uOttawa at SemEval-2023 Task 6: Deep Learning for Legal Text Understanding

Intisar Almuslim and **Sean Stilwell** and **Surya Kiran Suresh** and **Diana Inkpen**
School of Electrical Engineering and Computer Science
University of Ottawa
Ottawa, ON, K1N 6N5
{iamul068, sstil051, ssrure044, diana.inkpen}@uottawa.ca

Abstract

We describe the methods we used for legal text understanding, specifically Task 6 Legal-Eval at SemEval 2023. The outcomes could assist law practitioners and help automate the working process of judicial systems. The shared task defined three main sub-tasks: sub-task A, Rhetorical Roles Prediction (RR); sub-task B, Legal Named Entities Extraction (L-NER); and sub-task C, Court Judgement Prediction with Explanation (CJPE). Our team addressed all three sub-tasks by exploring various Deep Learning (DL) based models. Overall, our team's approaches achieved promising results on all three sub-tasks, demonstrating the potential of deep learning-based models in the judicial domain.

1 Introduction

The legal sector generates an overwhelming volume of information from many sources, including law firms, law courts, independent attorneys, and legislators. As a result, tools to help legal professionals manage large volumes of legal documents are becoming increasingly necessary.

The LegalEval 2023 shared task (Modi et al., 2023)[1] aims to build a legal research community by addressing three specific problems in the legal domain. First, Sub-task A, Rhetorical Roles Prediction (RR), involves identifying the rhetorical roles played by various sentences in a given text. Sub-task B, Legal Named Entities Extraction (L-NER), involves identifying distinct named entities that appear in legal texts. Finally, Sub-task C, Court Judgement Prediction with Explanation (CJPE), involves predicting the outcome of a given court case, along with an explanation of the reasoning behind the prediction.

Our team participated in all three sub-tasks of the shared task. In Sub-task A, we created a hierarchical BiLSTM-CRF model with randomly-initiated embeddings. In Sub-task B, we created a BERT-based named entity recognition model. In Sub-Task C, we used CaseLawBERT to create judgement prediction and explanation models.

The rest of this paper is organized as follows: in Section 2, we briefly review the literature background on each sub-task. The detailed descriptions of our approaches to the three sub-tasks are presented in Sections 3, 4, and 5, respectively. Section 6 concludes the paper and comments on future work directions.

2 Related Work

Rhetorical role labeling was first automated by (Saravanan et al., 2008), where Conditional Random Fields (CRF) developed a generic approach to perform segmentation using seven rhetorical roles. Nejadgholi et al. (2017) developed a method for the identification of factual and nonfactual sentences using fastText. Later, the automatic Machine Learning approaches and rule-based scripts for rhetorical role identification evolved which was compared to (Walker et al., 2019). Kalamkar et al. (2022b) created a large corpus of RRs and propose transformer-based baseline models for RR prediction. The use of the Bi-LSTM-CRF model with sent2vec features to label rhetorical roles in Supreme Court documents in (Bhattacharya et al., 2019) serves as a precursor for transformer models. In this sub-task, we extend the work in (Bhattacharya et al., 2019) and explore transformer models to see which performs better in the prediction of rhetorical roles.

Named entity recognition for other domains has been a significant area of research. In the biomedical field, Schneider et al. (2020) and Hakala and Pyysalo (2019) both describe NER models for the biomedical domain. Arkhipov et al. (2019) and (Emelyanov and Artemova, 2019) both describe a multilingual named entity recognition system using a pre-trained BERT model in several languages. Winastwan (2022) describes a procedure for the

[1] https://sites.google.com/view/legaleval/home

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 580–588
July 13-14, 2023 ©2023 Association for Computational Linguistics

creation of a customized named entity recognition model using BERT. The described procedure is adapted and used for this subtask.

