SeqTNS: Sequential Tolerance-based Classifier for Identification of Rhetorical Roles in Indian Legal Documents

Arjun T D[1], **Anand Kumar Madasamy**[1], **Sheela Ramanna**[2]
[1]Department of Information Technology
National Institute of Technology Karnataka (NITK), Surathkal, India
[2]Department of Applied Computer Science, University of Winnipeg, Canada
{arjuntd.243it001, m_anandkumar}@nitk.edu.in, s.ramanna@uwinnipeg.ca

Abstract

Identifying rhetorical roles in legal judgments is a foundational step for automating legal reasoning, summarization, and retrieval. In this paper, we propose a novel Sequential Tolerance-based Classifier (SeqTNS) for rhetorical role classification in Indian legal documents. The proposed classifier leverages semantic similarity and contextual dependencies by using label sequence aware BiLSTMs on top of word embeddings from finetuned InLegalBERT model. These enriched embeddings are clustered into tolerance classes via a tolerance relation using a cosine distance threshold, enabling the model to make flexible, similarity-based predictions. We evaluate SeqTNS on two benchmark datasets annotated with thirteen and seven rhetorical roles, respectively. The proposed method outperforms fine-tuned transformer baselines (LegalBERT, InLegalBERT) as well as the previously developed tolerance relation-based (TNS) model, achieving a weighted F1 score of 0.78 on thirteen class dataset and a macro F1 of 0.83 on the seven class dataset, while reducing training time by 39-40% compared to state of the art BiLSTM-CRF models. The larger of our two datasets is substantial, containing over 40,000 sentences and 1.3M tokens, and serves as a challenging real world benchmark. Additionally, we use LIME for explainability and t-SNE to validate the coherence of tolerance-based clusters.

1 Introduction

The backlog of legal cases in India over 52 million as of 2025 underscores the urgent need for automated tools to process and understand legal documents. Legal texts are lengthy, unstructured and contain many domain-specific jargons, making traditional NLP approaches inadequate. A promising direction in legal NLP is the segmentation of legal documents into semantically meaningful units called *Rhetorical Roles* (RRs), such as facts, issues, arguments, statutes, and judgments. This segmentation facilitates downstream applications including legal summarization, information retrieval, and judgment prediction.

Early research in this area focused on identifying rhetorical roles through manually created features. Saravanan et al. (Saravanan et al., 2008) were the first to take on this task using Conditional Random Fields (CRFs) along with crafted lexical and structural cues. They segmented Indian legal texts into seven rhetorical roles. While these methods worked well, they needed rules made by experts and did not apply broadly to different legal sub-domains. Bhattacharya et al. (Bhattacharya et al., 2019) created a corpus of 50 Supreme Court rulings that were annotated with seven rhetorical roles in order to overcome these constraints. They also suggested deep neural architectures, such as BiLSTM-CRF models,introduced by (Ma and Hovy, 2016) to automate role identification. This line of research, modeling rhetorical role identification as a sequence tagging problem, was further advanced by (Dutta, 2021), who achieved first place in the AILA-2021 shared task by combining embeddings from a domain-specific language model with a GRU-CRF sequence tagging classifier. The inherent subjectivity in annotating legal texts was brought to light by their work, which also introduced a robust analysis of inter-annotator agreement. This signaled a shift toward deep learning-based systems for feature extraction. The survey from (He et al., 2020) provides a scientific taxonomy of deep learning models for sequence labeling, systematically analyzing approaches based on their core components: the embedding, context encoder, and inference modules. In more recent attempts both the size of the dataset and the sophistication of the models have increased. Malik et al. (Malik et al., 2022) presented a corpus of 100 documents from the legal sub-domains of competition law and income tax, annotated with thirteen intricate rhetorical roles.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 837–847
December 20-24, 2025 ©2025 Association for Computational Linguistics

To capture topical coherence across sentences, they put forth a Multi-Task Learning (MTL) framework that combines auxiliary label shift detection with rhetorical role prediction.

With 354 Indian judgments annotated at the sentence level with thirteen rhetorical roles, Kalamkar et al. (Kalamkar et al., 2022) provided the largest dataset to date with rhetorical role labels. Their research showed how legal summarization and judgment prediction can be improved through rhetorical segmentation. The corpus provides a representative standard for legal NLP in India and includes rulings from the Supreme Court, High Courts, and district courts. Legal NLP has been transformed by recent advancements in domain-specific pre-trained language models. LegalBERT, which was developed using European legal corpora and showed excellent performance on a variety of downstream legal tasks, was introduced by Chalkidis et al. (Chalkidis et al., 2020). However, due to the significant differences in language structure and legal semantics, these models are not optimized for Indian legal texts. Paul et al. (Paul et al., 2023) published InLegalBERT, which was pre-trained on more than 50,000 Indian legal documents, to fill this gap. In order to improve contextual understanding and rhetorical role prediction, we use both LegalBERT and InLegalBERT in our research to embed legal sentences before classification.

The proposed method builds on previous works on TNS (tolerance near sets) models in classification tasks specifically in NLP where neural embeddings were used for short-text classification (Patel et al., 2022), later optimized and evaluated with different sentence embedding strategies (Hegde et al., 2023), and extended to multimodal classification task by combining textual and visual embeddings (Kelkar et al., 2024). The formal foundations for the original TNS model are tolerance relations (Zeeman, 1962), tolerance spaces as a framework for resemblance or similarity (Sossinsky, 1986) and descriptively near sets (Peters, 2007; Peters and Naimpally, 2012). The tolerance relation forms the basis for clustering similar samples into tolerance classes. Since elements of a tolerance class are considered "equivalent", it is then sufficient to choose a prototype element to represent the entire tolerance class. The TNS model uses both unsupervised (clustering) and supervised learning strategies (with labelled prototypes).

