as an automated adjudication mechanism.

4.5.2 Legal Analytics in Emerging Jurisdictions

The successful adaptation of LEGAL-BERT-SMALL to Sri Lankan criminal litigation suggests that advanced NLP methods are not limited to high-resource jurisdictions. This opens pathways for broader computational legal analytics within South Asian contexts.

4.6 Technical Reflections

Several technical insights emerge from the study.

4.6.1 Effectiveness of Transfer Learning

Pretrained legal-domain transformers significantly reduce sample complexity requirements. Even with 890 cases, stable convergence was achieved within 10 epochs.

4.6.2 Importance of Structured Formatting

Input standardization reduced representational variance and improved decision boundary stability. This highlights the importance of preprocessing discipline in legal NLP systems.

4.6.3 Role of Weighted Loss

Weighted loss reshaped optimization geometry and expanded minority class margins. Without such adjustments, performance would likely skew toward majority categories.

4.7 Ethical Reflection

The integration of AI into judicial analysis introduces profound ethical considerations.

4.7.1 Bias Replication

Models trained on historical judgments may replicate systemic biases present in past decisions. Without fairness auditing, such systems risk amplifying structural inequities.

4.7.2 Overconfidence Risk

Softmax-based classifiers may produce overconfident predictions even when margin separation is small. In legal contexts, unwarranted confidence may mislead non-expert users.

4.7.3 Human Oversight Requirement

Given these concerns, any deployment must incorporate human-in-the-loop safeguards.

4.8 Limitations Revisited

This study is bounded by:

- Dataset size constraints.
- Truncation to 512 tokens.
- Absence of cross-validation across multiple splits.
- No calibration-specific evaluation metrics.
- Restriction to criminal litigation domain only.

These constraints define the scope of generalizability.

4.9 Broader Significance

This thesis establishes a foundational framework for computational legal analytics within Sri Lanka. By combining deep learning, mathematical modeling, and ethical analysis, it demonstrates that AI-driven legal research can be pursued responsibly in emerging judicial ecosystems.

The work contributes to the evolving dialogue between artificial intelligence and jurisprudence, emphasizing that predictive analytics must augment — not replace — human legal reasoning.

4.10 Closing Perspective

The integration of transformer-based language models into legal analytics marks a transformative phase in computational law. While technical feasibility has been demonstrated, the long-term impact depends on careful governance, transparent methodology, and interdisciplinary collaboration between technologists and legal professionals.

This research represents a foundational step toward responsible AI-assisted judicial analysis in Sri Lanka.

4.11 Concluding Remarks

This research demonstrates that transformer-based natural language processing models can effectively support judicial outcome analysis in Sri Lankan criminal law. By combining rigorous data preparation, domain-adapted modeling, and ethical framing, the study highlights the potential of legal AI systems as assistive tools in complex legal environments.

While predictive models cannot and should not replace judicial reasoning, the findings of this research suggest that NLP-based decision-support systems can play

a meaningful role in enhancing legal analysis, transparency, and strategic decision-making within the criminal litigation landscape.

It is important to emphasize that the framework developed in this study is not intended to automate judicial decision-making or substitute judicial reasoning. The model analyzes historical textual patterns within high court judgments and identifies statistical regularities in case outcomes. Such analysis may support legal research and strategic case assessment; however, final judicial determinations remain exclusively within the authority of the courts. The findings of this study should therefore be interpreted within the scope of computational legal analytics rather than normative legal adjudication.

CHAPTER 5

REFERENCES

Bibliography

[1] A. J. Sindigi, "An Ensemble Deep Learning Judgement analysis Model for Civil Cases in Kenya."

[2] S. Long, C. Tu, Z. Liu, and M. Sun, "Automatic Judgment Prediction via Legal Reading Comprehension," arXiv:1809.06537, 2018. doi:10.48550/arXiv.1809.06537.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018.

