YNU-HPCC at SemEval-2024 Task 5: Regularized Legal-BERT for Legal Argument Reasoning Task in Civil Procedure

Peng Shi, Jin Wang and Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
Shipeng1@stu.ynu.edu.cn, {wangjin, xjzhang}@ynu.edu.cn

Abstract

This paper describes the submission of team YNU-HPCC to SemEval-2024 for Task 5: The Legal Argument Reasoning Task in Civil Procedure. The task asks candidates the topic, questions, and answers, classifying whether a given candidate's answer is correct (True) or incorrect (False). To make a sound judgment, we propose a system. This system is based on fine-tuning the Legal-BERT model that specializes in solving legal problems. Meanwhile, Regularized Dropout (R-Drop) and focal Loss were used in the model. R-Drop is used for data augmentation, and focal loss addresses data imbalances. Our system achieved relatively good results on the competition's official leaderboard. The code of this paper is available at https://github.com/YNU-PengShi/SemEval-2024-Task5.

1 Introduction

The task can be formulated as follows: given an introduction to the topic, a question, and an answer candidate, classify if the given candidate is correct (True) or incorrect (False) (Bongard et al., 2022). This task has two main difficulties: 1) The text length of the topic and question is much larger than 512 tokens. 2) The number of positive and negative samples in the data varies widely.

Initially, the online system represented the first attempt to utilize computational methods for addressing legal conundrums (VALENTE et al., 1999). Despite notable advancements in recent years, which have seen a concerted effort to establish objective benchmarks for natural language processing models in the domain of legal language comprehension (Chalkidis et al., 2022), a lack remains in the realm of complex tasks involving argumentative reasoning within legal contexts. However, Legal-BERT has emerged as a forerunner in this domain, demonstrating compelling performance (Chalkidis et al., 2020).

This paper proposes a model based on Legal-BERT. In processing tasks, we used sliding window simple (SWS) and sliding window complex (SWC) to process the original data and solved the problem of the token count of the original data being much larger than 512. In the subsequent process, we found that there was a significant imbalance in the dataset that resulted in the return of the most common label in the training set (in this case, 0). We added R-Drop (Wu et al., 2021) to the model to address this issue and changed the loss function from cross entropy to focal loss (Lin et al., 2017). In the end, we achieved a good result. The best submission for the test set has achieved 0.6166 and ranked 9th in this task.

The remainder of this paper is organized as follows. Section 2 describes the model and method used in our system, section 3 discusses the results of the experiments, and finally, the conclusions are drawn in section 4.

2 System Description

This section delves into the intricate design of the proposed model's architecture. The architecture comprises multiple essential components, namely the text cutting, the tokenizer, the pre-trained Legal-BERT model, the output layer, and the methods. Figure 1 illustrates the comprehensive system model that we have devised.

2.1 Text Preprocessing

Sliding Window Simple (SWS). The process involves dividing the combined question and introduction into discrete segments or chunks. These chunks are then submitted to a classification algorithm, which assigns a category or label to each segment based on its content. Once the classification is complete, the system calculates the average predicted output for all the chunks. This average serves as a comprehensive summary or representa-

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 757–762
June 20-21, 2024 ©2024 Association for Computational Linguistics