In traditional prototype-based classification methods such as *Prototypical Networks* (Snell et al., 2017), the support set is a fixed, pre-determined subset of labeled data from which class prototypes are computed. Each prototype is obtained as the mean of the embeddings of its support examples, and classification is performed by comparing query embeddings to these fixed prototypes using a distance metric (e.g., Euclidean or cosine distance).

In contrast, our **tolerance-based formulation** does not rely on such a pre-defined support set. Instead, tolerance relations dynamically form *overlapping groups* of instances under a similarity threshold ε, from which representative (prototype) embeddings are derived. These tolerance classes evolve based on the intrinsic structure of the embedding space rather than pre-specified class partitions, allowing instances to belong to multiple overlapping classes when semantically justified. This results in a threshold-driven, similarity-based, and interpretable clustering mechanism that is particularly suited for textual data exhibiting semantic overlap.

Furthermore, unlike traditional end-to-end sequence labeling approaches (e.g., BiLSTM-CRF or Transformer-CRF), which employ a parametric softmax or CRF classification head, our model replaces the classifier with a **tolerance-based decision process** that integrates sequential context. In this setup, the final classification depends on proximity to representative tolerance embeddings rather than on direct parameter optimization, providing both computational efficiency and robustness to label ambiguity.

Building on this foundation, in this paper, we introduce SeqTNS, a Sequential Tolerance Near Sets Classifier by combining the current sentence embedding with the embeddings of the previous and next predicted labels. Our sequential approach is motivated by trends in the broader legal AI domain, where tasks like legal judgment prediction are also modeled as sequential processes to capture dependencies between sub-tasks (Shang, 2022). Furthermore, the strategy of augmenting a sequence tagger with document-level context has proven effective for related tasks like catchphrase extraction from Indian legal texts, underscoring the importance of document-specific information (Mandal et al., 2022). Where the original TNS model (Patel et al., 2022), simply forms tolerance classes based on non sequential neural embeddings, the new SeqTNS is designed for capturing rhetorical flow in legal discourse and particularly suited for rhetorical role classification task. The sentence embeddings used

for SeqTNS are derived from a BiLSTM-based label modeling architecture illustrated in Figure 1. The model uses token-level representations from a fine-tuned InLegalBERT encoder and captures both contextual level and document level dependencies. The final embedding for each sentence is formed by concatenating its contextual BiLSTM output with predicted label embeddings from its surrounding sentences. Our experiments on two benchmark datasets demonstrate that SeqTNS outperforms fine-tuned transformer baselines (LegalBERT, InLegalBERT) as well as the original TNS model. Additionally, our proposed SeqTNS model outperforms the state of the art BiLSTM-CRF architecture on the seven class rhetorical role dataset introduced by Bhattacharya et al. (2019), achieving a macro F1 score of 0.83 compared to 0.82. This result not only demonstrates improved performance but also does so with 39% lower training time, underscoring the efficiency and scalability of SeqTNS for legal discourse analysis.

Our primary contribution is its novel adaptation and extension to a sequential labeling task within the legal domain. To our knowledge, this is the first work to apply and extend TNS for rhetorical role classification. Our innovation lies in integrating sequential context via neighboring label embeddings into the classification mechanism (SeqTNS).

This paper is organized as follows: Section 2 discusses materials and formal definitions for tolerance classes used in building SeqTNS. Section 3 presents the computational environment, results and discusses the findings followed by Section 4, which provides a conclusion of our study along with the future work and limitations.

2 Materials and Methods

2.1 Materials

Two publicly accessible Indian legal datasets are used in our research. The first dataset, created by Bhattacharya et al. (Bhattacharya et al., 2019), includes 50 Supreme Court rulings that are categorized using seven rhetorical roles such as Facts, Issues, Arguments, Precedent, Statute, Decision Ratio and Rulings over five different legal domains (e.g., criminal, consitutional, land and property) with 10 documents per domain for a total number of 9,380 sentences and 323,869 tokens. Presented by Kalamkar et al. (Kalamkar et al., 2022), the second, much larger dataset consists of 354 judgments from different levels of Indian courts that are annotated at the sentence level (40,315 sentences and more than 1.3M tokens) with thirteen rhetorical roles (12 + NONE). This dataset provides better generalizability across legal sub-domains and captures a greater variety of legal discourse structures.

2.2 Methods

Preliminaries We recall some useful definitions related to tolerance relations, and tolerance classes.
Definition [Tolerance Relation (Zeeman, 1962)]: Let O be a set of sample objects, and let τ be a binary relation (called a tolerance relation) on O ($\tau \subseteq O \times O$) that is reflexive (for all $x \in O$, $x\tau x$) and symmetric (for all $x, y \in O$, if $x\tau y$, then $y\tau x$) but transitivity of τ is not required. Based on E.C. Zeeman (Zeeman, 1962), every pseudometric space$\langle O, p\rangle$ determines tolerance relations with respect to some positive real threshold $\varepsilon(0, +\infty)$.

Definition [PreClass and Tolerance Classes (Wasilewski et al., 2011)]: A set $A \subseteq O$ is a *τ-preclass* (or briefly *preclass* when τ is understood) if and only if for any $x, y \in A$, $(x, y) \in \tau$. The family of all preclasses of a tolerance space is naturally ordered by set inclusion and preclasses that are maximal with respect to a set inclusion are called tolerance classes.

