LegalSeg: Unlocking the Structure of Indian Legal Judgments Through Rhetorical Role Classification

Anonymous ACL submission

Abstract

In this paper, we address the task of semantic segmentation of legal documents through rhetorical role classification, with a focus on Indian legal judgments. We introduce `LegalSeg`, the largest annotated dataset for this task, comprising over 7,000 documents and 1.4 million sentences, labeled with 7 rhetorical roles. To benchmark performance, we evaluate multiple state-of-the-art models, including Hierarchical BiLSTM-CRF, TransformerOverInLegalBERT (ToInLegalBERT), Graph Neural Networks (GNNs), and Role-Aware Transformers, alongside an exploratory `RhetoricLLaMA`, an instruction-tuned large language model. Our results demonstrate that models incorporating broader context, structural relationships, and sequential sentence information outperform those relying solely on sentence-level features. Additionally, we conducted experiments using surrounding context and predicted or actual labels of neighboring sentences to assess their impact on classification accuracy. Despite these advancements, challenges persist in distinguishing between closely related roles and addressing class imbalance. Our work underscores the potential of advanced techniques for improving legal document understanding and sets a strong foundation for future research in legal NLP.

1 Introduction

The increasing complexity of legal documents necessitates the use of advanced NLP techniques to aid in their understanding and analysis. Semantic segmentation of legal texts into rhetorical roles is essential for improving the efficiency of legal research, enhancing access to justice, and supporting automated legal decision-making systems. It also facilitates various downstream tasks, such as legal search, summarization, and case analysis. Traditional methods often struggle with the intricacies of legal language, making it imperative to develop models that can accurately classify and interpret these documents. This paper addresses the challenge of semantic segmentation in legal documents, with a focus on the Indian judiciary's legal judgments. Historically, the lack of large-scale annotated datasets has hindered the effective training of state-of-the-art ML models in this domain.

Corpus	Country	Language	# Cases	Total # Sentences	# Labels	Domain Coverage
(Bhattacharya et al., 2019)	India	English	50	9,380	7	Supreme Court
(Majumder and Das, 2020)	India	English	60	-	7	Supreme Court High Courts Tribunal Courts
(Malik et al., 2022)	India	English	100	21,184	13	Supreme Court Bombay High Court Kolkata High Court
(Kalamkar et al., 2022)	India	English	354	40,305	13	Supreme Court High Courts District Courts
(Marino et al., 2023)	India	English	275	31,865	13	Supreme Court High Courts District Courts
(Marino et al., 2023)	Italy	Italian	1,488	95,920	5	Civil Law of Italian Courts
(Modi et al., 2023)	India	English	265	26,304	13	Not Mentioned
`LegalSeg` (Ours)	India	English	7,120	1,487,149	7	Supreme Court High Courts

Table 1: Overview of Legal Corpora for Rhetorical Role Classification

Previous research in this domain has highlighted the importance of annotated datasets for training effective models. However, many existing studies have relied on relatively small annotated datasets, limiting their applicability and effectiveness in real-world scenarios. For instance, datasets such as those compiled by (Bhattacharya et al., 2019; Kalamkar et al., 2022) and (Malik et al., 2022) provided valuable insights but were constrained in size, thereby restricting the scope of their findings. In contrast, this study leverages a newly compiled dataset, `LegalSeg`, which consists of 7,120 annotated legal documents and 1,487,149 sentences. This dataset is considerably larger than those used in previous research, particularly in terms of its volume and diversity, as illustrated in Table 1, which summarizes various legal corpora for rhetorical role classification. An example of how a legal judgment is semantically segmented into rhetorical roles is illustrated in Figure 1. As shown, an unstructured legal document is broken down into coherent parts, each annotated with a rhetorical

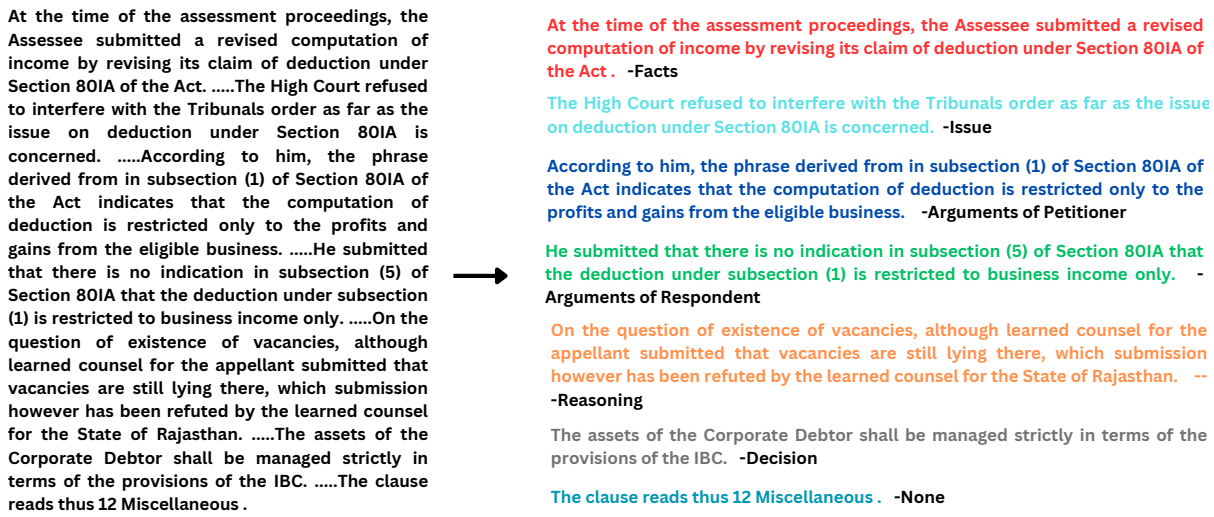

Figure 1: Example illustrating document segmentation using rhetorical roles. The left side shows an excerpt from a legal document, while the right side demonstrates the segmentation and labeling of sentences.

role label such as Facts, Reasoning, or Decision. This segmentation is critical for understanding the flow of arguments and supporting the automation of legal processes.

We implemented several SoTA models to evaluate the effectiveness of our dataset. Among these, the Hierarchical BiLSTM-CRF model (Bhattacharya et al., 2019) captures contextual information using a hierarchical approach, while MultiTask Learning (MTL) incorporates label shift prediction to refine the identification of rhetorical roles by considering role transitions (Malik et al., 2022). Additionally, we explored LEGAL-TransformerOverBERT (LEGALToBERT), a hierarchical architecture stacking a transformer encoder over a legal-domain-specific BERT model, which effectively captures sentence relationships and positional encoding within legal documents (Marino et al., 2023).

In addition to these models, we introduce novel approaches, including InLegalToBERT, Graph Neural Networks (GNNs), and Role-Aware Transformers, that have not been previously explored in the context of rhetorical role classification in legal texts. InLegalToBERT, a variant of LEGALToBERT, incorporates the total number of sentences as an additional feature to enhance the model's ability to capture positional information within documents. GNNs leverage the structural relationships between sentences by representing them as nodes in a graph, allowing for effective propagation of information across sentence pairs and capturing both local and global context. Role-Aware Transformers, on the other hand, utilize specialized embeddings to incorporate rhetorical role-specific information into pre-trained models, improving the model's ability to differentiate between closely related roles.

