Automatic structuring of Legal metadata using NLP with applications in RE

Surya Kiran Suresh
Faculty of Engineering
Master of Computer science
University of Ottawa
Ottawa, Canada
ssure044@uottawa.ca

Saranya Krishnasami
Faculty of Engineering,
Master of Computer science
University of Ottawa
Ottawa, Canada
saranya.k.balaji@gmail.com

Abstract— Requirements engineering is a labor-intensive and tedious task. The significance of this project is to address the pressing issue of annotating bulky and cluttered documents. Befitting the plot, legal metadata that is verbose is considered for interpretation and analysis for this activity. This project will take advantage of pre-trained transformer-based models such as Legal-BERT and Legal-RoBERTa-Base to annotate the legal metadata into coherent units. These texts are transformed into segmented, understandable, and interpretable building blocks such as preamble, facts, ruling by the lower court, issues, argument by petitioner, argument by the respondent, analysis, statute, precedent relied on, precedent not relied on, the ratio of the decision, ruling by the present court, and none.

The sentences in the legal text are sequentially classified by their semantic content using BERT models. The labeled segments can be used to predict court judgments, or for automatic charge identification. Similarly, requirements engineering documents such as software specification documents, development methodologies, system feasibility studies, and validation techniques to elicit the requirements, etc. are structural. As a future scope, this approach we used to annotate the legal text can be coupled with transfer learning to annotate the requirement specifications metadata through automation and transform the text into segmented, understandable blocks of interpretable labels.

Keywords—Sequential Sentence classification, legal text, transfer learning to RE, Rhetorical Roles (RRs)

I. Introduction

Legal document processing is arduous as legal texts are lengthy and are usually spanning tens to hundreds of pages. Longer the documents difficult is the automatic processing as the details are spanned throughout the document. Additionally, legal texts use different lexicons and legal connotations which poses a natural challenge for NLP models.

Earlier attempts in this field focused on automation using hand-crafted features such as the sequential order of annotations/labels or linguistic cues to indicate rhetorical roles. Deep learning neural network models such as BiLSTM that outperform the hand-crafted features approach by automatically learning the features using pre-trained legal embeddings. Later with the evolution of transformer models, there is a significant contribution to the legal domain that has produced state of art results.

This activity is to experiment with annotating the legal text and aiming to see if there is a natural pattern or sequence that the model can learn from the legal documents. As is, the current transformer models are driven by data and are trained on the annotated corpus.

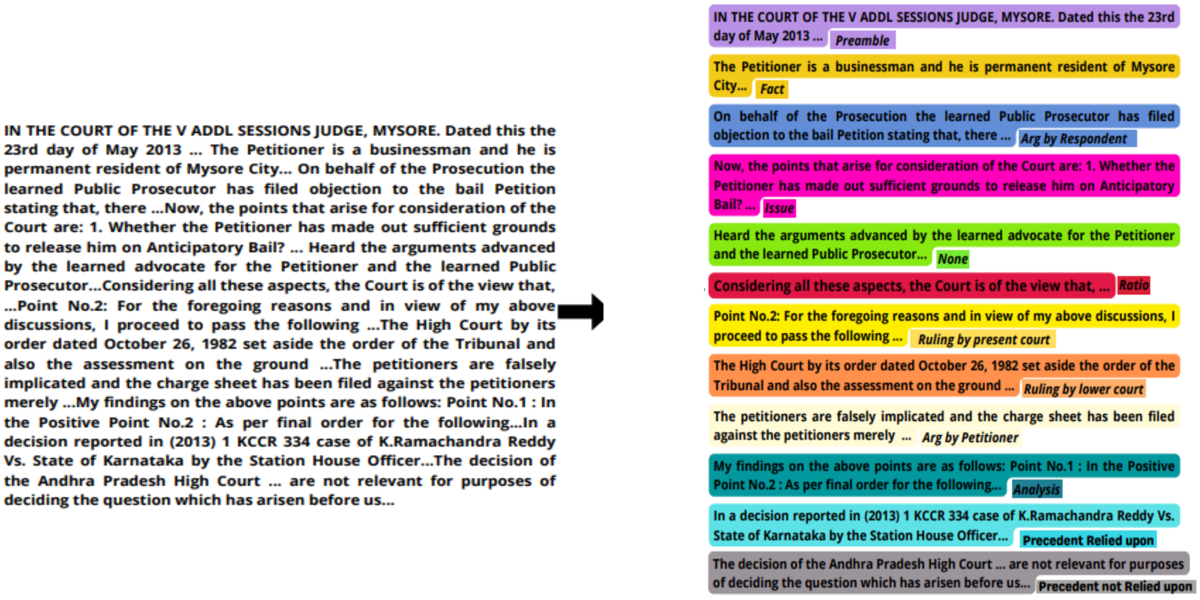
For automatically segmenting the legal documents, we experiment with the task of rhetorical role prediction from the given document and predict the text segments corresponding to various roles. Using the created corpus, we experimented extensively with various deep learning-based baseline models for the task. This will help the legal community in expediting the legal judgments and uncover the underlying pattern in court judgments which can be extended to generate automated court judgment predictions.

