Transformers at SemEval-2024 Task 5: Legal Argument Reasoning Task in Civil Procedure using RoBERTa

Kriti Singhal[1], **Jatin Bedi**[2]
Computer Science and Engineering Department,
Thapar Institute of Engineering and Technology, India
[1]kritisinghal711@gmail.com, [2]jatin.bedi@thapar.edu

Abstract

Legal argument reasoning task in civil procedure is a new NLP task utilizing a dataset from the domain of the U.S. civil procedure. The task aims at identifying whether the solution to a question in the legal domain is correct or not. This paper describes the team "Transformers" submission to the Legal Argument Reasoning Task in Civil Procedure shared task at SemEval-2024 Task 5. We use a BERT-based architecture for the shared task. The highest F1-score score and accuracy achieved was 0.6172 and 0.6531 respectively. We secured the 13th rank in the Legal Argument Reasoning Task in Civil Procedure shared task.

1 Introduction

Mastering the art of arguing a legal case is essential for lawyers. This necessitates deep knowledge of the particular area of law along with advanced reasoning capabilities, including drawing similarities and differences. Researchers have made significant efforts towards setting the benchmark models for the new Natural Language Processing (NLP) problems in the domain of legal language understanding (Chalkidis et al., 2022).

The task, Legal Argument Reasoning Task in Civil Procedure[1] (Held and Habernal, 2024), organized at SemEval-2024 aimed at classifying the solution to a given problem as right or wrong.

Classifying an answer to a given question as correct or incorrect is a new NLP task. In particular, in the legal domain limited number of publicly available corpora exist. This contributes to added difficulty of this task (Fawei et al., 2016).

Recent advances in the field of NLP have addressed various issues, such as long texts and under-resourced domains. These include Long Short Term Memory (Hochreiter and Schmidhuber, 1997), and Gated Recurrent Units (Chung et al., 2014). But transformers (Vaswani et al., 2017) have taken the performance to new heights which were not possible earlier.

In the past, various efforts have been made to perform domain-specific adaption of different existing techniques and models. Some of these adaptions include SciBERT which was pre-trained for scientific texts, specifically in the bio-medical domain (Beltagy et al., 2019). Similarly, BioBERT was created with special emphasis on the bio-medical area (Lee et al., 2019).

In this paper, we discuss our use of a transformer-based model, RoBERTa, in the shared task of Legal Argument Reasoning Task in Civil Procedure at SemEval-2024.

2 Related Work

Researchers have used and explored various techniques in the past. In the work done by Beltagy et al. (2019); Lee et al. (2019), it was found that BERT-based architectures did not perform very well on problems that required specialized domain knowledge. Two possible solutions were found to address this issue. The first was to further pre-train BERT on domain-specific corpora, and the second possible solution was to pre-train BERT from scratch on domain-specific corpora (Chalkidis et al., 2020).

Lee et al. (2019) performed domain-adaption of BERT in the bio-medical domain. The experiment explored the effect of further pre-training BERT base for 470,000 steps on biomedical articles. The performance of the resulting model, BioBERT, was evaluated on biomedical datasets. This led to an improvement in performance when compared to BERT base.

Beltagy et al. (2019) proposed a family of BERT-based models, SciBERT, for scientific texts with a special focus on the bio-medical domain. Two approaches were followed for SciBERT, the first was further pre-training BERT base, and the second

[1]https://github.com/trusthlt/semeval24

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 969–972

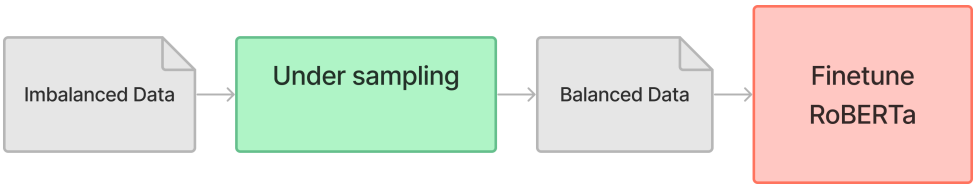

Figure 1: Proposed Methodology

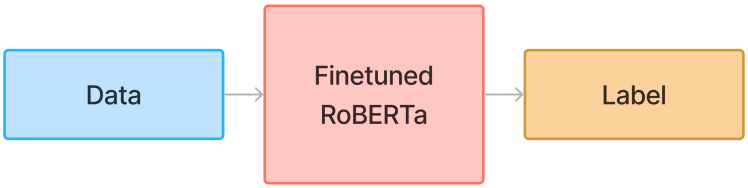

Figure 2: Label Generation for Unseen Data

approach was to pre-train BERT base on domain-specific corpora from scratch. In the second approach, random initialization of the model was performed, and a fresh new vocabulary was created. An improvement in performance was observed in the downstream tasks for both the approaches.