In an era dominated by LLMs, we also developed and investigated `RhetoricLLaMA`, a fine-tuned version of the open-source LLaMA-2-7B architecture specifically designed for the task of semantic segmentation in legal documents. While the initial performance of `RhetoricLLaMA` was lower than anticipated, it demonstrated the potential of instruction-tuned LLMs for handling complex legal language. This highlights the promise of LLMs for legal NLP, although further fine-tuning and domain-specific adaptations are needed to fully unlock their potential for tasks such as rhetorical role classification.

Our contributions to this work are as follows:

1. Introduction of the `LegalSeg` dataset, the largest annotated dataset for rhetorical role classification in legal documents.
2. Implementation and evaluation of SoTA models for semantic segmentation of legal texts.
3. The development of novel models, including InLegalToBERT, Graph Neural Networks (GNNs), and Role-Aware Transformers, which enhance representation and context handling for rhetorical role classification.
4. Exploration of instruction-tuned LLMs, through the development of `RhetoricLLaMA`, highlighting the potential and limitations of LLMs in rhetorical role classification.

To ensure reproducibility, we have made the `LegalSeg` dataset and the code for all our models accessible via an anonymous link[1].

[1] https://anonymous.4open.science/r/LegalSeg-68D5/

2 Related Work

Recent advancements in legal text processing have spurred significant research efforts aimed at automating various tasks such as semantic segmentation, judgment prediction, and summarization of legal documents. However, much of this work relies heavily on manual annotation, with many studies focusing on the intricacies of annotation processes, including the development of annotation guidelines, IAA studies, and the curation of gold standard corpora. For instance, the TEMIS corpus, which consists of 504 sentences annotated both syntactically and semantically, was developed to enhance understanding of legislative texts (Venturi, 2012). Additionally, an in-depth annotation study highlighted low assessor agreement for labels such as Facts and Reasoning (Wyner et al., 2013). In the Indian context, datasets like ILDC (Malik et al., 2021), PredEx (Nigam et al., 2024) and (Nigam et al., 2022; Malik et al., 2022; Nigam et al., 2023a,b) have highlighted the growing role of AI in legal judgments, with an emphasis on explainability. Research in LJP with LLMs, such as (Vats et al., 2023) and (Nigam et al., 2024), has experimented with models like GPT-3.5 Turbo and LLaMA-2 on Indian legal datasets.

Several efforts have been made to automate the annotation task itself. For example, (Wyner, 2010) discusses methodologies that employ NLP tools to analyze 47 criminal cases from California courts. Initial experiments aimed at understanding rhetorical roles within court documents were often intertwined with broader goals of document summarization (Saravanan et al., 2008).

Further contributions include segmenting documents into functional parts (e.g., Introduction, Background) and issue-specific sections (Šavelka and Ashley, 2018). A semi-supervised training method for identifying factual versus non-factual sentences was explored by (Nejadgholi et al., 2017) using a fastText classifier. The comparison between rule-based scripts and machine learning approaches for rhetorical role identification was conducted by (Walker et al., 2019) demonstrating the efficacy of both methodologies in this context.

In recent studies, (Bhattacharya et al., 2019) proposed a CRF-BiLSTM model specifically for assigning rhetorical roles to sentences in Indian legal documents (Bhattacharya et al., 2019; Malik et al., 2022) created a comprehensive rhetorical role corpus annotated with 13 fine-grained roles and developed a multi-task learning model for prediction tasks. (Kalamkar et al., 2022) constructed a corpus consisting of 354 Indian legal documents annotated with rhetorical roles across 40,305 sentences and introduced a transformer-based baseline model.

Moreover, (Malik et al., 2022) proposed an MTL framework that significantly improved classification scores by leveraging a Hierarchical BiLSTM with CRF architecture. (Marino et al., 2023) introduced LEGAL-ToBERT, which integrates a transformer encoder atop a legal-domain-specific BERT model tailored for both Italian and Indian datasets. More recently, the HiCuLR framework (Santosh et al., 2024) introduced hierarchical curriculum learning for rhetorical role labeling, progressively training models with a structured, easy-to-difficult learning strategy, which enhances performance across multiple rhetorical role datasets.

3 Task Description

The goal of this research is to develop models capable of performing semantic segmentation on legal documents by identifying and classifying rhetorical roles (RR) within the text. Let $D = \{d_1, d_2, \ldots, d_n\}$ represent a collection of legal documents, where $d_i \in D$ consists of a sequence of sentences $S_i = \{s_{i1}, s_{i2}, \ldots, s_{im}\}$, with m representing the number of sentences in document d_i. The task is to assign a rhetorical role label $y_{ij} \in Y$ to each sentence s_{ij}, where Y is the predefined set of 7 rhetorical role labels.

Formally, the task can be described as:

$$f : S_i \rightarrow Y$$

$$Y = \left\{ \begin{array}{l} \text{Facts, Issue, Arguments of Petitioner,} \\ \text{Arguments of Respondent, Reasoning,} \\ \text{Decision, None} \end{array} \right\}$$

where f is a function that maps each sentence s_{ij} in a document d_i to its corresponding rhetorical role label y_{ij}. Thus, the goal is to find:

$$f(s_{ij}) = y_{ij}, \quad \forall s_{ij} \in S_i, \quad y_{ij} \in Y$$

The input to the system is a legal document d_i, and the output is a sequence of rhetorical role labels corresponding to each sentence in the document:

$$f(S_i) = \{y_{i1}, y_{i2}, \ldots, y_{im}\}, \quad y_{ij} \in Y$$

4 Dataset

In this research, we present the LegalSeg Judgment Dataset, the largest annotated dataset of legal judgments in the English language, specifically focused on rhetorical role segmentation. This dataset represents a significant advancement in the field of legal Natural Language Processing (L-NLP), especially in the context of the Indian judiciary. It aims to address existing gaps in annotation comprehensiveness by offering a rich resource of annotated legal judgments designed to facilitate tasks such as sentence segmentation and semantic labeling.

4.1 Dataset Compilation

The Dataset comprises 16,000 legal judgments sourced from the IndianKanoon database, a widely used legal search engine for Indian legal documents. These judgments were sourced from the Supreme Court of India and High Courts, ensuring a diverse selection of cases across multiple domains of law, such as criminal, civil, and constitutional matters.

After the initial collection, we reduced the dataset to approximately 8,000 case files. During the annotation process, sentences were categorized based on their rhetorical roles, including facts, arguments, rulings, statutes, and more. After a preprocessing step, where cases deemed too short or lacking discernible segments were removed, the final dataset was reduced to 7,120 case files.

4.2 Dataset Preparation and Preprocessing

To train and evaluate models for this task, the dataset was divided into training, validation, and test sets using a 70-20-10 split, which comprises 4,984, 1,424, and 712 documents correspondingly. This split ensures a robust set of data for both training and evaluating models. Additionally, we computed various statistics regarding the documents and sentences within the dataset, including the average number of sentences per document and token counts presented in Table 2. Furthermore, the distribution of rhetorical roles within the dataset is visualized in a pie chart Appendix, Figure 13.

4.3 Annotation Process

The annotation process was performed by a group of 10 legal experts, consisting of third and fourth year law students selected for their strong academic backgrounds and familiarity with legal processes. The annotation process spanned from April 2022 to October 2023. Each annotator was assigned 10 judgments per week, ensuring detailed attention to every document.