II. Background and Related Work

The insufficiency of rule-based reasoning in weak-theory areas is illustrated by the domain of legal reasoning. Early background work was done to aid this argument structuring such as the Araucaria project [1], which visualizes the argument structure of a text. The manual structuring of an argumentative text into a graph visualization as is done in the Araucaria research was a very costly job. Projects such as the ACILA project (2006-2010) considered the practical need for automatic detection and classification of arguments in legal texts. Rhetorical role labeling was first automated by Saravanan et al. [2], where Conditional Radom Fields (CRF) developed a generic approach to perform segmentation using seven rhetorical roles. Nejadgholi et al. [3] 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. [5]. Kalamkar et al. [6] create 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. [7] serves as a precursor for transformer models.

Adding to the existing complexity, the legal domain has sub-domains (corresponding to varying laws, e.g., criminal law, income tax law) within it. Although some of the fundamental legal principles are common, the overlap between different sub-domains is low; hence systems developed on one law (e.g., income tax law) may not directly work for another law (e.g., criminal law), so there is the problem of a domain shift (Bhattacharya et al [7]; Malik et al., [8]; Kapoor et al. [9])



IN THE COURT OF THE V ADDL SESSIONS JUDGE, MYSORE. Dated this the 23rd day of May 2013 … The Petitioner is a businessman and he is permanent resident of Mysore City… On behalf of the Prosecution the learned Public Prosecutor has filed objection to the bail Petition stating that, there …Now, the points that arise for consideration of the Court are: 1. Whether the Petitioner has made out sufficient grounds to release him on Anticipatory Bail? … Heard the arguments advanced by the learned advocate for the Petitioner and the learned Public Prosecutor…Considering all these aspects, the Court is of the view that, …Point No.2: For the foregoing reasons and in view of my above discussions, I proceed to pass the following …The High Court by its order dated October 26, 1982 set aside the order of the Tribunal and also the assessment on the ground …The petitioners are falsely implicated and the charge sheet has been filed against the petitioners merely …My findings on the above points are as follows: Point No.1 : In the Positive Point No.2 : As per final order for the following…In a decision reported in (2013) 1 KCCR 334 case of K.Ramachandra Reddy Vs. State of Karnataka by the Station House Officer…The decision of the Andhra Pradesh High Court … are not relevant for purposes of deciding the question which has arisen before us…

IN THE COURT OF THE V ADDL SESSIONS JUDGE, MYSORE. Dated this the 23rd day of May 2013 … *Preamble*

The Petitioner is a businessman and he is permanent resident of Mysore City… *Fact*

On behalf of the Prosecution the learned Public Prosecutor has filed objection to the bail Petition stating that, there … *Arg by Respondent*

Now, the points that arise for consideration of the Court are: 1. Whether the Petitioner has made out sufficient grounds to release him on Anticipatory Bail? … *Issue*

Heard the arguments advanced by the learned advocate for the Petitioner and the learned Public Prosecutor… *None*

Considering all these aspects, the Court is of the view that, … *Ratio*

Point No.2: For the foregoing reasons and in view of my above discussions, I proceed to pass the following … *Ruling by present court*

The High Court by its order dated October 26, 1982 set aside the order of the Tribunal and also the assessment on the ground … *Ruling by lower court*

The petitioners are falsely implicated and the charge sheet has been filed against the petitioners merely … *Arg by Petitioner*

My findings on the above points are as follows: Point No.1 : In the Positive Point No.2 : As per final order for the following… *Analysis*

In a decision reported in (2013) 1 KCCR 334 case of K.Ramachandra Reddy Vs. State of Karnataka by the Station House Officer… Precedent Relied upon

The decision of the Andhra Pradesh High Court … are not relevant for purposes of deciding the question which has arisen before us… Precedent not Relied upon

Figure 1: Example of document segmentation via Rhetorical Roles labels. On the left is an excerpt from a legal document and on the right a is document segmented and labeled with rhetorical role labels.