In the work carried out by Chalkidis et al. (2020), BERT domain adaption was performed for the legal domain. A systematic analysis was performed for the three available techniques. The first technique was to use BERT out of box, the second technique was to perform additional pre-training on BERT using domain-specific corpora, and the third approach was to perform pre-training from the start using domain-specific corpora.

3 Dataset Description

The dataset provided by the organizers was selected from the domain of the U.S. civil procedure and is based on a book aimed at law students.

In the training set, there are 666 instances, out of which 505 are labeled as 0 and 161 are labeled as 1. For each instance in the training data, there is a general introduction to a case, a question from that case, a possible argument solution along with a detailed analysis of why the argument is valid for that case. The test set, on the other hand, contains a question, answer and an explanation on the basis of which a label needs to be assigned to each instance. The assigned label will indicate whether or not the answer to the question is right or not.

4 Methodology

It was observed that in the dataset provided by the organizers, the number of instances in class 0 was 505, while the number of instances labeled as 1 was 161. Hence, in order to address the data imbalance, minority sampling was performed by randomly picking 161 instances from those labeled as class 0. This ensured that no bias existed in the trained model.

For identifying whether the answer to a given problem was correct or not, the RoBERTa Large model was employed. The RoBERTa model was designed by Facebook AI in 2019 (Liu et al., 2019). RoBERTa is a pre-trained transformer model which was trained in a self-supervised manner, i.e. only raw texts were used to train it without the involvement of human labeling.

While training the model, all the fields present in the training data, namely, question, answer, and analysis, were used to predict the provided label. The weighted Adam optimizer along with cross-entropy loss was used as the optimizer and the loss

Table 1: RoBERTa Performance Comparison

Model	F1-Score	Accuracy
RoBERTa Base	0.5511	0.6020
RoBERTa Large	0.6172	0.6531

function respectively. The learning of the optimizer was set at 1e-5. The RoBERTa model was trained for 100 epochs with the aforementioned parameters with a batch size of 8.

The training procedure has been summarised in Figure 1. The fine-tuned transformer was used to then predict the label for the unseen data as shown in Figure 2.

5 Results and Discussion

A BERT-based transformer, RoBERTa was discussed to perform categorization of an answer as right or wrong given a case, question, and a possible answer.

The data imbalance was handled by performing under sampling on the majority class instances in a random fashion. This was followed by fine-tuning the RoBERTa Large model for 100 epochs. After fine-tuning, the model achieved an F1 score of 0.5511 and an accuracy of 0.6020.

As shown in Table 1, the RoBERTa Base model performed better than RoBERTa Large, when fine-tuned for 100 epochs using the same methodology and hyper parameters. And it achieved an F1 score of 0.6172 and an accuracy of 0.6531.

Overall, we achieved the 13th rank in the Legal Argument Reasoning Task in Civil Procedure shared task at SemEval-2024 out of the 21 participating teams.

6 Conclusion and Future Work

Legal argument reasoning is a new NLP task, aimed at classifying a candidate answer as correct or incorrect given an introduction to the topic, a question and a candidate answer.

In this work, we describe our use of a BERT-based architecture, RoBERTa in the Legal Argument Reasoning Task in Civil Procedure shared task at SemEval-2024.

Ensembling techniques have shown promising results on various NLP tasks in different domains. Using an ensemble approach of different transformers may hence improve the performance. Transformers trained specifically with a focus on legal transformation such as Legal-BERT (Chalkidis et al., 2020) can improve the performance further.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. 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.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Biralatei Fawei, Adam Wyner, and Jeff Pan. 2016. Passing a USA national bar exam: a first corpus for experimentation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3373–3378, Portorož, Slovenia. European Language Resources Association (ELRA).

Lena Held and Ivan Habernal. 2024. SemEval-2024 Task 5: Argument Reasoning in Civil Procedure. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.