This Zeeman tolerance definition is the basis for a feature-based tolerance relations and tolerance near sets (Peters, 2009) that induces tolerance classes using a distance measure. This reflexive and symmetric relation induces a set of *tolerance classes*, where each class contains instances that are pairwise similar within the chosen threshold. Additional instances can belong to more than one tolerance class which can helpful with ambiguous label boundaries.

Definition [Textual feature-based tolerance relation (Patel et al., 2022; Kelkar et al., 2024)]: A tolerance space $\langle O, \tau_\varepsilon, \mathcal{T}\rangle$ is defined with a binary relation τ_ε where O is the set of objects and $\mathcal{T} = \{\phi(t_i) \mid t_i \in O\}$ represents sentence embeddings obtained from a transformer-based encoder.

Then tolerance classes are formed using the following equation:

$$\tau_{\mathcal{T},\varepsilon} = \{(\phi(t_i), \phi(t_j)) \in \mathcal{T} \times \mathcal{T} \mid \text{dist}(\phi(t_i), \phi(t_j)) \leq \varepsilon\} \quad (1)$$

where ε is a user-defined tolerance threshold, and $\text{dist}(\phi(t_i), \phi(t_j)) = 1 - \frac{\phi(t_i)\cdot\phi(t_j)}{\|\phi(t_i)\|\|\phi(t_j)\|}$ is the cosine distance between the sentence embeddings.

Algorithm 1 Sequential Tolerance Near Set Classifier (SeqTNS) for Rhetorical Role Classification

Input: Labeled training set $\mathcal{D}_{train}$, test set $\mathcal{D}_{test}$, threshold ε
Output: Predicted labels for test instances (y_i)
Compute sentence embeddings $\phi(t_i)$ for each $t_i \in \mathcal{D}_{train}$ using BiLSTM with label-sequence architecture shown in Figure 1
Construct cosine distance matrix $\text{dist}(\phi(t_i), \phi(t_j))$ between all training instances
Form tolerance classes $\mathcal{TC}_k$ using threshold ε:
    $\mathcal{TC}_k = \{t_i \mid \text{dist}(\phi(t_i), \phi(t_j)) \leq \varepsilon\}$
Compute prototype vector p_k for each tolerance class:
    $p_k = \frac{1}{|\mathcal{TC}_k|} \sum_{t_i \in \mathcal{TC}_k} \phi(t_i)$
Assign label y_k to each prototype p_k using majority voting over class members
for each test instance $t_{\text{test}} \in \mathcal{D}_{test}$ **do**
    Compute sentence embedding $\phi(t_{\text{test}})$ using Figure 1 architecture, incorporating previous predicted label and masking next label
    Compute cosine distances to all prototypes
    Identify the closest prototype p^* with minimum distance
    Assign label of p^* to t_{test}
end for
Compare predicted vs. original labels to evaluate accuracy

To obtain the feature representations $\mathcal{T}$ used in our SeqTNS framework, we leverage two transformer-based models: LegalBERT (Chalkidis et al., 2020) and InLegalBERT (Paul et al., 2023). These models, pretrained on large corpora of legal texts, are further fine-tuned on our downstream rhetorical role classification task. The resulting sentence embeddings are then directly used within the TNS (non sequential) framework (Patel et al., 2022) to form tolerance classes. The TNS model serves as one of our baseline models without the sequential and discourse level context.

Inspired by the nature of legal documents, where the rhetorical role of a sentence is often dependent on its neighboring context, we construct sentence representations that incorporate embeddings of adjacent predicted labels.

Suppose we are the predicting label for a sentence and we know that the label for both its previous and next sentence is PREAMBLE then there is a high probability that this sentence is of the same label.

The architecture of the SeqTNS framework, illustrated in Figure 1, is designed to capture both sentence-level semantics and document-level rhetorical context. This architecture also aligns directly with the steps of Algorithm 1, and we now explain the key notations used throughout.

Let $W = \{w_{1,i}, w_{2,i}, \ldots, w_{m,i}\}$ denote the sequence of m tokenized words in the i-th sentence of a legal document. Each sentence is first passed through a fine-tuned InLegalBERT model to obtain contextualized word embeddings. These embeddings are then input into a sentence-level BiLSTM, followed by an attention pooling layer that aggregates them into a fixed-size vector s_i, representing the sentence embedding. Let $S = \{s_1, s_2, \ldots, s_n\}$ be the set of such sentence embeddings for a document with n sentences.

To capture document-level context, these sentence embeddings are further passed into a second BiLSTM that operates at the sentence level, producing context-enriched vectors $C = \{c_1, c_2, \ldots, c_n\}$, where c_i encodes both the content of sentence i and its relation to other sentences in the document.

Since the rhetorical role of a sentence is often influenced by the roles of its neighboring sentences, we incorporate label embeddings. For each sentence i, let l_{i-1} and l_{i+1} denote the embeddings of the rhetorical roles of the previous and next sentences, respectively. These are drawn from a label embedding matrix $L = \{l_1, \ldots, l_k\}$, where k is the total number of training samples.

The final embedding used for classification is constructed by concatenating c_i with l_{i-1} and l_{i+1}:

$$\phi(t_i) = l_{i-1} \oplus c_i \oplus l_{i+1} \tag{2}$$

The set of all such vectors $\phi(t_i)$ for training instances constitutes the feature space $\mathcal{T}$ described in the tolerance relation definition. This enriched vector $\phi(t_i)$ corresponds to the input representation used in the SeqTNS algorithm (Algorithm 1) to compute cosine distances and form tolerance classes.

