# Enhancing Rhetorical Role Identification in Legal Documents using Large Language Models and IN_place Data Augmentation

Reshma Sheik[1,2][0000−0003−3567−9757], Chandrababu Namani[3], Prasanna P[1], and S Jaya Nirmala[1][0000−0001−5432−4156]

[1] National Institute of Technology, Trichy, Tamil Nadu
[2] TKM College of Engineering, Kollam, Kerala
[3] MVSR Engineering College, Hyderabad, Telangana
{rezmasheik,chandranamani9271,prasannacsnitt}@gmail.com, sjaya@nitt.edu

**Abstract.** Understanding the rhetorical roles of sentences within legal documents is crucial for various downstream tasks, including semantic search, summarization, and case law analysis. However, the complex structure of legal case documents, coupled with the interplay of various themes, poses challenges even for human experts. In this paper, we perform data augmentation using open-sourced Large Language Models (LLMs) to automate the identification of rhetorical roles. Specifically, we explore data augmentation techniques to address class imbalance issues within datasets, introducing a novel augmentation technique called IN_Place Augmentation to mitigate linear dependency issues. Additionally, a comparative analysis of two neural network architectures, InLegalBERT and Hierarchical BiLSTM, combined with Conditional Random Fields (Hier-BiLSTM-CRF), integrating different types of sentence embeddings was conducted. We fine-tune the InLegalBERT model using our proprietary dataset and utilize the fine-tuned embeddings to train the Hier-BiLSTM-CRF model. Our evaluations demonstrate the efficacy of the fine-tuned InLegalBERT model across diverse legal contexts, showcasing significant improvements in the Land & Property domain with an increase in the weighted average F1 score from 0.729 to 0.884, and in the Criminal domain with an increase in the macro average F1 score from 0.631 to 0.817. In summary, our contributions include leveraging LLM-Mistral 7B for data augmentation, introducing IN_Place augmentation, utilizing domain-specific transformer InLegalBERT, and developing a hybrid model integrating InLegalBERT with Hier-BiLSTM-CRF.

**Keywords:** Rhetorical Roles· legal domain· Large Language Models· IN_place augmentation · InLegalBERT · Hier-BiLSTM-CRF · Mistral 7B.

## 1 Introduction

Labeling the rhetorical roles of sentences in legal documents involves discerning the semantic function associated with each sentence. These roles encompass

various themes, such as presenting case facts, articulating arguments of the involved parties, and delivering the court's final judgment. Accurately identifying these roles is pivotal for facilitating numerous downstream tasks, including semantic search [10], summarization [13]; [7], and case law analysis [14] and so on. Legal case documents exhibit significant structural variability [16], [1], with diverse themes frequently intertwining. For example, the rationale underlying a judgment (known as the Ratio of the decision) often intersects with references to precedents and statutes. Consequently, discerning the subtle distinctions between rhetorical roles can pose challenges, even for legal experts.

In this paper, we explore the utilization of the open-sourced Large Language Model (LLM) - Mistral [9] for addressing class imbalance issues within datasets through data augmentation techniques. Specifically, we investigate two distinct approaches to data augmentation. The first approach involves traditional data augmentation practices, where newly generated samples are conventionally appended to the end of the dataset. In contrast, our research introduces a novel augmentation technique termed IN_Place Augmentation, specifically designed to mitigate linear dependency issues among dataset samples. With IN_Place Augmentation, each newly generated sample is augmented beneath its corresponding original sample, offering a nuanced strategy for addressing data imbalances and enhancing dataset diversity. We perform the comparative analysis of two distinct neural network architectures: the InLegalBERT model [12] and the Hierarchical BiLSTM model combined with Conditional Random Fields (Hier-BiLSTM-CRF). Our study focuses on exploring the performance of the later model when integrated with different types of sentence embeddings, namely Sent2Vec[11] embeddings, InLegalBERT embeddings, and InLegalBERT embeddings after augmentation. To summarize, this work offers the following key contributions: Leveraged an Open-Source LLM for effective data augmentation. Introduced the innovative IN_Place Augmentation method to address linear dependency issues in the dataset. Employed a domain-specific transformer, InLegalBERT, for both training and evaluation A novel hybrid architecture incorporating the fine-tuned InLegalBERT embeddings into the Hier-BiLSTM-CRF architecture

The structure of our paper is organized as follows: In Section 2, we explore the classification of rhetorical labeling based on prior literature. In Section 3, we present the Indian dataset tailored for rhetorical role-labeled tasks, highlighting the necessity for data augmentation. Section 4 clarifies our methodology for data augmentation and the diverse models employed for training. Following this, in Section 5, we present our empirical findings, contrasting results with and without augmentation and providing classification scores for different models. Finally, Section 6 concludes our work by summarizing key insights and important findings.