Statistic	Train Set	Validation Set	Test Set
# Documents	4,984	1,424	712
Total # Sentences	1,087,740	258,649	140,760
Avg. # Sentences per Doc	218.25	181.64	197.70
Avg. # Tokens per Sentence	31.49	30.34	30.56
Sentence Count per Label			
Facts	167,858	51,279	24,573
Issue	12,557	4,201	1,831
AoP	61,158	20,353	13,157
AoR	49,361	15,735	9,421
Reasoning	200,558	66,461	36,354
Decision	19,372	7,523	3,794
None	576,876	93,097	51,630
Average Number of Tokens per Label			
Facts	31.30	29.69	29.39
Issue	37.36	37.93	41.24
AoP	35.65	33.84	32.86
AoR	34.78	32.57	32.03
Reasoning	31.73	31.27	30.84
Decision	23.33	22.57	22.56

Table 2: Dataset Statistics for LegalSeg Dataset

4.4 Quality Control

To ensure the annotation accuracy and consistency, we implemented multiple levels of quality control:

- Senior Expert Review: All disagreements in annotations were escalated to them for resolution.
- Regular Training: Annotators participated in biweekly training sessions, ensuring consistency in understanding and interpreting legal content. Ambiguous or difficult segments were regularly discussed and standardized.

4.5 Annotation Roles

Legal experts annotated each sentence with one of the following rhetorical roles:

- **Facts:** Sentences that describe the sequence of events that led to the case. These typically involve details of the circumstances and actions related to the case, providing a factual narrative of the case's background, and details about the parties involved, including key dates, events, and parties involved.
- **Issue:** Sentences that outline the legal issues or questions being addressed in the case. These often identify the core legal disputes or points of law that the court must resolve to make a ruling.
- **Arguments of Petitioner (AoP):** Sentences representing the arguments made by the petitioner (the party bringing the case to court). These include claims, reasoning, and justifications presented by the petitioner to support their position and persuade the court to rule in their favor.

- **Arguments of Respondent (AoR):** Sentences that summarize the arguments made by the respondent (the party defending against the case). Like the petitioner's arguments, these statements offer counterpoints, legal interpretations, and rebuttals designed to challenge the petitioner's claims and persuade the court to rule in the respondent's favor.
- **Reasoning:** Sentences that provide the rationale or reasoning behind the court's decision. This includes the application of legal principles and precedents, as well as the logic that connects the facts and arguments to the final ruling. This label captures how the court justifies its decision in light of the legal issues presented.
- **Decision:** Sentences that reflect the final ruling or judgment of the court. This label marks the conclusion of the case, where the court issues its verdict or order, stating the outcome of the case based on its reasoning, such as granting relief, compensation, or dismissing the case.
- **None:** Sentences that do not fall under any of the defined rhetorical roles. These sentences may include procedural information, non-substantive remarks, legal jargon, or content that is not directly relevant to the legal analysis or judgment.

This annotation schema follows closely with prior works in rhetorical role segmentation, as demonstrated by the datasets used in similar research efforts such as those by (Bhattacharya et al., 2019; Kalamkar et al., 2022; Malik et al., 2022).

5 Methodology

This section outlines the methodology employed for the task of semantic segmentation of legal documents via rhetorical roles. We implemented several SoTA methods while also exploring new techniques. These methodologies collectively aim to enhance the model's ability to understand and classify rhetorical roles in legal texts by incorporating structural, contextual, and sequential information. Each technique addresses different aspects of the complex relationships between sentences in legal documents, contributing to more accurate and context-aware classification outcomes.

5.1 TransformerOverInLegalBERT (ToInLegalBERT)

Inspired by the work of (Marino et al., 2023), the TransformerOverInLegalBERT (ToInLegalBERT) model follows a hierarchical architecture. It consists of four main components: (i) an InLegalBERT token-level encoder, (ii) a sentence-level positional encoder, (iii) a sentence-level encoder, and (iv) a prediction layer.

The process begins by splitting the document into sentences and tokenizing them. Each sentence is then input into the ToInLegalBERT token-level encoder, where the pooled output—specifically, the hidden representation of the [CLS] token—is extracted. These pooled outputs are subsequently fed into the positional layer to create a position-dependent encoding for each sentence within the document. The encoded representations are then passed to the sentence-level encoder, which captures the relationships between sentences in the document, and finally, these outputs are directed to the prediction layer for rhetorical role classification. This method incorporates both the local context of sentences and their position in the document, enabling better rhetorical role classification.

5.2 Hierarchical BiLSTM CRF Classifier

We also implemented the BiLSTM-CRF model proposed by (Bhattacharya et al., 2019), which combines Bidirectional Long Short-Term Memory (BiLSTM) with a Conditional Random Field (CRF) layer. The input to this model is sentence embeddings generated using a sent2vec model trained on Indian Supreme Court judgments. These sentence embeddings are passed through a BiLSTM model, which captures sequential dependencies between sentences. The CRF layer on top of the BiLSTM ensures that the predicted rhetorical role labels adhere to the structured nature of legal documents. This model predicts the rhetorical role for each sentence by considering the context provided by neighboring sentences.

5.3 Multi-Task Learning (MTL)

Inspired by the Multi-Task Learning framework proposed by (Malik et al., 2022), we adopt an MTL approach where rhetorical role prediction is the main task, and label shift prediction serves as the auxiliary task. The model consists of two components: a label shift detection component and a rhetorical role prediction component. The intuition is that the label shift between sentences (indicating a change in rhetorical role) helps improve role classification. The label shift detection component predicts whether a shift in rhetorical role occurs at the ith sentence, while the rhetorical role classification component predicts the rhetorical role for

that sentence. The output from both components is concatenated and passed to the CRF layer for final role predictions. The overall loss function for the MTL model is: $L = \lambda L_{\text{shift}} + (1 - \lambda) L_{\text{RR}}$, where L_{shift} corresponds to label shift prediction, L_{RR} corresponds to rhetorical role classification, and λ is a hyperparameter balancing the two tasks. This method allows the model to learn dependencies between sentences more effectively.

5.4 InLegalBERT Variants

We experimented with different configurations of the InLegalBERT (Paul et al., 2023) model to improve performance. These configurations vary in terms of the number of sentences provided as input during training and inference:

- InLegalBERT(i): The model is trained and tested using only the current sentence i.
- InLegalBERT(i-1, i): The model is trained with the previous sentence $i - 1$ and the current sentence i.
- InLegalBERT(i-2, i-1, i): The model is trained using the previous two sentences $i - 2, i - 1$ and the current sentence i.
- InLegalBERT(i-1, i, i+1): The model is trained with the previous sentence, current sentence, and the next sentence.

5.5 Incorporate Previous Sentence and Label

We further explored methods where we provide the model with additional contextual information. In one variant, we concatenate the current sentence with the previous sentence and the true label of the previous sentence during training. This approach allows the model to leverage contextual information from preceding sentences to make better predictions. Another variant replaces the true label with the predicted label of the previous sentence during inference, simulating real-world conditions where true labels are unavailable. This method helps the model handle prediction errors and learn sequential dependencies between rhetorical roles.

5.6 Self-Supervised Pre-Training with Role-Aware Transformers

We propose a novel method, Role-Aware Transformer model, a variant of the standard transformer architecture that incorporates role embeddings to represent different rhetorical roles (e.g., Facts, Issues, Arguments). The model is pre-trained in a self-supervised manner on a large corpus of legal documents. During pre-training, it predicts masked tokens while considering the rhetorical role of each sentence, thereby learning how these roles influence language within legal contexts.