In the final stage of the architecture, each $\phi(t_i)$ is passed through a multi-layer perceptron (MLP) trained using a cross-entropy loss to predict rhetorical roles. This MLP enables the network to learn discriminative features during training. Once trained, the sentence representations $\phi(t_i)$ extracted from the final hidden layer are used as fixed embeddings for tolerance based classification. Specifically, prototype vectors for each

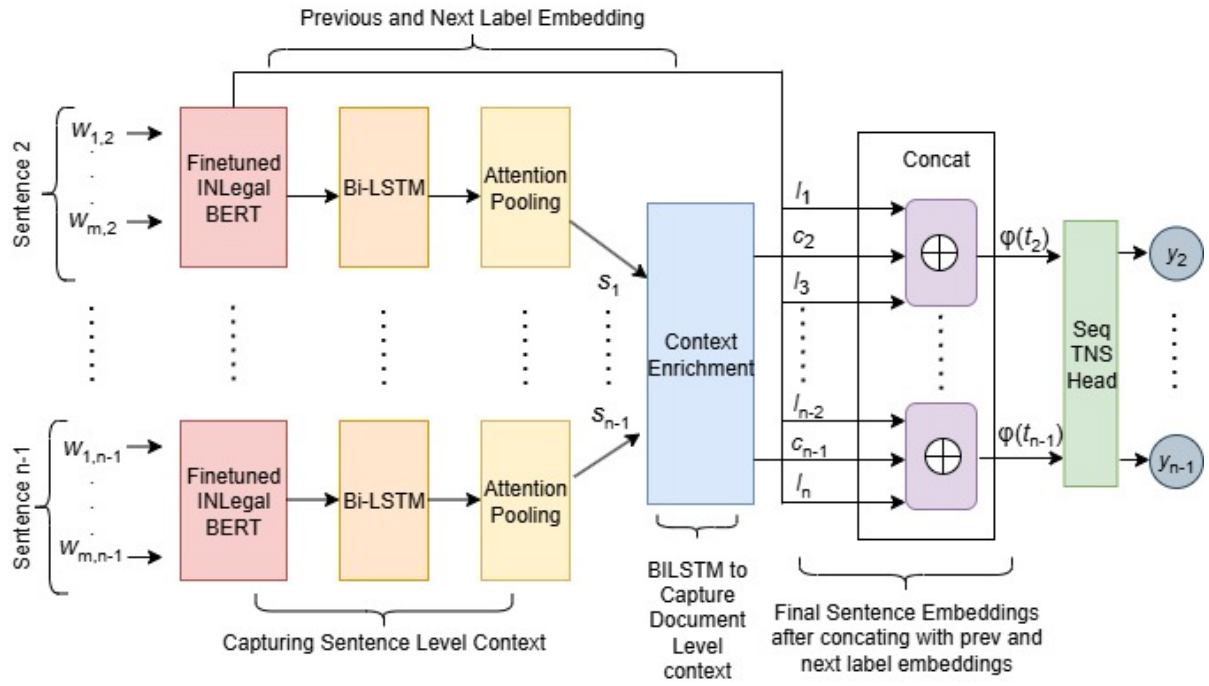

Figure 1: Architecture of the proposed SeqTNS model, where sentence embeddings are derived from a sequential InLegalBERT–BiLSTM model trained with an MLP head, and classification is performed using a tolerance-based (SeqTNS) head.

tolerance class are constructed by averaging the $\phi(t_i)$ embeddings of all its members, and test instances are classified based on proximity to the nearest prototype in this embedding space.

By incorporating both intra-sentence semantics and inter-sentence discourse-level context, as well as the influence of neighboring rhetorical roles, this architecture effectively supports tolerance-based classification for legal discourse analysis.

3 Results and Discussions

3.1 Computational Environment

All experiments were conducted using a virtual machine equipped with an Intel(R) Xeon(R) CPU @ 2.00GHz and an NVIDIA Tesla P100 GPU with 16GB VRAM. The models were implemented in Python using the PyTorch deep learning framework, along with the Hugging Face Transformers library for BERT-based architectures. All the experiments being reported in the paper including the comparative studies were done by us, in this computational setup.

3.2 Results

We tested all the baselines and the proposed model on the seven rhetorical role dataset introduced by Bhattacharya et al (Bhattacharya et al., 2019). As shown in Table 1, the proposed model achieves an overall accuracy of 0.84 and a macro F1 score of 0.83, with a 14.4% increase in F1 score compared to the best performing baseline with non sequential TNS.

Table 1: Performance Comparison on Seven Rhetorical Roles from (Bhattacharya et al., 2019). TNS refers to the previous non-sequential model.

Model	Test Accuracy	Test F1	Training Time
LegalBERT (Fine-tuned)	0.69	0.69	10 minutes
InLegalBERT (Fine-tuned)	0.71	0.70	10 minutes
LegalBERT + TNS	0.65	0.65	No training
Fine-tuned LegalBERT + TNS	0.70	0.68	10 minutes
InLegalBERT + TNS	0.68	0.67	No training
Fine-tuned InLegalBERT + TNS	0.72	0.71	10 minutes
BiLSTM + CRF	**0.83**	**0.82**	33 minutes
BiLSTM + Label Modelling + SeqTNS	**0.84**	**0.83**	20 minutes

This surpasses the best reported result in the original paper, where the state of the art BiLSTM-CRF model attained a macro F1 of 0.82. In addition to this performance gain, The proposed model also achieves 39% lower training time.

Table 2 shows the class-wise performance of the proposed model of the seven rhetorical role dataset. Macro F1 score of 0.83 indicates that the model performs well across all seven classes.

To further evaluate generalizability, we tested all the baselines and the proposed model on the thirteen class rhetorical role dataset introduced

Table 2: Classification Results of the Proposed SeqTNS Model on Seven Rhetorical Role Dataset from (Bhattacharya et al., 2019).