In the previous few years, researchers have successfully attempted Legal Judgment Prediction (LJP) tasks for the text of judicial cases in several languages and within different legal systems, such as the European, Chinese, and Indian systems. A new English LJP dataset containing 11,478 cases from the European Court of Human Rights (ECHR) was created by (Chalkidis et al., 2019), and the results of various DL architectures were reported on three tasks: binary classification, multi-class classification, and case importance prediction. Moreover, a hierarchical version of BERT (Devlin et al., 2019) has been proposed to overcome BERT's input token count limitation for the LJP task. A lot of work in the field of predicting law articles or charges is currently being done in China. In (Xiao et al., 2018), the Chinese AI and Law challenge dataset (CAIL 2018) was released for legal judgment prediction. The dataset includes annotations for the judgments of over 2.6 million criminal cases. It consists of detailed annotations of related law articles to cases, the prison terms, and the charges. For the Indian legal system, Malik et al. (2021) introduced the Indian Legal Document Corpus (ILDC) dataset and used it to experiment with the Case Judgment Prediction and Explanation (CJPE) task, providing various models and proposing a hierarchical occlusion-based model for explainability, which create the baseline for the current sub-task C.

3 Sub-task A: Rhetorical Roles Prediction

Legal documents are usually found to be lengthy, unstructured, and quite difficult to process. Reading and understanding legal text is arduous as they usually span from tens to hundreds of pages. In this subtask, we automatically annotate/classify sentences into semantically coherent units called rhetorical roles.

3.1 Dataset Details

The dataset for this subtask is provided by the organizers of LegalEval 2023. The dataset consists of Indian Court Judgements in which each line has been annotated by law students falling into one of the 13 different pre-defined rhetorical roles. The annotated predefined rhetorical roles are:

- Preamble (PREAMBLE)
- Facts(FAC)
- Ruling by Lower Court (RLC)
- Issues (ISSUE)
- Argument by Petitioner (ARG_PETITIONER)
- Argument by Respondent (ARG_RESPONDENT)
- Analysis (ANALYSIS)
- Statute (STA)
- Precedent Relied (PRE_RELIED)
- Precedent Not Relied (PRE_NOT_RELIED)
- Ratio of the decision (Ratio)
- Ruling by Present Court (RPC)
- NONE

More details on how the dataset was annotated and each of the rhetorical roles can be found in (Kalamkar et al., 2022b). These annotations are used to capture the metadata from the legal text and make it easier for understanding and summarizing the text. The data provided is already separated into training, development (dev), and test set. The training dataset contains 247 annotated court judgments, the dev dataset contains 30 annotated court judgments, and the test dataset contains 50 court judgments that are not annotated. Every sentence in these documents is annotated as one of the 13 rhetorical roles along with their start and end index in the document.

3.2 Preprocessing

Since the given training dataset has only 247 court judgments, we combine both the training and dev datasets. The combined dataset containing 277 documents is then shuffled and split into 70% for training and 30% for validation. The test data has 50 documents that are annotated with a DUMMY label that needs to be replaced with the predicted label. The dataset split is given in Table 1.

	Number of documents
Training	193
Validation	84
Test	50

Table 1: Dataset split for sub-task A (Rhetorical Roles Prediction)

As a part of the preprocessing of the data, we extract each annotated sentence from the documents and convert the text to lower case, remove the escape sequence character (\n) that is extracted from the text and replace it with space and remove any leading or trailing white spaces.

3.3 Methods

We adopted a variety of methods to get a good micro F1-score on the test dataset[2]. This includes using the Hierarchical BiLSTM-CRF model, transformer-based models like RoBERTa-Base (Liu et al., 2019) [3], LegalBERT (Chalkidis et al., 2020) [4], InLegalBERT (Paul et al., 2022) [5]. The transformer-based models were fine-tuned using AutoModelForSequenceClassification class with the hyperparameters indicated in Table 2.

Hyperparameter	Value
Models	legal-bert-base-uncased InLegalBERT roberta-base
Learning Rate	$2e-5$
Batch Size	8
Number of Epochs	5

Table 2: Hyperparameters of the fine-tuned transformer-based models for sub-task A

An experiment very similar to this subtask was carried out in (Bhattacharya et al., 2019) where the authors used a Hierarchical BiLSTM model with and without CRF (Conditional Random Fields) along with sent2vec embedding and a randomly initialized word embeddings using another BiLSTM model. In this subtask we incorporate a similar methodology by using sentence embeddings of GPT2 model (Radford et al., 2019) [6], SBERT model (Reimers and Gurevych, 2019a) [7] and sent2vec model (Pagliardini et al., 2018) as input to Hierarchical BiLSTM-CRF model. We have used the ***sentence-transformers/all-mpnet-base-v2*** model to generate the sentence embeddings using SBERT. A sent2vec model pre-trained on a legal corpus of 53K court case documents (Bhattacharya et al., 2019) was also used to generate sentence embeddings. We also used randomly initialized word embeddings similar to the one used in (Bhattacharya et al., 2019).