5.7 GNN with Document Context

To capture the structural relationships between sentences, we propose a method that leverages Graph Neural Networks (GNNs). In this approach, each sentence in a document is represented as a node in a graph, and the edges between nodes are based on sentence order or semantic similarity. Sentence embedding generated via InLegalBERT, a pre-trained language model on the Indian legal domain, serves as a node feature. The GNN processes the graph by propagating information between connected sentences, allowing the model to capture both local and global contextual dependencies. The GNN processes this graph, allowing for information propagation and aggregation across connected sentences, which enhances understanding of interdependencies between sentences.

5.8 `RhetoricLLaMA`

To leverage the power of LLMs for rhetorical role prediction, we implemented `RhetoricLLaMA`, an instruction-tuned model based on LLaMA-2-7B (Touvron et al., 2023). For this specific task, we fine-tuned the LLaMA-2-7B model on our `LegalSeg` dataset using instruction-tuning, a method designed to guide the model's understanding of specific tasks through a set of structured instructions.

To enhance the model's ability to segment legal documents accurately, we developed a set of 16 instruction sets tailored to the nature of rhetorical role classification in legal texts. These instructions provided the model with explicit guidance on how to handle the different rhetorical roles in a legal document. A complete list of these instruction sets can be found in Table 5 in the Appendix.

6 Evaluation Metrics

To evaluate the performance of models, we adopt a set of standard metrics commonly used in classification tasks. For each sentence in the dataset, the predicted label (rhetorical role) is considered correct if it matches the label assigned by the majority opinion of human annotators.

We utilize macro-averaged Precision, Recall, F-score, Accuracy, and Matthew Correlation Coefficient (MCC) (Chicco and Jurman, 2020) as our

Model	Precision	Recall	F1-Score	Accuracy	MCC
MTL	0.59	0.40	0.37	0.41	**0.78**
GNN	0.64	0.50	0.54	**0.64**	0.40
Role-Aware	0.21	0.20	0.14	0.50	0.04
ToInLegalBERT	0.67	0.60	0.62	**0.64**	0.52
`RhetoricLLaMA`	0.19	0.15	0.09	0.39	0.02
InLegalBERT(i)	0.57	0.45	0.49	0.53	0.45
InLegalBERT(i-1, i)	0.60	0.53	0.55	0.57	0.50
InLegalBERT(i-2, i-1, i)	0.62	0.56	0.58	0.59	0.52
InLegalBERT(i-1, i, i+1)	0.61	0.56	0.58	0.59	0.52
InLegalBERT(i-1, label_t, i)	0.63	0.32	0.34	0.45	0.22
InLegalBERT(i-1, label_p, i)	0.54	0.46	0.48	0.52	0.35
Hier_BiLSTM CRF	**0.78**	**0.77**	**0.77**	0.62	0.68

Table 3: Performance Comparison of Models on Rhetorical Role Classification. In the Model column, i indicates the current sentence, $i-1$ means the previous sentence, and $i+1$ means the next sentence. `label_t` and `label_p` refer to the true and predicted labels of the previous sentences. The best results are in bold.

primary evaluation metrics. Macro-averaging involves calculating these metrics for each class separately and then taking their average. This method is particularly beneficial as it prevents bias towards high-frequency classes, ensuring that all rhetorical roles are treated equally in the evaluation process.

7 Results and Analysis

In this section, we present the results of our experiments on rhetorical role classification and analyze the performance of different models. Table 3 summarizes the evaluation metrics for each model.

7.1 Model Performance

The Hierarchical BiLSTM-CRF model demonstrates the strongest overall performance. Its ability to capture sequential dependencies through BiLSTM and refine predictions using CRF layers enables it to handle the complexity of legal text and label transitions effectively. The model's high scores across all metrics suggest that it excels at identifying the various rhetorical roles while maintaining consistency in predictions.

The ToInLegalBERT model, which incorporates sentence-level positional encodings and hierarchical structure, also performs well, indicating that incorporating the document's structure into the model helps improve the accuracy of rhetorical role classification. This model's strong performance in both precision and recall highlights its ability to handle diverse sentence types in legal documents.

Graph Neural Networks (GNN) and the various InLegalBERT configurations show competitive results. The GNN approach leverages structural relationships between sentences by representing them as nodes in a graph, allowing for effective propagation of contextual information. Among the InLegalBERT variants, the model using the current sentence and two preceding sentences performs the best, indicating that a broader sentence context helps the model make more informed predictions about a sentence's rhetorical role.

The Multi-Task Learning (MTL) model, which incorporates label shift prediction as an auxiliary task, shows moderate performance. While the label shift information is expected to aid in rhetorical role classification, the added complexity may have introduced challenges in learning effective sentence-level dependencies. Nevertheless, this approach remains promising for capturing transitions between different rhetorical roles.

The `RhetoricLLaMA` model did not perform as strongly as expected. While LLMs have shown great promise in various NLP tasks, it appears that further fine-tuning and domain-specific adaptations are needed for rhetorical role classification in legal documents. Similarly, the Role-Aware Transformer model struggled to match the performance of more traditional models, suggesting that additional optimization is required to improve its understanding of the role-specific structure in legal texts.

7.2 Analysis of Contextual Influence

The InLegalBERT models with varying sentence input configurations provide valuable insights into the impact of sentence context on performance. The model using only the current sentence shows moderate results, but performance improves as the context is expanded to include preceding sentences. This improvement demonstrates the importance of using sentence context to better understand the flow of legal documents and the transitions between rhetorical roles.

Incorporating predicted or true labels of previous sentences also has an influence on model performance. The model that uses the predicted label of the previous sentence shows lower performance, which may indicate that errors in label prediction propagate through subsequent sentences, reducing overall accuracy. This observation suggests that the quality of the label predictions for preceding sentences is critical for improving performance in real-world applications where true labels are not available.

7.3 Impact of Instruction-Tuning in `RhetoricLLaMA`

The instruction-tuned `RhetoricLLaMA` model, despite leveraging large language models, did not

achieve the expected performance. Instruction-tuning provides explicit guidance to the model, but for highly specialized tasks such as rhetorical role classification in legal documents, more domain-specific training and refined instruction sets may be required. The complexity of legal text segmentation and the need for deeper contextual understanding could explain the model's performance, highlighting the challenge of applying general-purpose LLMs to specialized domains without extensive adaptation.

7.4 Error Analysis

Our error analysis revealed that the models struggled primarily with distinguishing between closely related rhetorical roles, such as Facts and Reasoning, due to the overlap in their language and structure within legal documents. This challenge is clearly illustrated in the confusion matrix of the Hierarchical BiLSTM-CRF model Figure 2, which shows frequent misclassifications between these roles. Similarly, confusion between Arguments of Petitioner and Arguments of Respondent was prevalent, as both often exhibit similar language patterns, further complicating accurate classification. Models that incorporated contextual information from preceding or following sentences demonstrated some improvement in reducing these errors, particularly for roles requiring a clear transition, such as Issue and Decision. However, despite this improvement, the context-aware models still encountered difficulties, suggesting that the rhetorical role boundaries within these transitions are not always well-defined. Another critical issue identified was class imbalance. More frequent labels like None and Facts were consistently overpredicted, leading to lower precision for less frequent labels such as Issue and Decision. This imbalance skewed the performance, resulting in models favoring high-frequency roles at the expense of accuracy for underrepresented roles. Figures 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. illustrating the confusion matrices for other models, are provided in the Appendix due to space constraints. These figures further highlight the patterns of misclassification and the impact of various model architectures on error distribution. Addressing these issues, particularly through improved handling of context, mitigating class imbalance, and minimizing the propagation of sequential errors, remains a critical area for future research and model refinement.