Class	Precision	Recall	F1 Score
Facts	0.82	0.87	0.84
Argument	0.69	0.84	0.76
Ratio of the Decision	0.90	0.85	0.88
Ruling by Lower Court	1.00	0.83	0.91
Ruling by Present Court	1.00	0.79	0.88
Precedent	0.94	0.83	0.88
Statute	0.57	0.82	0.67
Macro Avg.	0.85	0.83	0.83
Weighted Avg.	0.86	0.84	0.85

Table 3: Performance Comparison on Thirteen Rhetorical Roles (Test Set) from (Kalamkar et al., 2022). TNS refers to the previous non-sequential model.

Model	Test Accuracy	Test F1	Training Time
LegalBERT (Fine-tuned)	0.65	0.66	26 minutes
InLegalBERT (Fine-tuned)	0.67	0.67	26 minutes
LegalBERT + TNS	0.55	0.52	No training
Fine-tuned LegalBERT + TNS	0.69	0.69	26 minutes
InLegalBERT + TNS	0.63	0.61	No training
Fine-tuned InLegalBERT + TNS	0.70	0.70	26 minutes
BiLSTM + CRF	**0.79**	**0.79**	4h 20m
BiLSTM + Label Modelling + SeqTNS	**0.78**	**0.78**	2h 35m

by Kalamkar et al. (Kalamkar et al., 2022). The proposed BiLSTM with label sequence modelling combined with SeqTNS achieves 0.78 accuracy and 0.78 weighted F1 score, outperforming all other baselines. Compared to the best performing fine-tuned InLegalBERT+TNS setup, our model yields a relative F1 improvement of over 11%. Notably, SeqTNS achieves 1% less F1-score than state of the art BiLSTM-CRF model from the original dataset paper (Kalamkar et al., 2022), while reducing training time by 40%, demonstrating its efficiency and scalability across datasets with different rhetorical taxonomies.

Table 4: Classification Results of the Proposed SeqTNS Model on Thirteen Rhetorical Role Dataset from (Kalamkar et al., 2022).

Class	Precision	Recall	F1-Score	Support
ANALYSIS	0.80	0.88	0.84	1256
ARG_PETITIONER	0.65	0.66	0.65	253
ARG_RESPONDENT	0.42	0.71	0.53	82
FAC	0.83	0.77	0.80	933
ISSUE	0.95	0.84	0.89	43
NONE	0.89	0.76	0.82	250
PREAMBLE	0.96	0.76	0.85	723
PRE_NOT_RELIED	0.00	0.00	0.00	1
PRE_RELIED	0.78	0.57	0.66	141
RATIO	0.45	0.59	0.51	131
RLC	0.43	0.70	0.54	87
RPC	0.84	0.83	0.83	197
STA	0.51	0.75	0.61	61
Weighted Avg.	**0.80**	**0.78**	**0.78**	**4158**

Table 4 presents a detailed class-wise performance breakdown of the proposed model. SeqTNS classifier demonstrates strong generalization on dominant classes like ANALYSIS, FAC, PREAMBLE, and NONE, while maintaining reasonable performance across minority classes. As expected, the PRE_NOT_RELIED class, with only a single example, is not predicted correctly, a limitation inherent in extreme class imbalance.

Tolerance threshold (ε) is a critical hyperparameter in the SeqTNS classifier as it governs the size and granularity of the formed tolerance classes. Figure 2 shows the variation of both micro and weighted F1 scores with respect to ϵ across both the datasets.

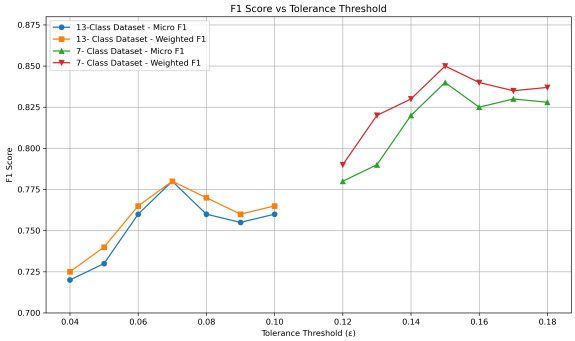

Figure 2: Effect of Tolerance Threshold (ε) on F1 Score for both datasets. 0.4 <= ε <= 0.10 for (Bhattacharya et al., 2019) dataset. 0.12 <= ε <= 0.18 for (Kalamkar et al., 2022) dataset.

For both datasets, performance improves with increasing ε up to a point, indicating that more inclusive similarity neighborhoods help generalization. However, very large values of ε can lead to overly broad tolerance classes, causing dissimilar instances to be grouped together, which may degrade precision. The best performance is observed at $\varepsilon = 0.07$ for the seven class dataset and $\varepsilon = 0.15$ for the thirteen class dataset.

To further investigate how well tolerance classes group semantically and contextually similar sentences, we visualize the class-wise distribution of feature vectors using t-SNE in the Discussion section. This helps illustrate the clustering quality under the optimal ε values.

3.3 Discussions

To better understand how the proposed SeqTNS framework organizes rhetorical role embeddings in the vector space, we visualize the class-wise distribution of training instances using t-SNE. Figures 3 and 4 show the 3D t-SNE projections of the prototype vectors for the thirteen class and seven class

datasets, respectively.