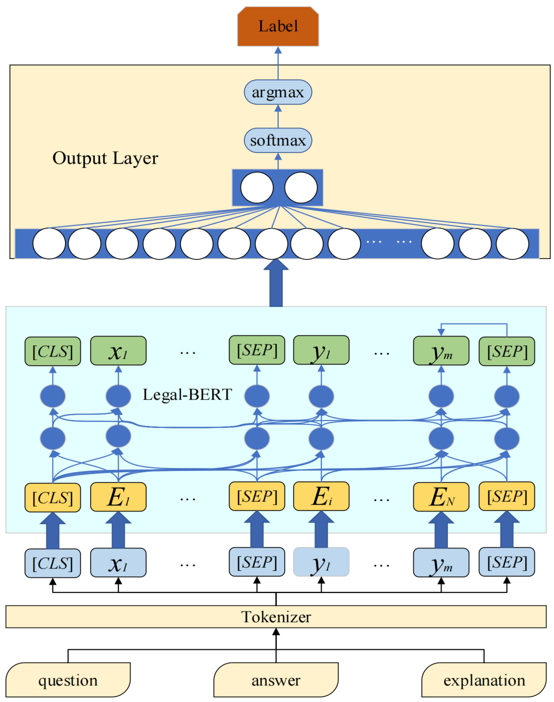

Figure 1: The structure of system

tion of the combined text, capturing the key themes and characteristics. It's a method that leverages machine learning techniques to distill the essence of a complex textual input into a single numerical value, which can be helpful in various applications such as summarization, sentiment analysis, and information retrieval.

Sliding Window Complex (SWC). In this sophisticated text processing workflow, the initial step decomposes the introductory text into discrete segments or chunks. Each chunk is meticulously constructed to include the complete question, flanked by the introduction's segments to provide context. This approach ensures that each chunk is a self-contained unit that retains the connectivity between the question and the supporting information in the introduction (Koay et al., 2021).

Subsequently, these meticulously crafted chunks are subjected to a comprehensive classification process. This process employs advanced machine learning algorithms to analyze the content of each chunk and assign it to one or more predefined categories or labels. The classification is nuanced and context-aware, considering the intricate details and subtle nuances present in the text (Kong et al., 2022).

The system employs a statistical aggregation technique to calculate the average of the predicted outputs for all the chunks. This average is a weighted sum of the individual predictions, giving more weight to chunks deemed more critical or relevant based on the specific application context.

The resulting average is a valuable metric that encapsulates the collective predictions of the model for the given question and introduction. It provides a robust summary of the model's understanding of the text, offering insights into the key themes and conclusions the model has extracted from the input. This average output can be used for various applications, such as generating summaries, making predictions, or informing decision-making processes.

2.2 Tokenizer

In many natural language processing (NLP) tasks, the original text must be processed into digital data before it can be processed by computer. Thus, the tokenizer was applied to divide the text into words and convert it into unique coding. Given a training data $\mathcal{D} = \{X^{(m)}, y^{(m)}\}_{m=1}^{M}$, $X^{(m)}$ is the processed input text. $y^{(m)}$is the corresponding ground-true label, the Bert tokenizer is applied to transform $X^{(m)}$ as,

$$X = [CLS]x_1x_2x_3...x_n[SEP]y_1y_2...y_m[SEP] \quad (1)$$

where x and y represent tokens, n and m represent the length of the first and second sentences, $[CLS]$ special mark indicates the beginning of the text sequence, $[SEP]$ indicates the separator between text sequences, respectively.

2.3 Legal-BERT Model

Legal-BERT is a specialized variant of the BERT model tailored for the legal domain, leveraging a corpus of legal text to facilitate advancements in legal natural language processing research, computational law, and legal technology applications (Chen et al., 2023). This model inherits the parameter weights from BERT-Base, ensuring a solid foundation for legal-specific tasks. In our study, we employed the pre-trained Legal-BERT model, built upon the Transformer library [1], to handle the complexities of legal language. The architecture of Legal-BERT mirrors that of the original BERT model, comprising an essential components: the Transformer encoder block (Vaswani et al., 2017). These blocks work to capture legal text's intricate

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

patterns and nuances. The model configuration used in our experiment features 12 layers, 768 dimensions, 12 self-attention heads, and a total of 109 million parameters. This configuration balances model complexity and computational efficiency, enabling us to tackle various legal NLP challenges effectively.

Encoder block. Firstly, Legal-BERT performs the embedding operation after receiving the processed raw data. Through the above processing, we obtained token embedding, segment embedding, and position embedding (Zhang et al., 2021), followed by a series of operations to obtain $\mathbf{H}$, as follows.

$$\mathbf{H} = \mathrm{Enc}(X; \theta) \tag{2}$$

where $\mathbf{H} \in \mathbb{R}^d$ is the logits with a dimensionality of 768.