In recent times, there has been a lot of work in legal text processing with the evolution of transformer models. Our project uses the Legal BERT and Legal-RoBERTa-Base uncased models for the annotation and sentence classification task.

We initially provided some general background on the problem of sentence classification in the legal text. The next sections describe the data set used, followed by our approach and methods that we use for annotation, comprising data preprocessing, technical approach to the implementation, our experiments, results, and their discussion.

III. Data Set

A. *Description of the dataset*

For this use case, we will be using the dataset provided by SemEval2023[4] competition. This is a dataset that comprises 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 pre-defined rhetorical roles are:

- **Preamble (PREAMBLE):** This section covers the metadata about the judgment document. A classic judgment would start with the court name, party details, lawyers and judge names, and a summary (headnote). This portion would typically end with keywords like (JUDGMENT or ORDER). Some documents also have HEADNOTES and ACTS sections in the beginning. These are also part of the Preamble.
- **Facts (FAC):** This section corresponds to the facts of the court case. It refers to the events in chronological order that led to the case filing. This section will contain information such as (e.g., First Information Report (FIR) at a police station, filing an appeal to the Magistrate) Depositions and proceedings of the current court, and a summary of lower court proceedings.
- **Ruling by Lower Court (RLC):** Cases are usually addressed at lower courts and then appealed at higher courts. This portion will have the judgments given by the lower courts (Trial court, High Court) in line with the current appeal to the supreme court. The lower court's verdict, analysis, and the ratio behind the judgment by the lower court are annotated with this label.
- **Issues (ISSUE):** Certain judgments mention the crucial points on which the decision or verdict needs to be delivered. Issues are the legal questions framed nu the court.
- **Argument by Petitioner (ARG_PETITIONER):** Arguments put forward by the petitioner's lawyers. Precedent cases argued by the petitioner's lawyer will be present in this category. The court discusses these arguments later and belongs as to be relied/not relied upon.
- **Argument by Respondent (ARG_RESPONDENT):** This section should hold the respondent's lawyer's arguments.
- **Analysis (ANALYSIS):** These are the court's views which include the court's discussion on the evidence, facts presented, prior cases, and statutes. Discussions on the relevance of the current case, and observations from the court. It is the parent tag for three tags: PRE-RELIED, PRE-NOT RELIED, and STATUTE i.e., every statement which belongs to these three tags should also be marked as ANALYSIS.
- **Statute (STA):** This section includes the court's discussion on the established laws that can come from

* Git hub Repository link https://github.com/skiran13/csi5137_final_project

several sources: Acts, Articles, Rules, Quotations, Order, Notices, and Notifications.

- **Precedent Relied (PRE_RELIED)**: Texts in which the court discusses prior case documents, discussions, and decisions that were relied upon by the court for final decisions.
- **Precedent not Relied (PRE_NOT_RELIED)**: These are the texts in which the court discusses prior case documents, decisions, and discussions that were not relied upon by the court for final decisions.
- **Ratio of the decision (RATIO)**: This includes the primary reason given for the application of any legal principle to the legal issue. This section appears right before the final decision by the court.
- **Ruling by the Present Court (RPC)**: Final decision of the court with a conclusion and court order following a natural logical outcome of the rationale.
- **NONE**: Any sentence that does not belong to any of the above categories will be labeled as NONE.