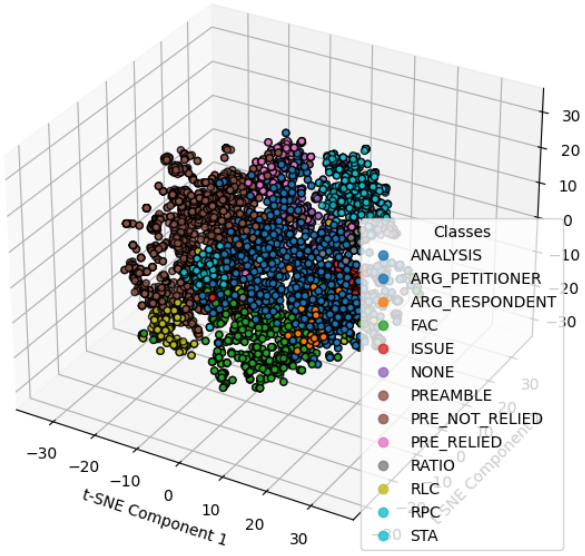

Figure 3: t-SNE visualization of prototype embeddings for thirteen class rhetorical role dataset from (Kalamkar et al., 2022).

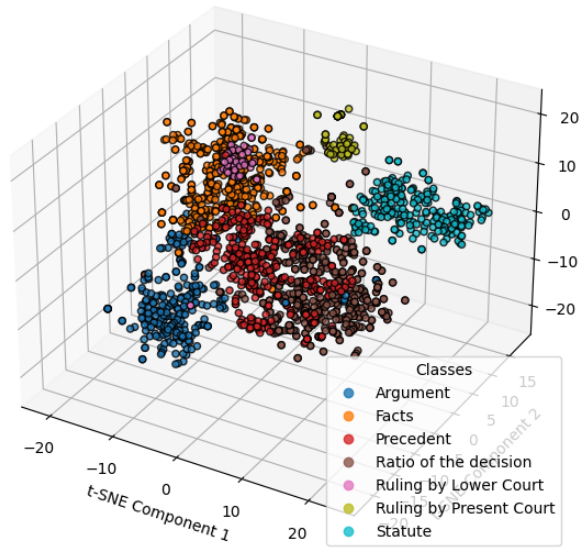

Figure 4: t-SNE visualization of prototype embeddings for seven class rhetorical role dataset from (Bhattacharya et al., 2019).

These visualizations highlight that the tolerance classes formed under the optimal ε threshold result in coherent clustering of instances belonging to the same rhetorical role. In both datasets, sentence embeddings associated with classes like `Facts`, `Precedent`, and `Ratio` are grouped into well-defined regions of the embedding space.

The clustering is especially pronounced in (Figure 4), where inter-class boundaries are clearer due to fewer overlapping semantic rhetorical roles. This suggests that SeqTNS is able to effectively group semantically and functionally similar sentences into shared tolerance classes, thereby supporting accurate and robust label assignment.

These insights offer an interpretable lens into the behavior of the soft clustering mechanism, validating the design choice of prototype based reasoning in SeqTNS.

Although the proposed SeqTNS classifier performs competitively on both rhetorical role datasets, interpretability is critical for adoption in real-world legal settings. Understanding why a sentence was assigned a particular rhetorical role helps improve trust and reliability.

To explain the predictions made by the TNS classifier, we employ the LIME framework (Local Interpretable Model-agnostic Explanations). The process is summarized in Algorithm 2.

Algorithm 2 LIME-Based Explanation of TNS Predictions

1: **Input:** Test sentence $s \in \mathcal{D}_{\text{test}}$, TNS classifier f, embedding model ϕ
2: Embed s: $\vec{s} = \phi(s)$
3: Predict rhetorical role: $y = f(\vec{s})$
4: Generate perturbed variants $\{s'_i\}$ by adding/removing words
5: Embed each variant: $\vec{s'_i} = \phi(s'_i)$
6: Predict labels for each variant: $y'_i = f(\vec{s'_i})$
7: Fit local surrogate model (logistic regression) on (s'_i, y'_i)
8: Extract top contributing words from the local model

LIME perturbs a sentence by randomly adding or removing words, embeds the variants using InLegalBERT, and observes how the TNS classifier responds to these changes. A simple logistic regression model is trained locally on the perturbed examples to identify which words contribute most to the classification decision.

This process is applied to all test set sentences. We then aggregate the top positively contributing words for each rhetorical role. Table 5 in the appendix section presents the top 10 most influential words per role.

The explanations highlight interpretable patterns: for instance, words like `court`, `judgment` and `date` contribute to `PREAMBLE`, whereas the word `accused` strongly contributes to `FAC` and `ANALYSIS`, while legal action words like `allowed`, `dismissed`, and `order` are dominant in `RPC`. Role specific

cues such as whether for ISSUE and submitted for ARG_PETITIONER and ARG_RESPONDENT further support the linguistic validity of the predictions.

These findings validate that SeqTNS along with embeddings from Figure 1 captures meaningful rhetorical structure and that LIME can effectively surface interpretable cues from its behavior.

4 Conclusion and Future Works

In this work, we introduce a novel tolerance based classifier (SeqTNS), that extends the Tolerance Near Sets framework by incorporating sequential label context for rhetorical role classification in Indian legal documents. The proposed method integrates semantic similarity with discourse-level context by combining sentence embeddings from fine-tuned LegalBERT and InLegalBERT with label aware BiLSTM representations. The classifier organizes instances into tolerance classes based on a cosine distance threshold, allowing flexible and interpretable classification.

Experiments on two benchmark datasets, one with thirteen and another with seven rhetorical roles. Demonstrate that the proposed model outperforms several strong baselines, including fine-tuned BERT classifiers and TNS without contextual information. On the thirteen class dataset, SeqTNS achieves a weighted F1 score of 0.78, significantly outperforming baseline transformer models while requiring 40% less training time compared to traditional BiLSTM-CRF architectures. Similarly, on the seven class dataset, SeqTNS achieves a macro F1 score of 0.83, exceeding the best reported performance in the original dataset paper (Bhattacharya et al., 2019).