2.4 Output Layer

The BERT model has two major pretraining tasks: mask language model (MLM) and next sentence prediction (NSP), and the text implication task usually uses the NSP method to predict, that is, use the hidden layer representation of $[CLS]$ bits to predict the text classification (Ma et al., 2021). In our proposed model, the output of the model is first to use a softmax function and then perform argmax on the results after softmax to obtain $\hat{y}$,

$$\hat{y} = \mathrm{argmax}(\mathrm{softmax}(W^o\mathbf{H} + h^o)) \tag{3}$$

The training objective is to optimize the focal loss between the true and predicted labels,

$$\mathcal{L}_{FL} = \begin{cases} -(1-\hat{y}^{(m)})^{\gamma}\log(\hat{y}^{(m)}) & if\ y^{(m)} = 1 \\ -\hat{y}^{(m)\gamma}\log(1-\hat{y}^{(m)}) & if\ y^{(m)} = 0 \end{cases} \tag{4}$$

where $W^o \in \mathbb{R}^d$ represents the weight of the fully connected layer, h^o represents the offset of the fully connected layer, $\mathbf{H} \in \mathbb{R}^d$ is the output representation of $[CLS]$ token in the L-th layer, γ is used to control the weight of difficult-to-classify samples, $y^{(m)}$ are respectively the true label, $\hat{y}^{(m)}$ are respectively the probability distribution of prediction.

2.5 Regularized Dropout (R-Drop)

To solve the problem of highly imbalanced data, R-Drop is added to the output layer of Legal-BERT. As shown in Figure 2, the same input can obtain two logits, $\mathbf{H}_1$ and $\mathbf{H}_2$, respectively, during the R-Drop process. Therefore, the model will output two predicted values $\hat{y}^{(1)}$ and $\hat{y}^{(2)}$, as follows.

$$\hat{y}^{(1)} = \mathrm{argmax}(\mathrm{softmax}(W^o\mathbf{H}_1 + h^o)) \tag{5}$$

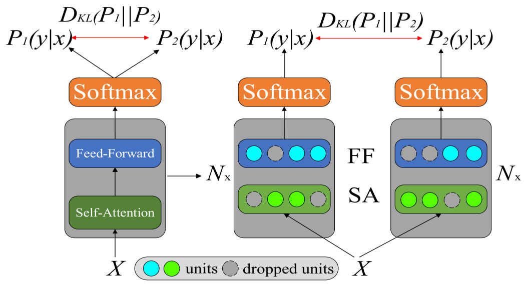

Figure 2: The structure of R-Drop

$$\hat{y}^{(2)} = \mathrm{argmax}(\mathrm{softmax}(W^o\mathbf{H}_2 + h^o)) \tag{6}$$

R-Drop uses a symmetrical Kullback-Leibler (KL) divergence to constrain $\hat{y}^{(1)}$ and $\hat{y}^{(2)}$, as follows.

$$\mathcal{L}^i{}_{KL} = \frac{1}{2}((D_{KL}(\hat{y}^{(1)}||\hat{y}^{(2)}) + D_{KL}(\hat{y}^{(2)}||\hat{y}^{(1)})) \tag{7}$$

Finally, the model will calculate the loss of two predicted values $\hat{y}^{(1)}$ and $\hat{y}^{(2)}$ using focal loss, as follows.

$$\mathcal{L}^1_{FL} = \begin{cases} -(1-\hat{y}^{(1)})^{\gamma}\log(\hat{y}^{(1)}) & if\ y^{(1)} = 1 \\ -\hat{y}^{(1)\gamma}\log(1-\hat{y}^{(1)}) & if\ y^{(1)} = 0 \end{cases} \tag{8}$$

$$\mathcal{L}^1_{FL} = \begin{cases} -(1-\hat{y}^{(2)})^{\gamma}\log(\hat{y}^{(2)}) & if\ y^{(2)} = 1 \\ -\hat{y}^{(2)\gamma}\log(1-\hat{y}^{(2)}) & if\ y^{(2)} = 0 \end{cases} \tag{9}$$