3.4 Results and Discussion

We found that variations using the Hierarchical BiLSTM-CRF model fetched far better results compared to the transformer-based models. We also found that all four different embeddings used in the experiment produced similar results, but the best score was obtained when randomly initialized word embeddings were used with the Hierarchical BiLSTM-CRF model. We also observed that all the transformer-based models showed a very low F1-score on the test dataset, whereas on the validation dataset, it obtained good results. The micro F1-scores on the test data for each of the models are shown in Table 3.

Model	F1-score
LegalBERT	0.27
RoBERTa-base	0.28
InLegalBERT	0.28
Hierarchical BiLSTM-CRF w/ GPT2 embeddings	0.59
Hierarchical BiLSTM-CRF w/ sent2vec embeddings	0.73
Hierarchical BiLSTM-CRF w/ randomly init embeddings	**0.74**
Hierarchical BiLSTM-CRF w/ SBERT embeddings	0.73
Best Performig Team	0.85

Table 3: Comparison of results of different models for sub-task A

4 Sub-task B: Legal Named Entities Extraction (L-NER)

This section describes our system for Sub-task B, which involves performing named entity extraction for the peculiar entity types that appear in legal documents.

We created a BERT-based named entity recognition model to accomplish this task, achieving an F1-score of 0.87 and finishing sixth in the shared task[8].

4.1 Dataset Details

We have carried out experiments using the datasets provided by the organizers of the shared task. The data has been split into a training, development, and test set. Each set consists of preambles, which consist in the formatted metadata of a judgement,

[2] Micro F1-score was the evaluation measure used for this sub-task of the shared task

[3] https://huggingface.co/roberta-base

[4] https://huggingface.co/nlpaueb/legal-bert-base-uncased

[5] https://huggingface.co/law-ai/InLegalBERT

[6] https://huggingface.co/gpt2

[7] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[8] This is a standard F1-score for strict detection of entities)

and the judgement itself. A breakdown of the number of preambles and judgements in each set is shown in Table 4.

Dataset	Preambles	Judgements	Entities
Training	1,560	9,435	29,964
Dev	125	949	3,216
Test	441	4,060	13,365

Table 4: Data Statistics for sub-task B (NER)

The entity types are described fully in (Kalamkar et al., 2022a). A preamble can contain a subset of five entities, while a judgement can contain any entity described excluding LAWYER. The total number of entities in each data set can be found in 4, while detailed counts per entity can be found in (Kalamkar et al., 2022a).

4.2 Preprocessing

For this sub-task, we began by converting the data to BIO format. Data was provided according to spaCy's JSON training format, which we converted into a dataframe containing the original text and corresponding labels in BIO format.

We then tokenize the data using the BertTokenizerFast class from the transformers library. For tokenization, we use the original LEGAL-BERT model proposed in (Chalkidis et al., 2020).

Following tokenization, we adjust the labels because the length of the sequence no longer matches the length of the original label, due to the splitting of certain words into multiple tokens or the use of special tokens by LEGAL-BERT, such as for padding. To align the tokens, we follow the procedure set out in (Winastwan, 2022) by matching the original labels to the tokenized sentence using the word IDs, which are consistent between the two formats.

4.3 Model Training

Our model uses the pre-trained LEGAL-BERT model described in (Chalkidis et al., 2020) as the base model for our training. Text is classified at the token level, so we use the BertForTokenClassification class with the output of each token classifier being equal to the number of unique entities in the data.

We then use a standard PyTorch training loop. Hyperparameters were tuned based on the trained model's performance on the development data provided in the shared task, the final values of the hyperparameters can be seen in Table 5.

Hyperparameter	Value
Learning Rate	$5e-3$
Batch Size	2
Number of Epochs	25
Model	legal-bert-base-uncased
Optimizer	SGD

Table 5: Hyperparameters of the fine-tuned model for sub-task B (L-NER)

4.4 Results and Discussion

Our best performance attained an F1-score of 0.87, ranking sixth overall for the shared task. The best-performing system in the shared task attained an F1-score of 0.91, while the baseline proposed in (Kalamkar et al., 2022a) also attained an F1-score of 0.91. The results of these models can be seen in Table 6.