Beyond performance, we emphasized model transparency through LIME-based explanations and t-SNE visualizations. These analyses reveal that tolerance classes not only align well with ground truth labels but also group semantically and contextually similar sentences together. The most influential tokens extracted by LIME offer valuable insights into the linguistic cues driving rhetorical role predictions.

As **future work**, we plan to directly address this limitation by evaluating SeqTNS on cross-jurisdictional legal data (from corpora such as US or UK law) and on other sequential NLP tasks outside the legal domain. We also intend to include LLM-based explanation methods (e.g., CoT or layer-probing) as part of our future work. Incorporating long-range and hierarchical dependencies remains an important next step. Additionally, we plan to investigate re-sampling, augmentation, and contrastive learning strategies to further address imbalance issues.

Limitations

While the proposed SeqTNS framework shows strong performance and interpretability, there are a few limitations to consider. First, the approach relies on a fixed tolerance threshold (ϵ), which requires tuning per dataset and may not generalize well to unseen distributions. Second, although label sequence modelling captures local context via neighboring label embeddings, it does not yet incorporate broader document-level discourse structure. Third, while LIME provides useful token-level insights, it operates as a post-hoc approximation and may not always reflect the true internal reasoning of the classifier. Finally, the proposed model is performing well for all classes of the seven class dataset, but not as well for the thirteen class dataset because the dataset is highly imbalanced and some classes having not enough instances for model to learn.

Ethics Statement

This research was conducted using publicly available legal datasets released for academic and research purposes. No private or personally identifiable information was involved at any stage. The primary goal of this work is to explore sentence level rhetorical role classification in legal documents, which can support downstream tasks like legal summarization or structuring. While the proposed models show promising results, they reflect patterns present in the training data. Any biases, inaccuracies, or limitations in the dataset may influence model predictions. Therefore, these models should not be seen as replacements for human legal reasoning. We strongly encourage users to apply this work responsibly and ethically, keeping in mind the sensitive nature of legal decision making.

Acknowledgments

Sheela Ramanna's work was supported by NSERC Discovery Grant #194376. This work was also supported by the ANRF-SERB-CRG project titled "A Deep Explainable Framework for Semantically Similar Document Retrieval and Summarization of Legal Text," under the supervision of Dr. Anand

Kumar M, Department of Information Technology, National Institute of Technology Karnataka, Surathkal.

References

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identifying rhetorical roles of sentences in indian legal judgments. In *Proceedings of the 32nd International Conference on Legal Knowledge and Information Systems (JURIX)*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Sourav Dutta. 2021. Categorizing roles of legal texts via sequence tagging on domain-specific language models. In *Forum for Information Retrieval Evaluation (FIRE)*, volume 3159 of *CEUR-WS.org*, India. CEUR Workshop Proceedings. ©2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Zhiyong He, Zanbo Wang, Wei Wei, Shanshan Feng, Xianling Mao, and Sheng Jiang. 2020. A survey on recent advances in sequence labeling from deep learning models. *arXiv preprint arXiv:2011.06727*. 16 pages.

Tejaswi Hegde, Kalyane Satyam Sanjay, Swetha Mary Thomas, Ranjana Kambhammettu, M Anand Kumar, and Sheela Ramanna. 2023. Impact of vector embeddings on the performance of tolerance near sets-based sentiment classifier for text classification. *Procedia Computer Science, KES 2023*, 225:645–654.

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*.

Siddharth Kelkar, Srinivasa Ravi, Sheela Ramanna, and Anand Kumar Madasamy. 2024. Multimodal propaganda detection in memes with tolerance-based soft computing method. In *International Joint Conference on Rough Sets (IJCRS)*, volume 14839 of *Lecture Notes in Computer Science (LNAI)*, pages 343–351. Springer. First Online: 25 July 2024.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Angshuman Hazarika, Shubham 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. Association for Computational Linguistics.

Arpan Mandal, Kripabandhu Ghosh, Saptarshi Ghosh, and Sekhar Mandal. 2022. A sequence labeling model for catchphrase identification from legal case documents. *Artificial Intelligence and Law*, 30(3):325–358.

Vrushang Patel, Sheela Ramanna, Ketan Kotecha, and Rahee Walambe. 2022. Short text classification with tolerance-based soft computing method. *Algorithms*, 15(8).

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*.

James F. Peters. 2007. Near sets. Special theory about nearness of objects. *Fundamenta Informaticae*, 75(1-4):407–433.

James F. Peters. 2009. Tolerance near sets and image correspondence. *Int. J. of Bio-Inspired Computation*, 1(4):239–245.

James F. Peters and Som Naimpally. 2012. Applications of near sets. *Notices of the American Mathematical Society*, 59:536–542.

M. Saravanan, B. 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*. IJCNLP.

Xuerui Shang. 2022. A computational intelligence model for legal prediction and decision support. *Computational Intelligence and Neuroscience*, 2022:1–8.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4080–4090.

Alexei B. Sossinsky. 1986. Tolerance space theory and some applications. *Acta Applicandae Mathematicae: An International Survey Journal on Applying Mathematics and Mathematical Applications*, 5(2):137–167.

Piotr Wasilewski, James F. Peters, and Sheela Ramanna. 2011. Perceptual tolerance intersection. *Transactions on Rough Sets Journal*, 13:159–174.

Eric C. Zeeman. 1962. The topology of the brain and visual perception. *Univ. of Georgia Institute Conf. Proc.*, pages 240–256. , M.K. Fort, Jr. (Ed.), Topology of 3-Manifolds and Related Topics, Prentice-Hall, Inc.