## 2 Related Work

Previous research has explored the application of deep learning (DL) methods in rhetorical role identification tasks. [10] utilized word embeddings in conjunction with the FastText classifier to perform binary classification, distinguishing between factual and non-factual sentences within legal documents. Their study demonstrated the effectiveness of deep learning techniques, particularly in the specialized domain of Immigration documents. Similarly, [19] employed neural architectures to label the rhetorical roles of sentences, further highlighting the potential of DL approaches in this field. This study investigates the efficacy of transformer architectures like BERT [6] and LegalBERT [5] across various legal tasks, including legal judgment prediction [4], prior-case retrieval [15], and crime classification [18]. It is worth noting that previous research has not addressed the task of rhetorical role labeling across multiple domains. In our current study, we showcase the effectiveness of our models across five different domains, filling a gap in the existing literature and expanding the scope of research in this area.

Earlier investigations focused mainly on automating the identification of rhetorical roles within legal documents. For instance, [7] introduced LetSum, which categorizes text structures into five themes—Introduction, Context, Juridical Analysis, and Conclusion—primarily focusing on Canadian case documents. They utilized 'section titles' inherent in these documents and curated linguistic phrases associated with each theme. Similarly, [13] employed Conditional Random Fields (CRF) for the same task, defining seven rhetorical roles. Their study, conducted on 200 Kerala High Court documents spanning rent control, income tax, and sales tax domains, involved labeling sentences as facts or principles using handcrafted rules and the Multinomial Bayes Classifier. Additionally, [16] explored sentence labeling as facts or principles through the utilization of handcrafted features and the Multinomial Bayes Classifier, focusing on a diverse set of legal documents. [14] investigated the segmentation of U.S. court documents into both functional sections and issue-specific segments. [10] developed a methodology for discerning factual and non-factual sentences in Canadian immigration case documents. They annotated 150 such documents, labeled sentences as either factual or non-factual, and trained word embeddings on a large legal corpus. Leveraging the FastText classifier, their approach aimed to identify documents containing fact-asserting sentences similar to a given query. [19] employed Bi-LSTM-CRF architecture along with heading encoders for rhetorical sentence labeling in Japanese documents. [16] focused on distinguishing "facts" and "legal principles" in cited cases within legal documents. They annotated 50 common law reports from the British and Irish Legal Institute (BAILII) website, considering paragraphs with at least one citation. Our study shares similarities with the research conducted by Bhattacharya et al. (2023), as referenced [3], in which they also explored the identification of rhetorical roles in Indian judicial decisions. However, their methodology did not prioritize enhancing the dataset through augmentation techniques, unlike our approach, which significantly emphasizes dataset refinement strategies using LLMs.

## 3 Dataset

The dataset [1] contains legal judgments from The Supreme Court (SC) of India. It has 50 documents split into sentences using SpaCy. The dataset has five domains, namely Land & Property, Constitutional, Criminal, Intellectual Property, Labour & Industrial Law. Each domain has seven labels: Ratio of the decision, Facts, Precedent, Argument, Statute, Ruling by Lower Court, and Ruling by Present Court. The documents split into sentences undergo a human annotation to assign labels to each sentence. Since the annotation process is subjective, the final label is obtained after an inter-annotation agreement [3]. Figure 1 shows the distribution of statements among different labels. It is evident from the figure that the distribution is extremely uneven, with maximum(78%) statements belonging to just three labels: Ratio of the decision, Facts, and precedent. This leads to the need for data augmentation, which is covered in depth in Section 4.1.