The training loss function for Legal-BERT is as follows.

$$\mathcal{L}_i = \mathcal{L}^1_{FL} + \mathcal{L}^2_{FL} + \mathcal{L}^i_{KL} \tag{10}$$

3 Experimental Result

Datasets. The Legal Argument Reasoning Task in Civil Procedure shared task data set is composed of three CSV files: the size of the training set train.csv sorted by expert comments is 666, the size of the developing set dev.csv is 84, the size of test set test.csv is 98. The data part of the train and dev set mainly includes idx, question, answer, label, analysis, complete analysis, and explanation. The data part of the test set mainly includes idx, question, answer, and explanation. Idx is used to represent the number of each sample. The question is made in the context of the content of the explanation. The answer is a candidate answer in the sample. Label indicates whether the question and candidate

Question: 8. Technical fouls. Eban brings suit against Lorenzo for interference with business relations. Eban's lawyer, Darrow, calls Lorenzo's lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?
Answer: Darrow files the complaint and mails a copy of it by certified mail to Lorenzo with two copies of a proper request for waiver. He receives the green postal receipt back in the mail. Darrow files the postal receipt with the court.
BERT Predicted label: 1
Legal-BERT Predicted label: 0
True label: 0

Question: 8. Technical fouls. Eban brings suit against Lorenzo for interference with business relations. Eban's lawyer, Darrow, calls Lorenzo's lawyer, Sadecki, and tells her that he is filing suit that afternoon. In which of the following scenarios may the case proceed without formal service of process?
Answer: Darrow files the complaint and mails a copy of it by certified mail to Lorenzo with two copies of a proper request for waiver. He does not send a summons, signed and sealed by the court, with the waiver request. He receives the signed waiver form back and files it with the court.
BERT Predicted label: 0
Legal-BERT Predicted label: 1
True label: 1

Figure 3: Examples of different models on the dev set

answer match, 0 for mismatch, and 1 for matching. Analysis and complete analysis are used for experimenters to understand why the label is 0 or 1. Explanation is used to indicate the subject of the sample to which it belongs.

Evaluation Metrics. The Legal Argument Reasoning Task in Civil Procedure shared tasks are evaluated using the standard evaluation indicators, including Macro F_1-*score* and Accuracy. The submissions of all teams are ranked according to the F_1-*score*. The metrics will be calculated as follows.

$$Precision = \frac{true\ positives}{true\ positives + false\ positives} \quad (11)$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \quad (12)$$

$$F_1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (13)$$

Implementation Details. Initially, explanation and question are concatenated when processing data. The BERT (Devlin et al., 2018) is used as the first model to solve this task. However, without any treatment, the predicted value of the BERT is all 0, and the effect is not ideal. Next, we used the larger models RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020), but the predictions and $F1$_scores were identical to BERT. Due to the extreme data imbalance, we found that the cross-entropy loss function could not calculate the loss correctly. Therefore, we changed the loss function for BERT, RoBERTa, and DeBERTa to focal loss and dice loss. The results show that modifying the loss function can slightly improve the score, but the effect is not ideal. To solve the problem of extreme data imbalance further, we change their loss functions to focal loss and dice loss (Li et al., 2020) based on supervised contrastive learning (SCL) (Khosla et al., 2020) and R-Drop. The results show that the combination of pairs can effectively solve the problem of extreme data imbalance, and the score has also been significantly improved. During the experiment, we found that due to the large number of proprietary legal terms in the data text, the above model could not fully segment professional vocabulary using the corresponding tokenizer. Therefore, we believe that the Legal-BERT is the most suitable choice. As expected, Legal-BERT has achieved good results in adding R-Drop and focal Loss technologies, as shown in Figure 3.