These annotations are used to capture the metadata from the legal text and make it easier for requirements engineering. The data is already separated for both train and dev. The training dataset contains 247 Court Judgements that are annotated. The data is in JSON format and is formatted the following way:

- **id**: a unique id for this data point. This is useful for evaluation.
- **annotations**: (list of dict) The items in the dict are:
    - **result:** A list of dictionaries containing sentence text and corresponding labels pair. The keys are:
        - **id**: unique id of each sentence
        - **value**: a dictionary with the following keys:
            - **start**: (integer) starting index of the text
            - **end**: (integer) end index of the text
            - **text**: (string) The actual text of the sentence
            - **labels**: (list) the labels that correspond to the text
- **data**: the actual text of the judgment.

meta: (a string) It talks about the category of the case (Criminal, Tax)

IV. TECHNICAL APPROACH

For this activity, we aim to resolve the problem of reducing the complexity of decoding the long text of legal data into structured labels. To facilitate the labeling of the legal text and to capture the metadata transformer-based models are used and perform Sequential Sentence Classification. Processing of Legal documents being an active research field caters to a lot of transformer-based models that have been trained on the legal text and can be used for fine-tuning. Text preprocessing is done using general NLP techniques and then the classification task is performed using Transformer models such as Legal BERT and Legal-RoBERTa. These models are fine-tuned on the given dataset to reduce the loss. The given dataset needs to be tokenized and padded to be fed into transformer-based models.

As remarked by Hugging Face, Legal-Bert and Legal-Roberta-base are lightweight models pre-trained from scratch on legal data, which achieve comparable performance to larger models, while being much more efficient (approximately 4 times faster) with a smaller environmental footprint.

In recent times, there has been a lot of work in legal text processing with the evolution of transformer models. In this project, we intend to perform a survey study of using two base models for RR prediction and then compare their performance.

- The goal of Sequential Sentence Classification (SSC) is to classify each sentence in a sequence. The technical approach to SSC based on BERT is to encode sentences in context.
- The BERT model architecture consisting of multiple layers of Transformers uses a specific input representation, with two special tokens, [CLS] and [SEP], added at the beginning of the input sentence pair and between the sentences.
- The multi-layer transformer architecture allows the BERT models to contextualize the input over the entire sequence permitting it to capture the required information for the correct sentence classification
- Given the sequence of sentences S = (S1, ..., Sn) we concatenate the first sentence with BERT's delimiter, [SEP], and repeat this process for each sentence, forming a large sequence containing all tokens from all sentences. After inserting the standard [CLS] token at the beginning of this sequence, we feed it into BERT. BERT intuitively learns the sentence structure and the correlation between continuous sentences.
- Then the model is fine-tuned on task-specific data. During finetuning the model learns appropriate weights for the [SEP] token to allow it to capture contextual information for classifying sentences in the sequence.

The next sections describe about the data set used, the technical approach to the implementation, our experiments, results, and their discussion.

V. IMPLEMENTATION

To enable automatic segmenting of the legal documents, we experiment with the task of rhetorical role prediction given a document and predict the text segments corresponding to various roles. Using the created corpus, we experiment with two text classification and baseline models for the task.

In this project, we intend to perform a survey study of using two base models for RR prediction and then compare their performance.

- LEGAL-BERT, a family of BERT models intended to assist legal NLP research. We have taken the Legal-BERT uncased version. Another variation for comparison is LEGAL-ROBERTA, which is a domain-specific language representation model fine-tuned on large-scale legal corpora
- The pre-trained base models are taken and fine-tuned on the dataset using hold out method
- Then the labeling and sentence classification task is performed on the legal text

The model performance is compared for the models and the evaluation is done for the same. Legal-BERT and Legal-RoBERTa-Base models that are pre-trained and fine-tuned are attempted to leverage unlabeled data into sequentially annotated segments.

With the challenges explained earlier, the transformer models can't handle large texts because of the token length. If the number of tokens in a piece of text is longer than the max token length of the BERT model being used the classification cannot happen in a single go.

The implementation of the models is presented in the Git hub repository*.

A. *Preprocessing*

Raw text data might contain unwanted or unimportant text due to which our results might not give efficient accuracy and might make it hard to understand and analyze. While dealing with legal texts, there is a need for extensive preprocessing with the presence of noise (Malik et al. [8]; Kapoor et al. [9]). During pre-processing of the escape sequence tags(\n), multiple empty spaces were removed. The text was converted to lowercase as the models used were the uncased variants and text case was not a necessary feature.

B. *Approach*

Task Definition: Given 'D' a legal document, with sentences [s_1, s_2, ...s_n], the task of rhetorical role prediction is to predict y_i the label (or role) for each sentence $s_i \in$ D.