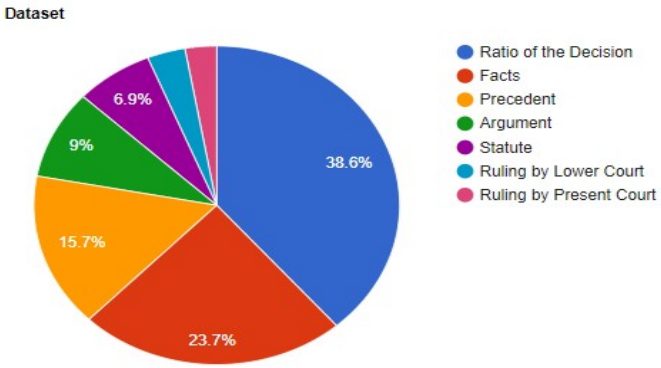


**Fig. 1.** Statistics of the Dataset

## 4 Methodology

Figure 2 explains the overall framework. The training data is sourced from the rhetorical role labeling dataset [2]. This dataset undergoes a dynamic data augmentation (refer to section 4.1) process leveraging one-shot learning and prompt engineering techniques. The augmented dataset thus created is used by various models (discussed in section 4.2) for training and classification.

### 4.1 Data Augmentation

For the augmentation, the documents available in the dataset were segregated into five folds, and the number of statements on each label was calculated. The

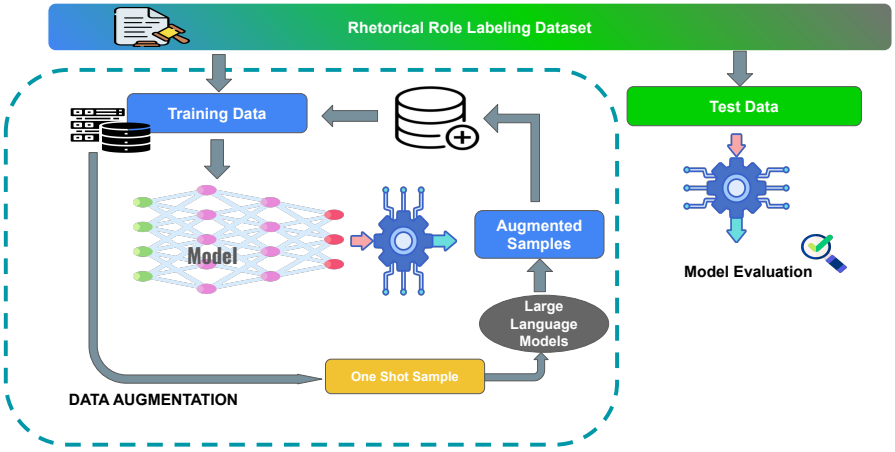


**Fig. 2.** Overall Framework

data augmentation was done dynamically for each fold depending on the distribution of the statements in the corresponding fold. Statistical features for each fold were calculated, and the number of new samples to be generated was discussed based on AugGenCount defined in equation 1.

$$AugGenCount_i = \frac{median_{fold_i}}{count_{label_i}} \tag{1}$$

where:

- $AugGenCount_i$ is the AugGenCount calculated for label $i$ in the corresponding fold.
- $median_{fold_i}$ is the overall median of fold $i$.
- $count_{label_i}$ is the count of label $i$.

The $AugGenCount$ value indicates the number of synthetic samples to be generated for each label. If $AugGenCount$ is less than 2, the generation of synthetic samples can be avoided. If $AugGenCount$ is from 2 up to 5, the number of samples to be generated is $AugGenCount - 1$. For any $AugGenCount$ value greater than 5(inclusive), the number of samples to be generated is 3. Thus, the data augmentation is performed dynamically depending on the statistics of the data samples.

Large language models such as Falcon[4], LLaMa [17], and Mistral were tested for data generation. While Falcon was prone to hallucinations, LLaMa produced inaccurate results, and Mistral produced accurate and semantically similar data through a prompt engineering method. Hence, Mistral 7B was used for the gen-

[4] https://falconllm.tii.ae/index.html

eration of data. Our project utilizes the Mistral 7B Instruct v0.1 [5] model, undergoes quantization involves "type-1" 4-bit quantization within super-blocks consisting of 8 blocks, each containing 32 weights. Scales and mins are quantized with 6 bits. With a size of 4.37 GB and a maximum RAM requirement of 6.87 GB, this model is tailored for medium-sized use cases, offering balanced quality. This quantization method results in one bit per weight (bpw) usage of 4.5. One-shot learning was employed, where an example from the underrepresented label was given to the model, and similar statements were generated. To ensure the semantic similarity of the generated samples, prompt engineering techniques were employed, where the model was prompted to assume the role of an annotator with a detailed description provided. This resulted in creating artificial samples that exhibited similarity to those produced by a human annotator. Two techniques were employed to integrate the synthetic data with the original dataset. Firstly, newly generated samples were appended at the end. Secondly, augmented data was integrated into the document using a method known as IN_Place augmentation, wherein the generated data was appended to the document alongside the statement from which it originated. Specifically, the generated data was appended next to the original statement rather than at the end of the document.