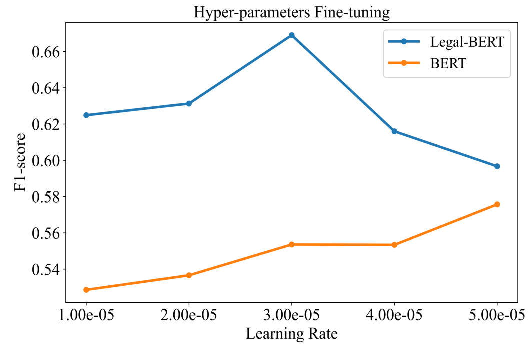

Figure 4: The performance of different learning rates on the F_1-*score*

Hyper-parameters Fine-tuning. We adjusted different learning rates and epochs to adapt to different models to achieve the expected results. Legal-BERT is better than BERT regardless of the learning rate, as shown in Figure 4. The optimal F_1-*score* was found at 4 with the batch size constantly changing, as shown in Figure 5. We set the best parameters in the final submitted results: warmup steps are 10, weight decay is 0.01, the learning rate is 3e-5, train batch size is 4, and epoch is 100.

Comparative Results and Discussion. The test is first carried out on the development set, whose size is 84. Facing the different predicted results of other models and Legal-BERT, it is clear that Legal-BERT performs better. Regardless of the

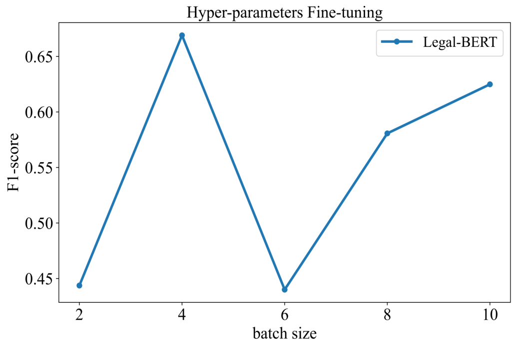

Figure 5: The performance of different batch sizes on the F_1-*score*

Model	Loss	F_1-*score*	Accuracy
BERT	Cross-Entropy	0.4437	0.7976
RoBERTa	Cross-Entropy	0.4437	0.7976
DeBERTa	Cross-Entropy	0.4437	0.7976
Legal-BERT	Cross-Entropy	0.4437	0.7976
BERT	Focal Loss	0.4688	0.8095
RoBERTa	Focal Loss	0.4437	0.7976
DeBERTa	Focal Loss	0.4956	0.7976
Legal-BERT	Focal Loss	0.5599	0.6548
BERT	Dice Loss	0.5468	0.6548
RoBERTa	Dice Loss	0.4830	0.7738
DeBERTa	Dice Loss	0.4830	0.7738
Legal-BERT	Dice Loss	0.4943	0.7421

Table 1: models and methods.

model, as long as the loss function is cross entropy, the final predicted value will be 0. Both dice loss and focal loss can solve the problem of imbalance in data, but focal loss is more effective. When SCL and R-Drop were introduced, R-Drop achieved significantly better results. Legal-BERT can deal with legal vocabulary more thoroughly than other models. Overall, Legal-BERT+R-Drop+focal Loss is the best combination obtained after experiments. The F_1-*score* obtained from the experiments of several models and methods is summarized in Table 1, Table 2, and Table 3, and the result of the best submission is shown in Table 4. Although the sliding window approach helps alleviate the token limitations of Legal-BERT, models specifically designed to handle longer documents, such as Longformer (Beltagy et al., 2020) or Big Bird (Zaheer et al., 2020), might offer superior efficiency. In the future, our team will also use the above model to solve the problem of long text.