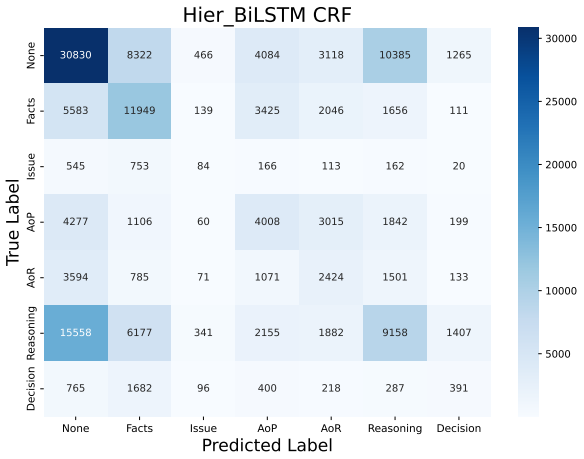

Figure 2: Confusion matrix for rhetorical role classification using Hierarchical BiLSTM-CRF model.

8 Conclusion and Future Work

In this work, we addressed the challenging task of rhetorical role classification in legal documents by introducing the `LegalSeg` dataset, the largest annotated dataset for this task. `LegalSeg`, provides a significant resource for advancing research in this domain. We evaluated multiple models, including `RhetoricLLaMA`, ToInLegalBERT, Role-aware, and GNNs. Our results show that models incorporating both sequential and contextual information, such as Hierarchical BiLSTM-CRF and ToInLegalBERT, perform best in identifying and classifying rhetorical roles in legal texts. We also demonstrated that adding sentence-level context improves the model's ability to capture transitions between rhetorical roles, reducing errors caused by the inherent similarity between roles like Facts and Reasoning.

Despite these advancements, our error analysis revealed several challenges, such as misclassification between similar roles and the cascading effect of label prediction errors.

For future work, we aim to explore more sophisticated techniques to handle class imbalance, such as advanced sampling strategies and loss function adjustments. Additionally, refining models' ability to capture long-range dependencies and leveraging more robust pre-training strategies could further enhance the performance of LLMs like `RhetoricLLaMA`. We also plan to incorporate more domain-specific knowledge into the models and experiment with cross-domain transfer learning to improve their adaptability across different legal contexts.

Limitations

While this study makes significant strides in rhetorical role classification for legal documents, a few areas remain where further refinement could enhance the approach. These areas are opportunities for future work rather than major limitations and are not expected to diminish the contribution of this research.

The `LegalSeg` dataset, while being the largest and most comprehensive of its kind for Indian legal judgments, is understandably specialized in the context of the Indian judiciary system. This focus provides unique insights into this particular legal domain. However, it is acknowledged that the models may require adaptation to handle legal documents from different jurisdictions. This does not limit the validity of our findings but opens a path for future research into cross-jurisdictional generalization using transfer learning techniques or domain adaptation strategies, which are common challenges in domain-specific NLP.

The class imbalance in the dataset, which is inherent in most real-world legal corpora, reflects the natural distribution of rhetorical roles in judgments. While some roles like Issue and Decision are less frequent, this mirrors their actual occurrence in legal texts. We have taken steps to mitigate this issue through advanced modeling techniques such as label shift prediction and the incorporation of contextual information. Future work could explore further enhancements, such as data augmentation or more refined class-weighting techniques, to boost performance on the less frequent roles.

Additionally, the computational requirements of models like ToInLegalBERT and `RhetoricLLaMA` are justified given the complexity and the high accuracy they provide. These models are aligned with state-of-the-art practices in NLP, which involve significant computational demands. While this may pose a challenge for deployment in low-resource environments, it is important to note that high-performance models are typically developed on powerful infrastructures and then optimized for more practical use cases through techniques such as model pruning, quantization, or distillation, which can be addressed in future work.

The overlap in rhetorical roles, such as between Facts and Reasoning, is an inherent challenge in legal discourse due to the intertwined nature of legal arguments and fact presentation. The models already handle these overlaps competently, and our use of sequential and contextual information improves performance. However, we recognize that future refinements, such as more sophisticated context-aware mechanisms or hybrid models that integrate symbolic reasoning with machine learning, could offer even greater differentiation between closely related roles.

In conclusion, the challenges discussed here are not insurmountable and represent common issues in the evolving field of legal NLP. This work provides a strong foundation for addressing these aspects, and we are confident that the solutions proposed will inspire future innovations and improvements. This manuscript significantly advances the state of the art in rhetorical role classification, and any remaining opportunities for refinement will only serve to further enhance the impact of this research.

Ethics Statement

This research was conducted with a strong commitment to ethical standards. The `LegalSeg` dataset comprises publicly available Indian legal judgments, ensuring no private or sensitive data was included. Anonymization was applied where necessary, and all data was collected in compliance with legal and privacy regulations.

In the process of annotating the dataset, law students from various institutions were involved. These annotators were treated with fairness and respect, and they were appropriately compensated for their contributions. Informed consent was obtained from all participants before their involvement in the project. While the students provided essential support in developing the dataset, they are not listed as co-authors of this manuscript to maintain the academic integrity of the publication. Instead, their valuable contributions are acknowledged separately.

The models developed are designed to support, not replace, legal professionals. We advocate for their responsible use, emphasizing the need for human oversight when applied in real-world legal contexts. This research adheres to ethical guidelines in authorship, data handling, and participant involvement, ensuring that all contributions are treated with fairness and respect.

References

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of rhetorical roles of sentences in indian legal judgments. In *Legal Knowledge and Information Systems*, pages 3–12. IOS Press.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13.

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022. Corpus for automatic structuring of legal documents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4420–4429, Marseille, France. European Language Resources Association.

Soumayan Bandhu Majumder and Dipankar Das. 2020. Rhetorical role labelling for legal judgements using roberta. In *FIRE (Working Notes)*, pages 22–25.

Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Angshuman Hazarika, Shubham Kumar Nigam, Arnab Bhattacharya, and Ashutosh Modi. 2022. Semantic segmentation of legal documents via rhetorical roles. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 153–171, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Gabriele Marino, Daniele Licari, Praveen Bushipaka, Giovanni Comandé, Tommaso Cucinotta, et al. 2023. Automatic rhetorical roles classification for legal documents using legal-transformeroverbert. In *CEUR WORKSHOP PROCEEDINGS*, volume 3441, pages 28–36. CEUR-WS.

Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik Kumar Guha, Sachin Malhan, and Vivek Raghavan. 2023. SemEval-2023 task 6: LegalEval - understanding legal texts. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2362–2374, Toronto, Canada. Association for Computational Linguistics.

Isar Nejadgholi, Renaud Bougueng, and Samuel Witherspoon. 2017. A semi-supervised training method for semantic search of legal facts in canadian immigration cases. In *Legal knowledge and information systems*, pages 125–134. IOS Press.

Shubham Nigam, Anurag Sharma, Danush Khanna, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2024. Legal judgment reimagined: PredEx and the rise of intelligent AI interpretation in Indian courts. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4296–4315, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Shubham Kumar Nigam, Aniket Deroy, Noel Shallum, Ayush Kumar Mishra, Anup Roy, Shubham Kumar Mishra, Arnab Bhattacharya, Saptarshi Ghosh, and Kripabandhu Ghosh. 2023a. Nonet at semeval-2023 task 6: Methodologies for legal evaluation. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1293–1303.