A Appendix

A.1 LIME based Word Importance Explanations

Table 5 highlights the top words that the classifier associates with each rhetorical role, as identified through LIME explanations. These words provide insight into the linguistic cues the model uses to differentiate between rhetorical functions in legal texts.

For PREAMBLE sentences, common words include judgment, date, court, versus, and advocate, which are typically found in the opening section of a judgment, where basic case metadata and party names are introduced.

The NONE category features terms like judge, heard, list, and signature, often appearing in procedural or administrative content that does not fit into other specific rhetorical roles.

FACT sentences tend to include words such as accused, filed, assessee, complainant, and appeal, reflecting the narration of events, background context, and involved parties.

In ARG_PETITIONER, words like submitted, counsel, learned, contended, and argued point to sentences containing the petitioner's legal arguments or claims.

ANALYSIS sections frequently contain words like accused, also, therefore, stated, and evidence, which are commonly used in judicial reasoning or evaluative commentary.

For ARG_RESPONDENT, similar argumentative words appear, including submitted, learned, counsel, submits, and urged, indicating the respondent's side of legal reasoning.

PRE_RELIED is marked by citations and references, with frequent use of terms such as case, held, supreme, court, and principle, reflecting reliance on prior judgments and legal doctrines.

The RATIO category contains words like view, opinion, therefore, failed, and circumstances, typically associated with the core reasoning behind a judicial decision.

In RULING BY PRESENT COURT (RPC), the model identifies words like allowed, dismissed, order, result, and accordingly, which are indicative of the final judgment or conclusion delivered by the court.

ISSUE sentences are characterized by question-oriented words such as whether, points, determination, question, and examined, aligning with how judges formally pose legal questions that need to be resolved in the case.

RULING BY LOWER COURT (RLC) is often signaled by words like held, sentenced, convicted, acquitted, and tribunal, which refer to the outcomes or opinions expressed by subordinate courts.

The STATUTE (STA) role includes legal and legislative terms like section, offences, defines, act, and provision, typically pointing to references to specific statutory laws or provisions.

Finally, PRE_NOT_RELIED contains uncommon or obscure legal terms like jurisprudence, quoted, apposite, and question, often referring to precedents that were mentioned but not relied upon by the court in the final decision.

Overall, this analysis illustrates that the classifier captures intuitive and legally meaningful cues for each rhetorical role. The words highlighted by LIME align well with the expected language used in different sections of a judgment, supporting the interpretability and trustworthiness of the SeqTNS model's predictions.

Table 5: Top Contributing Words for Each Rhetorical Role Based on LIME Explanations.

Rhetorical Role	**Top Contributing Words (with Importance Scores)**
PREAMBLE	judgment (0.5883), date (0.5135), court (0.5063), versus (0.4945), held (0.3530), advocate (0.3529), respondent (0.3314), commissioner (0.2972), high (0.2845), petitioner (0.2541)
NONE	judge (0.4559), heard (0.3422), ext (0.2536), list (0.2379), signature (0.1985), delivered (0.1750), appellate (0.1346), court (0.1147), appeal (0.0841), marked (0.0599)
FAC	accused (0.7206), said (0.3747), filed (0.3598), hence (0.3332), assessee (0.3330), complainant (0.3088), also (0.2898), appeal (0.2584), appellant (0.2415), petitioner (0.2305)
ARG_PETITIONER	submitted (0.2926), counsel (0.1363), learned (0.1360), submission (0.1068), contended (0.0575), submits (0.0489), petitioner (0.0314), urged (0.0281), argued (0.0251), petitioners (0.0204)
ANALYSIS	accused (0.5806), also (0.5069), therefore (0.3259), said (0.3165), stated (0.2535), prosecution (0.2458), even (0.2310), fact (0.2256), cannot (0.2111), evidence (0.1647)
ARG_RESPONDENT	submitted (0.0794), learned (0.0364), counsel (0.0252), submits (0.0202), urged (0.0109), petitioner (0.0108), argued (0.0081), submission (0.0070), dismissed (0.0067), bail (0.0065)
PRE_RELIED	case (0.1097), held (0.1093), supreme (0.0893), court (0.0700), apex (0.0526), observed (0.0381), decision (0.0332), supra (0.0328), see (0.0315), principle (0.0285)
RATIO	view (0.0094), find (0.0085), opinion (0.0065), therefore (0.0040), failed (0.0038), prosecution (0.0036), cannot (0.0029), hence (0.0029), circumstances (0.0028), substantial (0.0027)
RPC	allowed (0.5888), dismissed (0.4244), shall (0.2068), appeal (0.1621), order (0.1596), result (0.1545), disposed (0.1371), enter (0.1256), accordingly (0.1247), costs (0.1163)
ISSUE	whether (0.1125), points (0.0267), determination (0.0148), accused (0.0120), question (0.0097), ipc (0.0062), questions (0.0059), point (0.0056), seized (0.0043), examined (0.0039)
RLC	held (0.0790), sentenced (0.0616), convicted (0.0537), acquitted (0.0407), appellant (0.0315), dismissed (0.0274), conviction (0.0243), tribunal (0.0200), court (0.0196), high (0.0175)
STA	section (0.0704), offences (0.0157), shall (0.0155), defines (0.0112), offence (0.0080), provides (0.0075), act (0.0068), provision (0.0060), state (0.0057), goods (0.0056)
PRE_NOT_RELIED	jurisprudence (0.0005), negativing (0.0003), quoted (0.0002), arose (0.0001), connection (0.0001), apposite (0.0001), brannan (0.0001), 341 (0.0001), oft (0.0001), question (0.0001)