4 Conclusion

In this research paper, we introduce a system submitted for evaluation in SemEval-2024 Task 5. Leveraging the powerful pre-trained Legal-BERT model as its foundation, our system underwent essential modifications to enhance performance. Specifically, we refined the loss function and incorporated the R-Drop technique to determine the alignment between questions and their corresponding answers accurately. The empirical results obtained from our experiments demonstrate the effectiveness of our proposed system, showcasing its strong performance capabilities. However, when benchmarked against the leading systems in the competition, it becomes evident that there are still notable areas for further improvement. Looking ahead, we are eager to explore the integration of alternative legal-specific models and innovative length text processing strategies. By pursuing these avenues, we aim to achieve even more promising results that can contribute significantly to advancing the field.

Model	Loss	F_1-*score*	Accuracy
BERT + SCL	Cross-Entropy	0.4437	0.7976
RoBERTa + SCL	Cross-Entropy	0.4437	0.7976
DeBERTa + SCL	Cross-Entropy	0.4437	0.7976
Legal-BERT + SCL	Cross-Entropy	0.4437	0.7976
BERT + SCL	Focal Loss	0.5625	0.6428
RoBERTa + SCL	Focal Loss	0.5460	0.8095
DeBERTa + SCL	Focal Loss	0.4247	0.7381
Legal-BERT + SCL	Focal Loss	0.5296	0.6706
BERT + SCL	Dice Loss	0.4892	0.7302
RoBERTa + SCL	Dice Loss	0.4437	0.7976
DeBERTa + SCL	Dice Loss	0.4437	0.7976
Legal-BERT + SCL	Dice Loss	0.5299	0.6508

Table 2: models and methods.

Model	Loss	F_1-*score*	Accuracy
BERT + R-Drop	Cross-Entropy	0.4437	0.7976
RoBERTa + R-Drop	Cross-Entropy	0.4437	0.7976
DeBERTa + R-Drop	Cross-Entropy	0.4437	0.7976
Legal-BERT + R-Drop	Cross-Entropy	0.4437	0.7976
BERT + R-Drop	Focal Loss	0.5637	0.6746
RoBERTa + R-Drop	Focal Loss	0.4437	0.7976
DeBERTa + R-Drop	Focal Loss	0.5650	0.6510
Legal-BERT + R-Drop	Focal Loss	0.6690	0.8210
BERT + R-Drop	Dice Loss	0.4824	0.6310
RoBERTa + R-Drop	Dice Loss	0.4437	0.7976
DeBERTa + R-Drop	Dice Loss	0.5155	0.6310
Legal-BERT + R-Drop	Dice Loss	0.4437	0.7976

Table 3: models and methods.

F_1-*score*	Accuracy
0.6166	0.6837

Table 4: best submission result.

Acknowledgement

The authors would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos.61966038 and 62266051.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Leonard Bongard, Lena Held, and Ivan Habernal. 2022. The legal argument reasoning task in civil procedure. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Yu Chen, You Zhang, Jin Wang, and Xuejie Zhang. 2023. YNU-HPCC at SemEval-2023 task 6: LEGAL-BERT based hierarchical BiLSTM with CRF for rhetorical roles prediction. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2075–2081, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Jia Jin Koay, Alexander Roustai, Xiaojin Dai, and Fei Liu. 2021. A sliding-window approach to automatic creation of meeting minutes. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 68–75, Online. Association for Computational Linguistics.

Jun Kong, Jin Wang, and Xuejie Zhang. 2022. Hierarchical bert with an adaptive fine-tuning strategy for document classification. *Knowledge-Based Systems*, 238:107872.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

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. *arXiv preprint arXiv:1907.11692*.

Xinge Ma, Jin Wang, and Xuejie Zhang. 2021. YNU-HPCC at SemEval-2021 task 11: Using a BERT model to extract contributions from NLP scholarly articles. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 478–484, Online. Association for Computational Linguistics.

ANDRÉ VALENTE, JOOST BREUKER, and BOB BROUWER. 1999. Legal modeling and automated reasoning with on-line. *International Journal of Human-Computer Studies*, 51(6):1079–1125.

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.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

You Zhang, Jin Wang, and Xuejie Zhang. 2021. Personalized sentiment classification of customer reviews via an interactive attributes attention model. *Knowledge-Based Systems*, 226:107135.