Model	F1-score
Pinal-Patel (best performing)	0.91
en_legal_ner_trf (baseline)	0.91
uOttawa	**0.87**

Table 6: Comparison of results with the baseline and best performing model for sub-task B (L-NER)

Error analysis of the model on the development data showed a few recurring issues. One recurring issue was the mislabelling of the names of people, as names could be any of six possible entities for people (PETITIONER, RESPONDENT, JUDGE, LAWYER, WITNESS, and OTHER-PERSON).

Another recurring error involves a prefix of "dt" that was applied to some dates. The model tended to omit this prefix despite it being included in the DATE label.

5 Sub-task C: Court Judgement Prediction with Explanation

The aim of this sub-task is to automatically predict the outcome (accepted or rejected) of a legal case given the case document (Sub-task C-1) and also to provide explanations for the prediction (Sub-task C-2) by selecting the relevant sentences in the document that contribute to the decision.

5.1 Dataset Details

We have carried out experiments using the dataset released by the organizers of the LegalEval2023 shared task. The dataset is a subset of the one that was introduced by Malik et al. (2021), and it contains the case proceedings from the Supreme Court of India. It has two parts namely ILDC-single and ILDC-multi.

For ILDC-single, one decision is reached for a petition or for all the petitions together. While ILDC-multi documents include a variety of petitions involving different decisions. Document labelling in ILDC-multi sets the Label to '1' if a petition is accepted among multiple petitions, otherwise '0' to represent a 'rejected' document. The dataset statistics are given in Table 7.

	Number of documents
Train	ILDC-single: 4,982 ILDC-multi: 5,082
Validation	994
Test	Prediction Task C-1: 1,500 Explanation Task C-2: 50

Table 7: Data Statistics for Sub-task C (LJP)

5.2 Sub-task C-1: Legal Judgment Prediction (CJP)

The aim of this task is to predict the outcome of accepting or rejecting the appeal-petition case by classifying it into labels such as '1' for 'accepted' and '0' for 'rejected.' Hence, the task can be framed as a binary classification problem.

5.2.1 Methods

The ILDC dataset was initially introduced by Malik et al. (2021), where several baseline models have been developed for the judgment prediction task. The conducted experiments yielded four different types of models: classical models, sequential models, transformer models, and hierarchical transformer models.

According to Malik et al. (2021), domain-specific transformers like LEGAL-BERT (Chalkidis et al., 2020) or CaseLawBERT (Zheng et al., 2021) were not used in their experiments because they were trained on legal texts from other legal systems such as US or EU, so it was not known if they would function effectively on Indian law data. It is possible, however, to fine-tune these domain-specific model, even if they are not trained on Indian legal texts. This can be done in much the same way that other pre-trained models trained on general texts then utilized (with fine-tuning) to perform downstream tasks specific to the domain (Prasad et al., 2022). Therefore, we experimented with several Legal-BERT models such as Legal-BERT (Chalkidis et al., 2020)[9], CaseLawBERT (Zheng et al., 2021)[10], and InCaseLawBERT (Paul et al., 2022)[11] witch is a CaseLawBERT trained on on a large corpus of Indian legal corpus.

Similar to (Malik et al., 2021), we use two strategies to fine-tune these transformer models:

- Fine-tune these models and use them as classifiers.
- Fine-tune these models and use them as encoders to get inputs embeddings, then use neural networks such as: BiGRU and RCNN as classifiers.

We follow the same procedure set out in (Malik et al., 2021) by chunking documents and fine-tuning the transformers on ILDC-single corpus. Then, we use the fine-tuned transformers to extract the [CLS] token embeddings for each chunk in all the documents of ILDC-multi corpus. Each of these [CLS] representations are accumulated together to form the new sequence to be used as the input for either the transformer encoder models or the sequence encoder layers (Bi-GRU, RCNN) for classification.

5.2.2 Results and Discussion

The performance metrics of our models on the validation and test data are shown in Table 8 and Table 9 respectively. Out of 11 teams in sub-task C-1, our team placed third. Our best-performing model, CaseLawBERT + BiGRU, achieved an F1-score of 0.68, while the top-performing team achieved a score of 0.75. The baseline model, XLNet + BiGRU, had an F1-score of 0.78[12].