For the Sequential Sentence Classification task, the BERT model architecture consists of multiple layers of Transformers and uses a special input representation, with two special tokens, [CLS] and [SEP], added at the beginning of the input sentence pair and between the sentences (or bag of sentences) respectively.

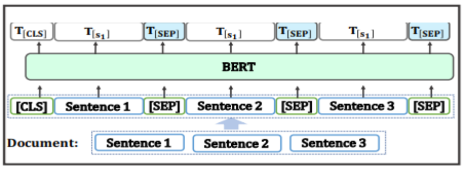


Figure 2: Overview of our model. Each [SEP] token is mapped to a contextualized representation of its sentence and then used to predict a label for the sentence.

C. *Pretraining*

Legal Bert: This is a lightweight pre-trained model on legal data, achieving comparable performance to larger models and being more efficient (approximately 4 times faster) with a smaller environmental footprint. This model is having 12-layers, 768 hidden layers, 12 heads, and 110M parameters. The pretrain configuration used for Legal-Bert is learning rate at 2e-5, batch size at 8, epochs at 5, and weight decay at 0.01. The pretraining was done on 194 annotated legal documents and 84 documents were retained for testing.

Legal_RoBERTa-Base: This model is built from the ROBERTA-BASE model and fine-tuned on the legal corpus. The fine-tuning configurations are learning rate 5e^{-5} (with learning rate decay, ends at 4.95e-8), epoch at 3, total steps at 446500, with loss starting at 1.850 and ending at 0.880. The noted perplexity after fine-tuning on legal corpus was 2.2735.

VI. Empirical evaluation

A. *Research questions*

The project primarily aims at answering the below research questions,

1. How efficiently can we annotate the domain-specific corpus using sequential sentence classification?
2. Is the annotated multiple classes domain representative?
3. Has the annotation activity contributed to the RE literature?
4. Does this annotation activity help in understanding the requirements specification promptly?

B. *Results and discussion*

The performance of pre-trained Bert models was tested on the test(in-domain) data and results are tabulated in Table 1.0. We use the standard F1 score metric for evaluation.

Step	Training Loss	Validation Loss	F1
1000	1.4111	1.318136	0.589281
2000	1.1484	1.188574	0.617505
3000	0.998	1.234505	0.61893
4000	0.8983	1.247883	0.624442
5000	0.9201	1.196241	0.622731
6000	0.7114	1.313335	0.61893
7000	0.683	1.33412	0.624062
8000	0.6848	1.349922	0.622256
9000	0.495	1.532845	0.614749
10000	0.4985	1.498666	0.6195
11000	0.4026	1.633226	0.623396
12000	0.3975	1.687387	0.617314
13000	0.3374	1.726642	0.618835

Table 1.0 Training and Validation Loss for Legal-bert

* Git hub Repository link https://github.com/skiran13/csi5137_final_project

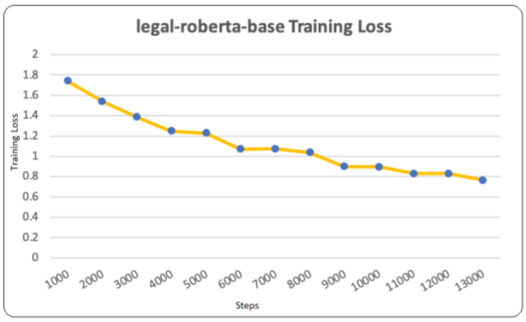
The graphs plotted for the F1, validation loss, and Training loss for the models are given below. The graph depicts that the models tend to overfit as there is a tradeoff between decreasing loss and validation increasing.

During the training phase, when the loss is low and stable training can be halted, this is usually known as early stopping. Early stopping is one of the many approaches used to prevent overfitting. Further, for Legal Bert, a checkpoint at 8000 was fixed and a checkpoint at 10000 was fixed for Legal-Roberta-Base. This is because the models tend to overfit as there is a tradeoff between decreasing loss and validation increasing.