Shubham Kumar Nigam, Navansh Goel, and Arnab Bhattacharya. 2022. nigam@ coliee-22: Legal case retrieval and entailment using cascading of lexical and semantic-based models. In *JSAI International Symposium on Artificial Intelligence*, pages 96–108. Springer.

Shubham Kumar Nigam, Shubham Kumar Mishra, Ayush Kumar Mishra, Noel Shallum, and Arnab Bhattacharya. 2023b. Legal question-answering in the indian context: Efficacy, challenges, and potential of modern ai models. *arXiv preprint arXiv:2309.14735*.

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2023. Pre-trained language models for the legal domain: A case study on indian law. In *Proceedings of 19th International Conference on Artificial Intelligence and Law - ICAIL 2023*.

TYSS Santosh, Apolline Isaia, Shiyu Hong, and Matthias Grabmair. 2024. Hiculr: Hierarchical curriculum learning for rhetorical role labeling of legal documents. *arXiv preprint arXiv:2409.18647*.

M Saravanan, Balaraman Ravindran, and S Raman. 2008. Automatic identification of rhetorical roles using conditional random fields for legal document summarization. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Jaromír Šavelka and Kevin D Ashley. 2018. Segmenting us court decisions into functional and issue specific parts. In *Legal Knowledge and Information Systems*, pages 111–120. IOS Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shaurya Vats, Atharva Zope, Somsubhra De, Anurag Sharma, Upal Bhattacharya, Shubham Nigam, Shouvik Guha, Koustav Rudra, and Kripabandhu Ghosh. 2023. Llms–the good, the bad or the indispensable?:

A use case on legal statute prediction and legal judgment prediction on indian court cases. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12451–12474.

Giulia Venturi. 2012. Design and development of temis: a syntactically and semantically annotated corpus of italian legislative texts. In *proceedings of the workshop on semantic processing of legal texts (SPLeT 2012)*, pages 1–12. Citeseer.

Vern R Walker, Krishnan Pillaipakkamnatt, Alexandra M Davidson, Marysa Linares, and Domenick J Pesce. 2019. Automatic classification of rhetorical roles for sentences: Comparing rule-based scripts with machine learning. *ASAIL@ ICAIL*, 2385.

Adam Wyner, Wim Peters, and Daniel Katz. 2013. A case study on legal case annotation. In *Legal Knowledge and Information Systems*, pages 165–174. IOS Press.

Adam Z Wyner. 2010. Towards annotating and extracting textual legal case elements. *Informatica e Diritto: special issue on legal ontologies and artificial intelligent techniques*, 19(1-2):9–18.

A Experimental Setup and Hyper-parameters

We conducted experiments across several models, utilizing different architectures and training techniques tailored to rhetorical role classification tasks. Below, we provide an overview of the key experimental setups and hyper-parameters used.

A.1 `RhetoricLLaMA` Training Procedure

`RhetoricLLaMA`, built on the LLaMA-2-7B model, was fine-tuned with Bfloat16 precision using a single A100 GPU with 40GB memory. Given the computational constraints, the model was optimized for efficiency, with training lasting 48 hours. A maximum token length of 1000 was used, and Low-Rank Adaptation (LoRA) was employed with a rank of 16, alpha set to 64, and a dropout rate of 0.1. The model leveraged flash-attention 2 for faster training. We applied a Paged Adam optimizer with a learning rate of 1e-4 and a cosine learning rate scheduler, along with gradient accumulation steps of 4. The model trained for 52,617 steps, corresponding to 3 epochs.

A.2 Transformers Training Hyper-parameters

For the Role-Aware Transformers, built upon the InLegalBERT model, pre-training involved self-supervised tasks such as Masked Language Modeling (MLM) with role embeddings added. The model processed a maximum sequence length of 512 tokens with a batch size of 4, running for 20 epochs. The learning rate was set to 2e-5, using the AdamW optimizer. Class weights were applied to handle the imbalance in rhetorical roles, and early stopping was used to prevent overfitting.

A.3 Graph Neural Networks (GNN) with Document Context

We utilized a Graph Neural Network (GNN) architecture to model sentence relationships within legal documents. A two-layer Graph Convolutional Network (GCN) processed sentence embeddings from InLegalBERT. The first and second GCN layers both had output dimensions of 128, using ReLU activations. The model was trained for 10 epochs with a learning rate of 1e-4 and employed a Cross-Entropy Loss function. Graphs were constructed with edges between consecutive sentences, capturing both sequential and semantic relationships.

A.4 Incorporating Previous Sentence and Actual Label

In this method, the input to the model combined the current sentence with the previous sentence and its actual rhetorical role label. The model used InLegalBERT with a maximum sequence length of 512 tokens, trained for 5 epochs with a learning rate of 2e-5. This approach provided explicit sequential context and utilized Cross-Entropy Loss with class weights to manage class imbalance.

A.5 Incorporating Previous Sentence and Predicted Label

Extending the previous approach, this variant incorporated the predicted label of the previous sentence, simulating real-world conditions. The same configuration was used as in the previous model, but the predicted label replaced the actual label during both training and inference.

A.6 Common Settings Across Models

All models were evaluated using Accuracy, Precision, Recall, F1 Score, and Matthews Correlation Coefficient (MCC). The experiments utilized PyTorch and Hugging Face Transformers libraries, with PyTorch Geometric handling the graph data in the GNN method. All models were trained on machines with NVIDIA GPUs for parallel computation.

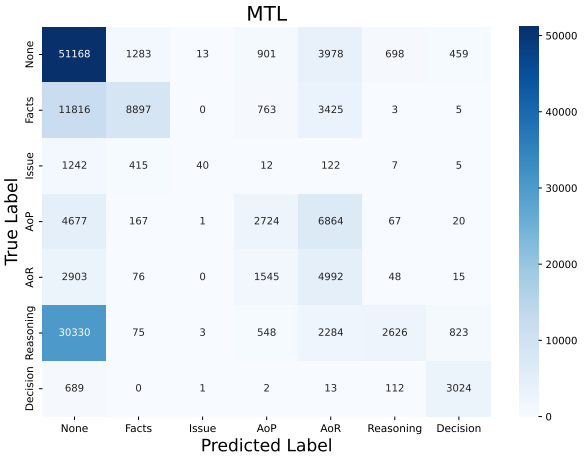

Figure 3: Confusion matrix for rhetorical role classification using the Multi-Task Learning (MTL) model.

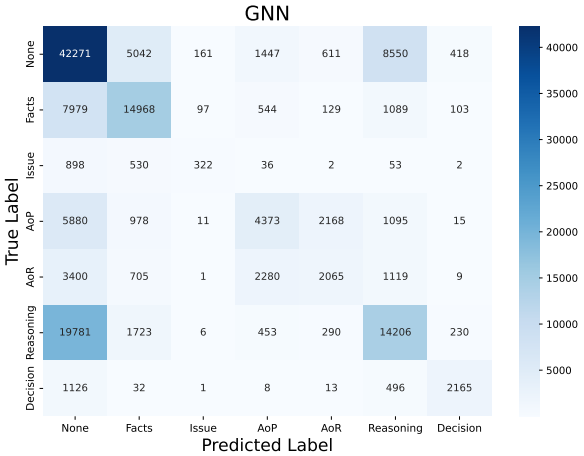

Figure 4: Confusion matrix for rhetorical role classification using GNN.