Out of all of our submitted runs, we found that the combination of CaseLawBERT and BiGRU performed the best and outperformed the other models, similar to what was mentioned in (Malik et al., 2021). This may prove that the combination of

[9] https://huggingface.co/nlpaueb/legal-bert-base-uncased

[10] https://huggingface.co/zlucia/legalbert

[11] https://huggingface.co/law-ai/InCaseLawBERT

[12] Average F1-score was used for this sub-task at the shared task

transformers with sequential models has the potential to be more effective in legal text classification tasks than using transformers alone.

Moreover, the successful performance of CaseLawBERT over the Indian CaseLawBERT can be attributed to the pretraining corpus used, which was constructed by ingesting the entire Harvard Law case corpus, amounting to 37GB in size and containing legal decisions across all US federal and state courts. On the other hand, the InCaseLawBERT model, which is trained on an Indian legal corpus of around 27GB, did not perform as well. This could be due to the fact that the model did not have access to the same amount of data as CaseLawBERT.

It is worth noting that the validation results are better than the test results, as the texts in the test set are much longer. The average numbers of tokens in the training datasets, single and multi, are 1,910 and 3,969, respectively, while there are, on average, 6,419 tokens in the test dataset. This suggests that the models may have struggled to generalize to longer texts.

5.3 Sub-task C-2: Court Judgement Prediction with Explanation (CJPE)

The objective of this task is to get explanations for a decision by identifying the significant sentences in a case that led to the decision, given the case document and the predicted decision for the case. The training process does not involve annotated explanations, as it is assumed that a model designed for prediction should be able to explain its decisions without explicit training on explanations.

5.3.1 Methods

Using the best judgment prediction model (CaseLawBERT + BiGRU) from task C-1, we experimented with the following method inspired from (Malik et al., 2021) and (Khan et al., 2020) in order to extract explanations. Just like (Malik et al., 2021), we apply a masking technique to the chunk embeddings of each document one by one in the BiGRU part of the model. The BiGRU is used to process the masked input, and the resulting probability of the label is compared with that of the original unmasked model to calculate the explainability score for each chunk. To extract explanatory sentences from the transformer section of the model, using the chunks that received positive scores, we explore the approach suggested in (Khan et al., 2020). The idea is to cluster the sentences that are contextually similar and pick one or two sentences from each cluster closest to the mean (centroid). We experimented with two variations as follows:

- Approach 1: cluster each chunk in the document and select one sentence from each cluster. Then combine all sentences from all chunks to form the explanation.

- Approach 2: combine all positive chunks into one chunk, cluster it, then select two sentences from each cluster. Then combine all sentences to form the explanation.

In order to split the text into sentences, we utilized the NLTK sentence tokenizer[13]. To obtain contextual embeddings of the sentences, we used Sentence Transformer(Reimers and Gurevych, 2019b)[14]. For the purpose of clustering, we applied K-means clustering[15]. Following the approach of (Bhattacharya et al., 2019), which employs seven rhetorical (semantic) roles to segment Indian case documents, we selected the number of clusters K as 7.

5.3.2 Results and Discussion

To evaluate the performance of the submitted work, The primary measure was the ROUGE score, as reported by the organizers. The machine explanations were evaluated with respect to the gold annotations (Malik et al., 2021). The results of our proposed approaches on the test set are reported in Table 10.

Although our team ranked 10th in this sub-task, the released results of all participants show that the best-performing method achieved a ROUGE-2 score of 0.047, which is within the same range as our obtained score of 0.040, but far from the baseline in (Malik et al., 2021) where the score was 0.303. After performing a simple error analysis, we conclude that the reason for our low score is the way in which we chose sentences for each chunk. When compared with Malik et al. (2021), where the top k sentences (40%) in each chunk were selected, we only selected one or two sentences per chunk. Therefore, there is a lower possibility of overlapping in our approach. In spite of this, the results confirm that the task of explaining legal outcomes is certainly challenging.

[13] https://www.nltk.org/.