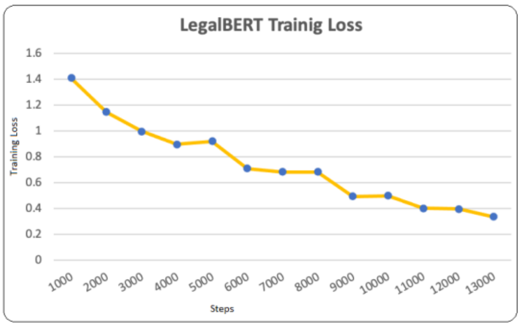
Step	Training Loss	Validation Loss	F1
1000	1.7415	1.684011	0.450442
2000	1.5423	1.493979	0.530172
3000	1.3856	1.417865	0.546517
4000	1.249	1.422216	0.55526
5000	1.2292	1.341946	0.576642
6000	1.0712	1.374896	0.572175
7000	1.0733	1.392038	0.568754
8000	1.0374	1.353413	0.577117
9000	0.8995	1.407402	0.576737
10000	0.8953	1.399892	0.579778
11000	0.8281	1.467968	0.57208
12000	0.8296	1.473147	0.569229
13000	0.7635	1.467962	0.578447

TABLE 2.0 TRAINING AND VALIDATION LOSS FOR LEGAL-ROBERTA-BASE

Training Loss: Training loss is calculated by taking the sum of errors for each instance in the training set and is a metric to calculate how the model fits the training data. From the graph, we can notice that the loss gradually decreases with the increase in steps for both models.

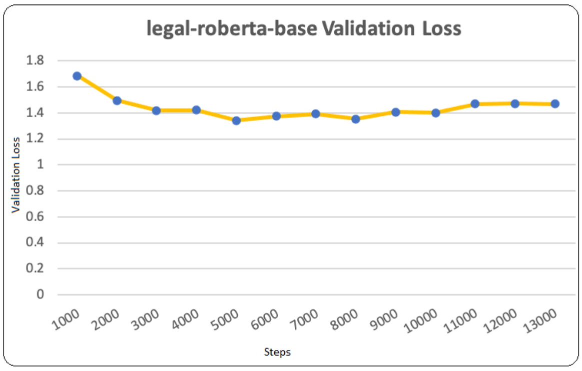


a) Training loss plotted for Legal-Roberta-Base

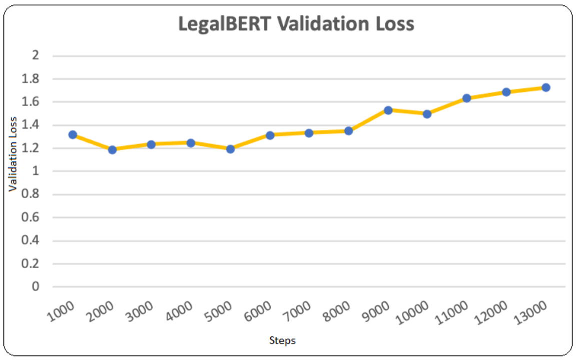


b) Training loss plotted for LegalBERT

Validation Loss: Validation loss is the evaluation metric used to assess the model performance on the validation set.

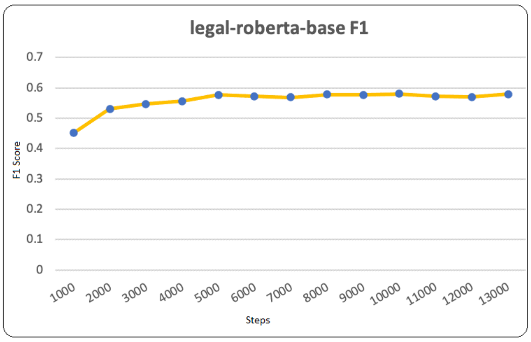


c) Validation loss plotted for Legal-RoBERTa-Base

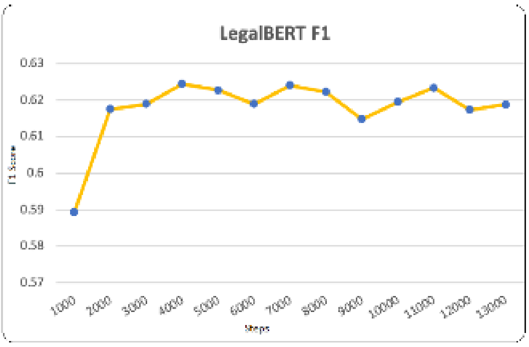


d) Validation loss plotted for LegalBERT

F1 Score:



e) F1 Metrics plotted for Legal-RoBERTa-Base



f) F1 metric plotted for LegalBERT

Figure 3: Training, validation, and F1 plots for Legal-BERT and Legal-RoBERTa-Base