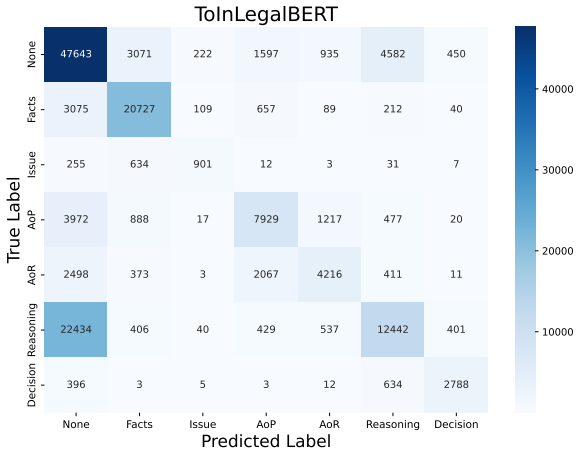

Figure 5: Confusion matrix for rhetorical role classification using TransformerOverInLegalBERT (ToInLegalBERT).

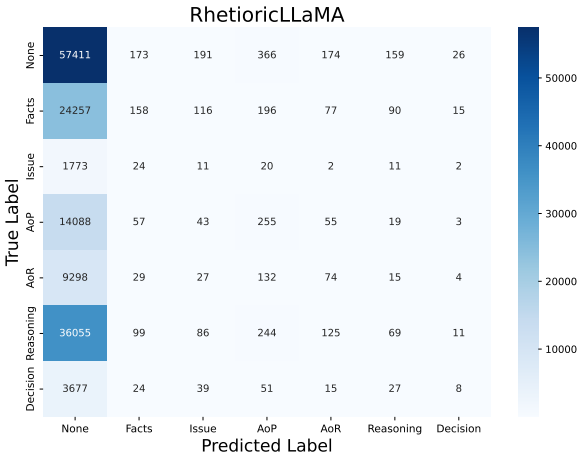

Figure 6: Confusion matrix for rhetorical role classification using `RhetoricLLaMA`, an instruction-tuned large language model.

Label	Sentences
Fact	For the sake of convenience, we are referring to the facts of Civil Appeal No.1328 of 2021.
Fact	At the time of the assessment proceedings, the Assessee submitted a revised computation of income by revising its claim of deduction under Section 80IA of the Act.
Issue	The Income Tax Appellate Tribunal (hereinafter the Tribunal), upheld the decision of the Appellate Authority on the issue of deduction under Section 80IA.
Issue	The High Court refused to interfere with the Tribunals order as far as the issue on deduction under Section 80IA is concerned.
Arguments of Petitioner (AoP)	Mr. Arijit Prasad, learned Senior Counsel appearing on behalf of the Revenue, submitted that Section 80AB of the Act contemplates deductions in respect of incomes against income of the nature specified in the relevant section.
Arguments of Petitioner (AoP)	According to him, the phrase derived from in subsection (1) of Section 80IA of the Act indicates that the computation of deduction is restricted only to the profits and gains from the eligible business.
Arguments of Respondent (AoR)	In response, the Assessee supported the order passed by the Appellate Authority which was upheld by the Tribunal and the High Court.
Arguments of Respondent (AoR)	He submitted that there is no indication in subsection (5) of Section 80IA that the deduction under subsection (1) is restricted to business income only.
Reasoning	As stated above, Section 80AB was inserted in the year 1981 to get over a judgment of this Court in Cloth Traders (P) Ltd. (supra).
Reasoning	On the question of existence of vacancies, although learned counsel for the appellant submitted that vacancies are still lying there, which submission however has been refuted by the learned counsel for the State of Rajasthan.
Decision	For the aforementioned reasons, the Appeal is dismissed qua the issue of the extent of deduction under Section 80IA of the Act.
Decision	The assets of the Corporate Debtor shall be managed strictly in terms of the provisions of the IBC.
None	Clause 11(b) reads as follows 11.
None	The clause reads thus 12 Miscellaneous.

Table 4: Example sentences for each label.

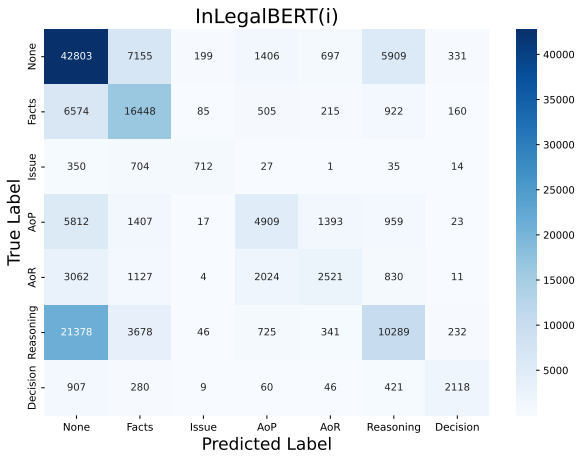

Figure 7: Confusion matrix for rhetorical role classification using InLegalBERT model with the current sentence (i) as input.

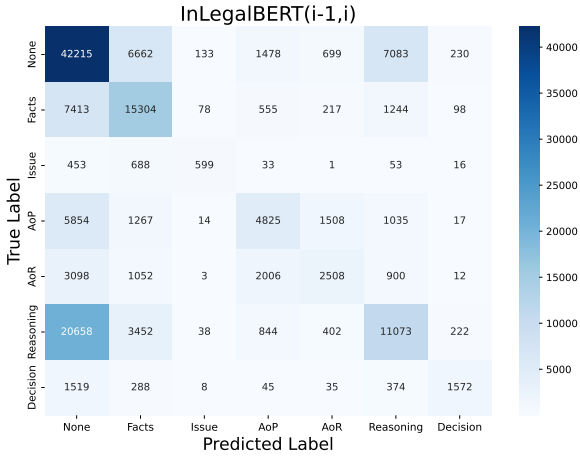

Figure 8: Confusion matrix for rhetorical role classification using InLegalBERT model with the current sentence (i) and the previous sentence (i-1) as input.

	Instruction Sets
1	Analyze the given legal sentence and predict its rhetorical role as a number: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
2	Determine the rhetorical function of this sentence from a court case and provide its corresponding number: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
3	Based on the content of the following legal text, classify its rhetorical role by selecting the appropriate number: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
4	Identify the rhetorical category of this legal statement and provide the number that represents it: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
5	Evaluate the rhetorical purpose of the provided legal sentence and label it with the correct number: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
6	Assign a number to the rhetorical role of this sentence from a legal case, choosing from: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
7	Review the legal statement and predict its rhetorical function using the corresponding number: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
8	Examine this legal text and determine its rhetorical role by outputting the appropriate number: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
9	Categorize the rhetorical purpose of the following sentence from a court proceeding with a number: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
10	Analyze the provided legal sentence and classify it into its rhetorical role, outputting only the number: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
11	Determine the appropriate number for the rhetorical category of this legal text: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
12	Assign a numerical label to the rhetorical role of this statement in a legal case: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
13	Predict the number that corresponds to the rhetorical function of the following legal sentence: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
14	Identify the number that represents the rhetorical role of this legal text: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
15	Analyze this legal statement and assign the number that best matches its rhetorical function: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.
16	Classify the following sentence from a court case by selecting its rhetorical role number: None-0, Facts-1, Issue-2, Arguments of Petitioner-3, Arguments of Respondent-4, Reasoning-5, Decision-6.