### 4.2 Models for Training

**InLegalBERT model:** We trained the Indian legal domain-specific transformer, InLegalBERT, using our data and evaluated them on five different domains. Following the training phase, we conducted comprehensive evaluations of the InLegalBERT[6] model across five different legal domains. InLegalBERT is a BERT model trained on an extensive corpus of legal documents sourced from the Indian SC and numerous High Courts from 1950 to 2019. These documents cover various legal domains, including Civil, Criminal, and Constitutional. The dataset comprises approximately 5.4 million Indian legal documents. InLegalBERT incorporates domain-specific knowledge tailored to the intricacies of the Indian legal context, thereby enhancing the original BERT model's applicability to legal text analysis tasks.

**Neural Model Hier-BiLSTM-CRF with pre-trained sentence embeddings** We use the hierarchical BiLSTM architecture [8] used in [3] to automatically extract features for identifying the rhetorical roles. We chose baseline models from [3] for comparison purposes. Different sentence embeddings were incorporated to this architecture as detailed below.

*Baseline Model-Pre-trained embeddings from a Sent2Vec model:* Sent2Vec is an unsupervised model for generating sentence embeddings based on the Continuous Bag-of-Words model. It creates embeddings by averaging unigram and n-gram

[5] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1
[6] https://huggingface.co/law-ai/InLegalBERT

embeddings from the sentence. During training, the entire sentence serves as the context window. The model predicts words within the sentence, treating all vocabulary words as class labels, with the whole sentence as the context. In [3], a Sent2Vec model was trained on a large collection of legal court case documents, specifically from the Indian Supreme Court. The training excluded 50 test documents to ensure model independence during rhetorical role labeling tests. Once trained, the Sent2Vec model was employed to infer embeddings for sentences within the test documents.

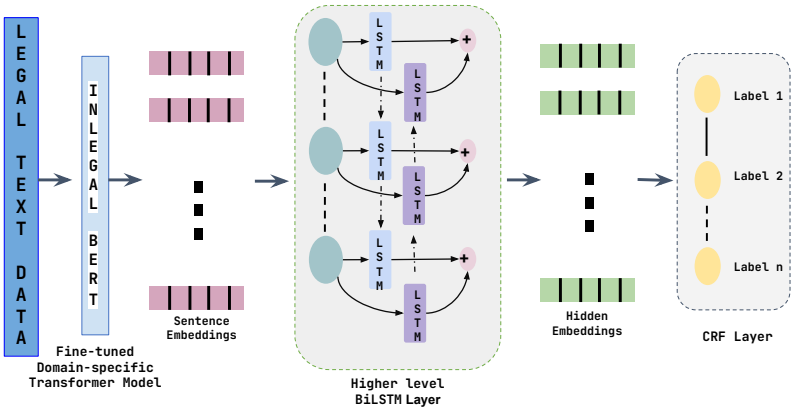


**Fig. 3.** Model architecture using InLegalBERT embeddings to Hier BiLSTM-CRF

*Using Fine-tuned InLegalBERT Embeddings in Hier BiLSTM-CRF:* InLegalBERT_FT denotes a fine-tuned iteration of the InLegalBERT model specifically tailored to our dataset. Through fine-tuning, our model undergoes a process of learning the intricacies and nuances of our dataset, thereby enhancing its ability to generate embeddings that are more relevant and contextually rich. These fine-tuned embeddings serve as a foundation for subsequent training using the Hier-BiLSTM-CRF model. Upon training the Hier-BiLSTM-CRF model with our fine-tuned embeddings, we conducted comprehensive evaluations to assess its performance. Our results surpassed those of the baseline model utilizing Sent2Vec embeddings, demonstrating the superior performance of our approach. This achievement underscores the efficacy of leveraging domain-specific embeddings and fine-tuning techniques in legal text analysis tasks, showcasing the potential of our model to outperform conventional methods. Figure 3 shows the model architecture, which uses the InLegalBERT embeddings to Hier BiLSTM-CRF model

*Using Fine-tuned InLegalBERT Embeddings from IN_Place augmented dataset:* In our research, we conducted fine-tuning of the InLegalBERT model using our

proprietary IN_place augmented dataset to enhance its capacity to generate accurate sentence embeddings. This process aimed to optimize the model's performance specifically for our target domain, legal text analysis. By fine-tuning InLegalBERT with our IN_Place augmented dataset, we aimed to improve its ability to capture domain-specific nuances and semantic features relevant to legal documents.