VII. TRANSFER LEARNING

Transfer learning enables a target task to take advantage of the knowledge from another source task to achieve better prediction accuracy. The tasks can have training data from different domains and vary in their objectives. Fine-tuning a pre-trained language model is a popular approach for sequential transfer learning in NLP [12, 11]. Here, the source task involves learning a language model (or a variant of it) using a large unlabeled text corpus. Then, the model parameters are fine-tuned with labeled data of the target task. Pruksachatkun et al. [13] improve these language models by intermediate task transfer learning where a language model is fine-tuned on a data-rich intermediate task before fine-tuning on the final target task.

Experiments [14] suggest that if the classes of the different dataset annotation schemes are semantically related, even though the datasets come from different domains and have different text types (e.g., abstract, or full papers). This semantic relatedness is an important prerequisite for transfer learning in Natural Language Processing tasks [15, 16]

VIII. LESSONS LEARNT

Hands-on experience with Hugging Face transformers and learned to use Trainer API to fine-tune models. Found that higher epochs do not necessarily have to improve the scores.

IX. CONCLUSION

The main aim of the project is to annotate the legal text and sequentially classify the legal documents. The experiments show a survey report comparing two pre-trained models for the classification task, comparing the results and identifying the scope for improvement.

The experiment has yielded a good score of up to 62% in successfully labeling the legal texts thus answering the research questions to effectively label Legal text using transformer models. For the dataset, we have referred to the Sem Eval dataset, pre-processing is done on the legal text and the models are fine-tuned to perform the classification. The annotation will surely aid in judgments also one could extract the appropriate portions of the case that contributes towards the final decision.

Finally with transfer learning strategies such as Sequential Transfer Learning the task can be extended to requirement engineering artifacts.

ACKNOWLEDGMENT

We would like to thank professor Mehrdad Sabetzadeh for the continuous support during the course that motivated us toward the project.

* Git hub Repository link https://github.com/skiran13/csi5137_final_project

REFERENCES

[1] C. Reed and G. Rowe. Araucaria: Software for argument analysis, diagramming, and representation. International Journal of AI Tools, (14 (3 - 4)):961 – 980, 2004

[2] 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.

[3] Isar Nejadgholi, Renaud Bougueng, and Samuel Witherspoon. 2017. A semi-supervised training method for semantic search of legal facts in Canadian immigration cases. In JURIX, pages 125–134.K. Elissa, "Title of paper if known," unpublished.

[4] https://github.com/Legal-NLP-EkStep/rhetorical-role-baseline

[5] 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.

[6] 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.

[7] Paheli Bhattacharya, Shounak Paul, Kripabandhu Ghosh, Saptarshi Ghosh, and Adam Wyner. 2019. Identification of rhetorical roles of sentences in Indian legal judgments. CoRR, abs/1911.05405.

[8] 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.

[9] Arnav Kapoor, Mudit Dhawan, Anmol Goel, Arjun T H, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, and Ashutosh Modi. 2022. HLDC: Hindi legal documents corpus. In Findings of the Association for Computational Linguistics: ACL 2022, pages 3521–3536, Dublin, Ireland. Association for Computational Linguistics.

[10] A. Sleimi, N. Sannier, M. Sabetzadeh, L. Briand and J. Dann, "Automated Extraction of Semantic Legal Metadata using Natural Language Processing," *2018 IEEE 26th International Requirements Engineering Conference (RE)*, 2018, pp. 124-135, doi: 10.1109/RE.2018.00022.

[11] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[12] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[13] Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work?. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (Eds.). Association for Computational Linguistics, 5231–5247. https://doi.org/10.18653/v1/2020.aclmain.467

[14] A. Brack, A. Hoppe, P. Buschermöhle and R. Ewerth, "Cross-Domain Multi-Task Learning for Sequential Sentence Classification in Research Papers," 2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2022, pp. 1-13.

[15] Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How Transferable are Neural Networks in NLP Applications?. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 479–489. https://doi.org/10.18653/v1/d16-1046

[16] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. IEEE Trans. Knowl. Data Eng. 22, 10 (2010), 1345–1359. https://doi.org/10.1109/TKDE.2009.191