[14] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[15] https://www.nltk.org/_modules/nltk/cluster/kmeans.html

	P	R	F1	Acc
Base LegalBERT	0.63	0.58	0.61	0.59
CaseLawBERT	0.64	0.61	0.62	0.63
InCaseLawBERT	0.63	0.61	0.62	0.60
Base LegalBERT + BiGRU	0.70	0.66	0.67	0.66
CaseLawBERT + BiGRU	**0.78**	**0.74**	**0.76**	**0.74**
InCaseLawBERT + BiGRU	0.70	0.69	0.69	0.69
Base LegalBERT + RCNN	0.76	0.72	0.74	0.72
CaseLawBERT + RCNN	0.77	0.71	0.74	0.71
InCaseLawBERT + RCNN	0.74	0.70	0.72	0.70

Table 8: Court Judgement Prediction results on the validation dataset

	P	R	F1	Acc
CaseLawBERT + BiGRU	**0.68**	**0.67**	**0.68**	**0.68**
CaseLawBERT + RCNN	0.55	0.54	0.54	0.55
InCaseLawBERT + BiGRU	0.64	0.56	0.60	0.59
InCaseLawBERT + RCNN	0.63	0.58	0.60	0.60
XLNet + BiGRU (baseline)			0.78	
Best Performing Team			0.75	

Table 9: Prediction results on the test dataset, as well as the rank of our best model compared to the best team with respect to F1-score

	P	R	F1	Acc	ROUGE-2
Approach1	0.47	0.48	0.47	0.48	0.034
Approach2	0.47	0.48	0.47	0.48	**0.040**
Best Performing Team					0.047
Baseline					0.303

Table 10: Explanation results of the proposed approaches on the test set, as well as the rank of our best model compared to the best team with respect to ROUGE score

6 Conclusion and Future Work

Our work performed well on the LegalEval 2023 shared tasks, placing 15th, 6th, 3rd, and 10th on subtasks A, B, C-1, and C-2, respectively. In Subtask A, we created a hierarchical BiLSTM-CRF model with randomly initiated embeddings. In Sub-task B, we created a BERT-based named entity recognition model. Finally, in Sub-Task C-1 and C-2, we use CaseLawBERT to create judgement prediction and explanation models.

The results demonstrate the feasibility of using deep learning models and domain-specific contextualized language models to understand legal texts and develop legal NLP applications. As part of future research, we plan to investigate the usefulness of combining some legal tasks, such as LegalNER and CJPE. The identification of entities within the legal text may facilitate a better interpretation of legal judgments. In addition, we will investigate the effectiveness of our models by considering new emerging AI-powered solutions such as legal prompting.

References

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of Rhetorical Roles of Sentences in Indian Legal Judgments. In *Proceedings of the 32nd International Conference on Legal Knowledge and Information Systems (JURIX)*.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anton Emelyanov and Ekaterina Artemova. 2019. Multilingual named entity recognition using pretrained embeddings, attention mechanism and NCRF. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 94–99, Florence, Italy. Association for Computational Linguistics.

Kai Hakala and Sampo Pyysalo. 2019. Biomedical named entity recognition with multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 56–61, Hong Kong, China. Association for Computational Linguistics.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022a. Named entity recognition in indian court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 184–193, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022b. 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.

Atif Khan, Qaiser Shah, M Irfan Uddin, Fasee Ullah, Abdullah Alharbi, Hashem Alyami, and Muhammad Adnan Gul. 2020. Sentence embedding based semantic clustering approach for discussion thread summarization. *Complexity*, 2020:1–11.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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.

Ashutosh Modi, Prathamesh Kalamkar, Saurabh Karn, Aman Tiwari, Abhinav Joshi, Sai Kiran Tanikella, Shouvik 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), Toronto, Canada. Association for Computational Linguistics (ACL).

Isar Nejadgholi, Renaud Bougueng Tchemeube, and Samuel Witherspoon. 2017. A semi-supervised training method for semantic search of legal facts in canadian immigration cases.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*.

Shounak Paul, Arpan Mandal, Pawan Goyal, and Saptarshi Ghosh. 2022. Pre-training transformers on indian legal text. *arXiv preprint arXiv:2209.06049*.

Nishchal Prasad, Mohand Boughanem, and Taoufiq Dkaki. 2022. Effect of hierarchical domain-specific language models and attention in the classification of decisions for legal cases. In *Proceedings of the CIRCLE (Joint Conference of the Information Retrieval Communities in Europe), Samatan, Gers, France*, pages 4–7.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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*.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

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. In *ASAIL@ICAIL*.

Ruben Winastwan. 2022. Named entity recognition with bert in pytorch. *Towards Data Science*.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. 2018. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*.

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