In the variations as mentioned above, sentence embeddings generated by the upper-level BiLSTM layer are fed into a subsequent feed-forward network, where probability scores for each associated label are computed. The dimensionality of these embeddings differs across models, with 200 dimensions for the Sent2Vec model and 768 dimensions for the BERT models. Each neural model undergoes 200 epochs of training with a fixed learning rate of 0.001, determined based on performance evaluations on the validation set. A dropout rate of 0.5 and a regularization strength of 0.0005 are uniformly applied to mitigate overfitting. Batch sizes are chosen differently, with a size of 32 for the Sent2Vec model and 16 for all BERT models, considering computational efficiency and empirical performance. The label assigned to each sentence is determined by the highest predicted probability score, ensuring straightforward classification interpretation.

## 5 Results

We employ a rigorous cross-validation strategy where validation data is exclusively sourced from gold-standard datasets. To achieve this, we utilize two distinct sets of data: augmented and un-augmented files. The training data is derived from the augmented files, while the validation data is sourced from the un-augmented files. By segregating the training and validation datasets in this manner, we aim to provide a robust assessment of model performance and validate its efficacy under real-world conditions. The evaluation includes weighted average F1 score and macro average F1 score.

**Table 1.** Table shows the evaluation of weighted average F1 score and macro average F1 score on InLegalBERT model improvement after augmentation.

| Domain | Weighted Average F1 | Macro Average F1 |
|---|---|---|
| Land & Property | 0.65 ->0.67 | 0.40 ->0.55 |
| Criminal | 0.64 ->0.65 | 0.50 ->0.57 |
| Labour & Industrial Law | 0.61 ->0.64 | 0.34 ->0.47 |
| Constitutional | 0.69 ->0.66 | 0.42 ->0.47 |
| Intellectual Property | 0.65 ->0.62 | 0.48 ->0.51 |

Table 1 shows using the InLegalBERT model, there is an improvement in the F1 scores. The highest weighted average F1 score, 0.67, is for the Land & Property domain.

**Table 2.** Table shows weighted average F1 scores for Hier BiLSTM-CRF model across diverse legal domains for different sentence embeddings

| Domain | Sent2Vec | InLegalBERT_FT | InLegalBERT_FT (Augmented) |
|---|---|---|---|
| Land & Property | 0.729 | 0.884 | 0.822 |
| Criminal | 0.781 | 0.852 | 0.727 |
| Labour & Industrial Law | 0.649 | 0.750 | 0.700 |
| Constitutional | 0.644 | 0.724 | 0.759 |
| Intellectual Property | 0.625 | 0.743 | 0.770 |

**Table 3.** Table shows macro average F1 scores for Hier BiLSTM-CRF model across diverse legal domains for different sentence embeddings.

| Domain | Sent2Vec | InLegalBERT_FT | InLegalBERT_FT (Augmented) |
|---|---|---|---|
| Land & Property | 0.641 | 0.805 | 0.747 |
| Criminal | 0.631 | 0.817 | 0.665 |
| Labour & Industrial Law | 0.504 | 0.684 | 0.565 |
| Constitutional | 0.530 | 0.626 | 0.651 |
| Intellectual Property | 0.486 | 0.674 | 0.742 |

Table 2 displays weighted average F1 scores from the evaluation of a Hier BiLSTM-CRF model across diverse legal domains. It compares three types of sentence embeddings: Sent2Vec (Baseline), InLegalBERT Fine-Tuned (InLegalBERT_FT), and Augmented InLegalBERT Fine-Tuned (InLegalBERT_FT). InLegalBERT_FT embeddings outperform Sent2Vec, with the highest score (0.884) in Land & Property. Augmented InLegalBERT_FT embeddings excel in the Constitutional and Intellectual Property domains, achieving scores of 0.759 and 0.770, respectively. Similarly, Table 3 shows the macro average F1 scores of Hier-BiLSTM-CRF results indicating improvements in scores achieved with InLegalBERT_FT embeddings of the highest 0.817 for the Criminal domain. Augmented InLegalBERT_FT embeddings excel in the Constitutional and Intellectual Property domains, achieving scores of 0.651 and 0.742, respectively. Additionally, we observed a significant enhancement in the rhetorical role classification task across Land & Property, Intellectual Property, and Constitutional domains when utilizing the Hier-BiLSTM-CRF model trained with IN_Place augmentation compared to augmenting sentences to the end.