Table 5: Instruction Sets for Predicting the Roles

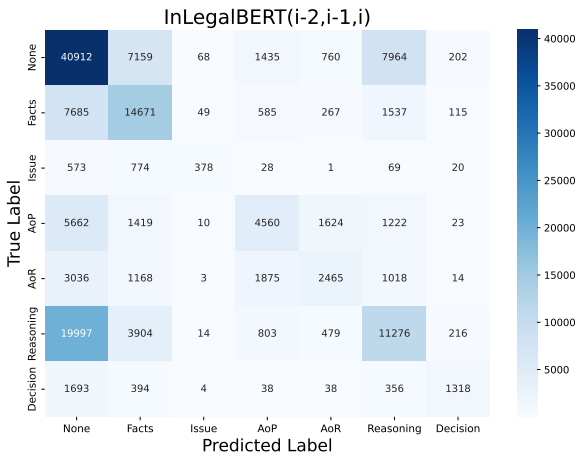

Figure 9: Confusion matrix for rhetorical role classification using InLegalBERT model with the previous-to-previous sentence (i-2), previous sentence (i-1), and the current sentence (i) as input.

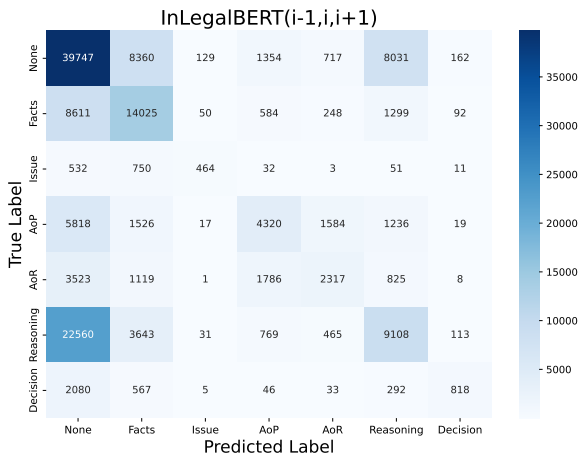

Figure 10: Confusion matrix for rhetorical role classification using InLegalBERT model with the current sentence (i), previous sentence (i-1), and next sentence (i+1) as input.

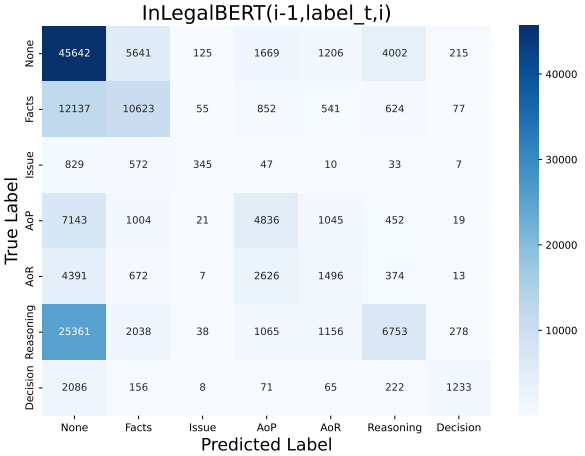

Figure 11: Confusion matrix for rhetorical role classification using InLegalBERT model with the true label of the previous sentence (i-1) and the current sentence (i) as input.

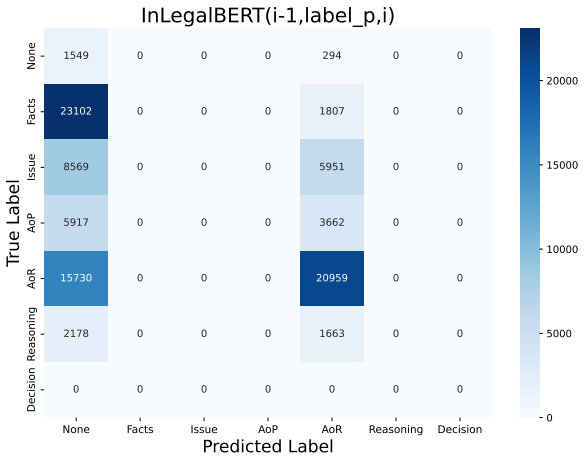

Figure 12: Confusion matrix for rhetorical role classification using InLegalBERT model with predicted label of the previous sentence (i-1) and the current sentence (i) as input.

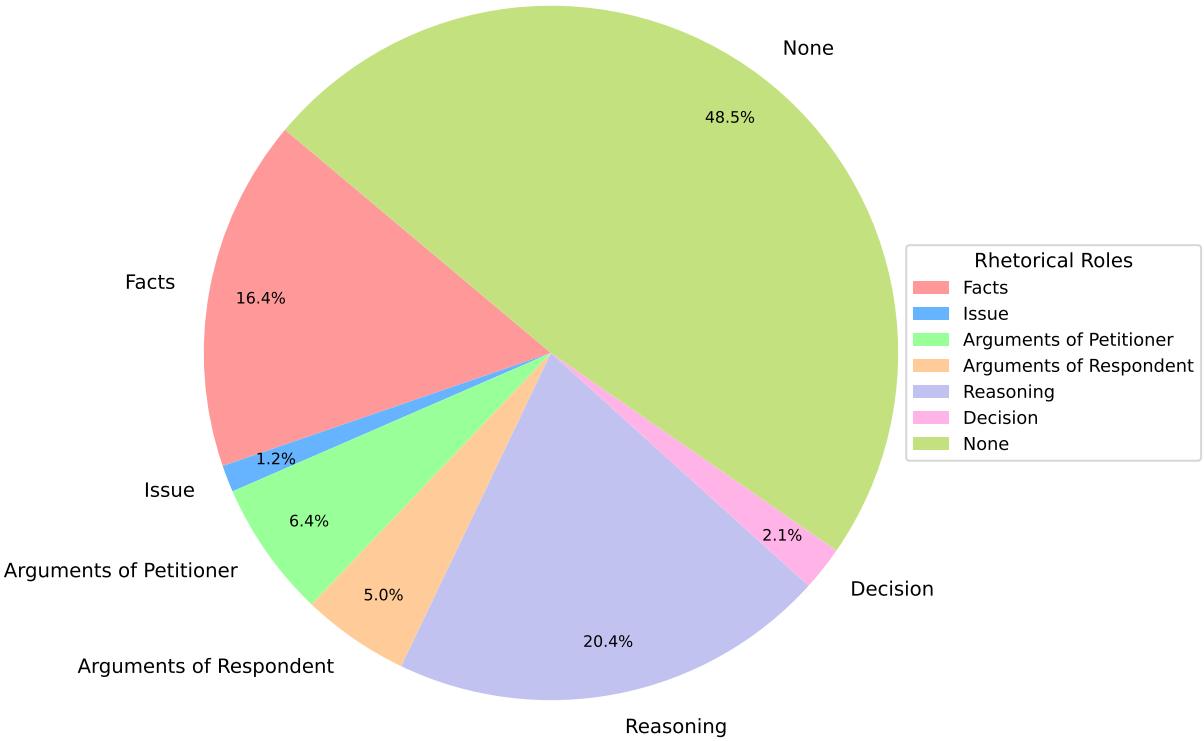

Figure 13: Distribution of Rhetorical Roles within the Dataset.