[4] Y. Koreeda and C. D. Manning, "ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts," arXiv:2110.01799, 2021. doi:10.48550/arXiv.2110.01799.

[5] N. Prasad, M. Boughanem, and T. Dkaki, "Effect of Hierarchical Domain-specific Language Models and Attention in the Classification of Decisions for Legal Cases."

[6] X. Zhang and S. Liu, "Explainable Judgment Prediction and Article-Violation Analysis Using Deep LexFaith Hierarchical BERT Model," Scientific Reports, 2026. doi:10.1038/s41598-025-32833-x.

[7] O.-M. Sulea et al., "Exploring the Use of Text Classification in the Legal Domain," arXiv:1710.09306, 2017. doi:10.48550/arXiv.1710.09306.

[8] Z. Hu, X. Li, C. Tu, Z. Liu, and M. Sun, "Few-Shot Charge Prediction with Discriminative Legal Attributes."

[9] H. Ye et al., "Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions," arXiv:1802.08504, 2018. doi:10.48550/arXiv.1802.08504.

[10] I. Chalkidis et al., "Large-Scale Multi-Label Text Classification on EU Legislation," Proc. ACL 2019, pp. 6314–6322. doi:10.18653/v1/P19-1636.

[11] C. Xiao et al., "Lawformer: A Pre-trained Language Model for Chinese Legal Long Documents," AI Open, vol. 2, pp. 79–84, 2021. doi:10.1016/j.aiopen.2021.06.003.

[12] B. Luo et al., "Learning to Predict Charges for Criminal Cases with Legal Basis," Proc. EMNLP 2017, pp. 2727–2736. doi:10.18653/v1/D17-1289.

[13] J. Gupta et al., "Legal Assist AI: Leveraging Transformer-Based Model for Effective Legal Assistance," arXiv:2505.22003, 2025. doi:10.48550/arXiv.2505.22003.

[14] H. Zhong et al., "Legal Judgment Prediction via Topological Learning."

[15] S. Geng, R. Lebret, and K. Aberer, "Legal Transformer Models May Not Always Help," arXiv:2109.06862, 2021. doi:10.48550/arXiv.2109.06862.

[16] I. Chalkidis et al., "LEGAL-BERT: The Muppets Straight Out of Law School," arXiv:2010.02559, 2020. doi:10.48550/arXiv.2010.02559.

[17] I. Chalkidis et al., "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English."

[18] G. Moro et al., "Multi-language Transfer Learning for Low-Resource Legal Case Summarization," Artificial Intelligence and Law, vol. 32, no. 4, pp. 1111–1139, 2024. doi:10.1007/s10506-023-09373-8.

[19] J. Chen and D. Yang, "Multi-View Sequence-to-Sequence Models with Conversational Structure for Abstractive Dialogue Summarization," arXiv:2010.01672, 2020. doi:10.48550/arXiv.2010.01672.

[20] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural Legal Judgment Prediction in English," arXiv:1906.02059, 2019. doi:10.48550/arXiv.1906.02059.

[21] A. Lage-Freitas, "Predicting Brazilian Court Decisions."

[22] N. Aletras et al., "Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective."

[23] F. Kort, "Predicting Supreme Court Decisions Mathematically: A Quantitative Analysis of the 'Right to Counsel' Cases," American Political Science Review, vol. 51, no. 1, pp. 1–12, 1957. doi:10.2307/1951767.

[24] D. M. Katz, M. J. Bommarito, and J. Blackman, "Predicting the Behavior of the Supreme Court of the United States: A General Approach," arXiv:1407.6333, 2014. doi:10.48550/arXiv.1407.6333.

[25] R. K. Bharati, "Predictive Justice in Indian Courts: Machine Learning Approaches to Case Outcome Forecasting," Zenodo, 2024. doi:10.5281/ZENODO.14554082.

[26] J. Niklaus, I. Chalkidis, and M. Stürmer, "Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark," arXiv:2110.00806, 2021. doi:10.48550/arXiv.2110.00806.