## 6 Conclusion

In conclusion, this paper addresses two significant challenges within the legal domain. Firstly, it tackles the need for high-quality data to train large deep neural network models. Acquiring such labeled datasets in the legal domain is challenging due to manual annotation processes and the potential for misleading cross-validation of labels. Secondly, this work addresses the critical issue of rhetorical role labeling, which is important for legal analysis and various downstream tasks while improving the readability of lengthy judicial statements. To

overcome the scarcity of labeled data, a combination of prompt engineering, one-shot learning, and data augmentation techniques were employed. The resultant dataset facilitated the training of deep neural network models and significantly improved classification scores for rhetorical role-labeling tasks. Using the domain-specific transformer model InLegalBERT, which was fine-tuned on the augmented dataset, enhanced its capacity to generate sentence embeddings accurately and was further utilized for training the Hier-BiLSTM-CRF model. Fine-tuned InLegalBERT Embeddings from IN_Place augmented dataset exhibited superior performance compared to the baseline model, showcasing significant advancements across most of the domains in the dataset.

## References

1. Bhattacharya, P., Hiware, K., Rajgaria, S., Pochhi, N., Ghosh, K., Ghosh, S.: A comparative study of summarization algorithms applied to legal case judgments. In: Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41. pp. 413–428. Springer (2019)
2. Bhattacharya, P., Paul, S., Ghosh, K., Ghosh, S., Wyner, A.: Identification of rhetorical roles of sentences in indian legal judgments. In: Legal Knowledge and Information Systems, pp. 3–12. IOS Press (2019)
3. Bhattacharya, P., Paul, S., Ghosh, K., Ghosh, S., Wyner, A.: Deeprhole: deep learning for rhetorical role labeling of sentences in legal case documents. Artificial Intelligence and Law pp. 1–38 (2023)
4. Chalkidis, I., Androutsopoulos, I., Aletras, N.: Neural legal judgment prediction in english. arXiv preprint arXiv:1906.02059 (2019)
5. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: Legal-bert: The muppets straight out of law school. arXiv preprint arXiv:2010.02559 (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
7. Farzindar, A.: Atefeh farzindar and guy lapalme,'letsum, an automatic legal text summarizing system'in t. gordon (ed.), legal knowledge and information systems. jurix 2004: The seventeenth annual conference. amsterdam: Ios press, 2004, pp. 11-18. In: Legal knowledge and information systems: JURIX 2004, the seventeenth annual conference. vol. 120, p. 11. IOS Press (2004)
8. Graves, A., Fernández, S., Schmidhuber, J.: Bidirectional lstm networks for improved phoneme classification and recognition. In: International conference on artificial neural networks. pp. 799–804. Springer (2005)
9. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
10. Nejadgholi, I., Bougueng, R., Witherspoon, S.: A semi-supervised training method for semantic search of legal facts in canadian immigration cases. In: JURIX. pp. 125–134 (2017)

11. Pagliardini, M., Gupta, P., Jaggi, M.: Unsupervised learning of sentence embeddings using compositional n-gram features. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 528–540 (2018)
12. Paul, S., Mandal, A., Goyal, P., Ghosh, S.: Pre-trained language models for the legal domain: a case study on indian law. In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. pp. 187–196 (2023)
13. Saravanan, M., Ravindran, B., Raman, S.: 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 (2008)
14. Savelka, J., Ashley, K.D.: Segmenting us court decisions into functional and issue specific parts. In: JURIX. pp. 111–120 (2018)
15. Shao, Y., Mao, J., Liu, Y., Ma, W., Satoh, K., Zhang, M., Ma, S.: Bert-pli: Modeling paragraph-level interactions for legal case retrieval. In: IJCAI. pp. 3501–3507 (2020)
16. Shulayeva, O., Siddharthan, A., Wyner, A.: Recognizing cited facts and principles in legal judgements. Artificial Intelligence and Law **25**(1), 107–126 (2017)
17. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
18. Wang, P., Fan, Y., Niu, S., Yang, Z., Zhang, Y., Guo, J.: Hierarchical matching network for crime classification. In: proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. pp. 325–334 (2019)
19. Yamada, H., Teufel, S., Tokunaga, T.: Neural network based rhetorical status classification for japanese judgment documents. In: Legal Knowledge and Information Systems, pp. 133–142. IOS Press (2019)