Research on the Application of Deep Learning-based BERT Model with Additional Pretraining and Multitask Fine-Tuning

Stanford CS224N Default Project

Muran Yu
Department of Civil and Environmental Engineering
Stanford University
yumuran@stanford.edu

Ricky Liu
Department of Management Science and Engineering
Stanford University
ricky916@stanford.edu

Abstract

In this study, we improve the performance of BERT (Bidirectional Encoder Representations from Transformers) in Natural Language Processing (NLP) tasks such as sentiment analysis, paraphrase detection, and semantic textual similarity (STS). Despite BERT's significant advancements, it faces challenges like overfitting and high computational costs. To mitigate these issues, we implemented additional pretraining and multitask fine-tuning using a miniaturized version of BERT (mini-BERT). Our approach involved pretraining on domain-specific data and multitask fine-tuning on the three downstream tasks, employing gradient surgery to manage gradient conflicts. The experimental results demonstrate substantial improvements in accuracy and correlation metrics, showcasing enhanced generalization and performance across multiple NLP tasks. However, we observed overfitting, indicating the need for further regularization. This work underscores the potential of advanced pretraining and multitask learning techniques in enhancing BERT's efficacy for diverse linguistic applications. Future directions include addressing overfitting, expanding pretraining data, and refining handling of nuanced inputs to build more robust models.

1 Key Information to include

- Mentor: Timothy Dai
- External Collaborators (if you have any): NO
- Sharing project: NO

2 Introduction

Natural Language Processing (NLP) tasks like sentiment analysis, paraphrase detection, and semantic textual similarity (STS) require a deep understanding of language nuances. BERT (Bidirectional Encoder Representations from Transformers)Devlin et al. (2019) has significantly advanced these tasks by providing contextual embeddings that improve model performance. However, BERT's reliance on large-scale pretraining and fine-tuning poses challenges, such as overfitting and high computational costs. These issues limit its generalization across diverse tasks and datasets, making further improvements necessary.

Our project addresses these challenges by implementing additional pretrainingSun et al. (2020) and multitask fine-tuningBi et al. (2022) for BERT. We used a miniaturized version of BERT (minBERT) to efficiently experiment with these techniques. By pretraining on domain-specific data and fine-tuning on sentiment analysis, paraphrase detection, and STS simultaneously, we aimed to enhance BERT's ability to generalize while managing gradient conflicts with gradient surgeryBi et al. (2022). This approach leverages shared representations across tasks, improving overall performance. Our findings demonstrate that these enhancements significantly boost BERT's effectiveness in multiple NLP applications, showcasing the potential of advanced pretraining and multitask learning methods.

3 Related Work

In recent years, significant advancements have been made in the field of AI-generated text detection using various machine learning and deep learning techniques. This section reviews the most relevant literature, focusing on methods leveraging BERT, query optimization techniques, hybrid models, analysis of BERT's inner workings, adversarial training, and fine-tuning frameworks.

Hoang et al. (2019) explored the application of BERT for aspect-based sentiment analysis (ABSA), demonstrating its potential to outperform previous state-of-the-art results on several benchmarks like SemEval-2015 and SemEval-2016. This study showcases how BERT's contextual word representations can significantly improve the accuracy of ABSA, particularly in out-of-domain settings. One key strength of this approach is leveraging BERT's pre-trained knowledge, reducing the need for extensive labeled data. However, the complexity and resource demands of fine-tuning BERT can be substantial, posing challenges in managing computational overhead.

Further advancing the capabilities of BERT, Haviv et al. (2021) presented an innovative approach to extracting knowledge from large pre-trained language models by automatically rewriting queries into a form that these models can better understand, termed "BERTese." This method optimizes queries to enhance the accuracy of knowledge extraction, which is particularly relevant to our project on AI-generated text detection. The ability to automatically optimize queries without complex paraphrasing pipelines simplifies the process and achieves higher accuracy. However, the resource-intensive nature of this approach and the need for substantial computational resources for training and fine-tuning are notable limitations.

Building on these foundational models, Rahman et al. (2024) introduced a hybrid model, RoBERTa-BiLSTM, for sentiment analysis, combining the strengths of the RoBERTa model and Bidirectional Long Short-Term Memory (BiLSTM) networks. This combination addresses several challenges in sentiment analysis, including lexical diversity and long dependencies within the text. The hybrid nature of this model is particularly relevant to our AI-generated text detection project, as it underscores the importance of leveraging both context-aware embeddings and sequential processing capabilities. However, the complexity and computational resource requirements of this approach can be barriers to its application.

A detailed analysis of BERT's inner workings by Kovaleva et al. (2019) focused on the self-attention mechanism, revealing that a limited set of attention patterns are repeated across different heads, suggesting that BERT is overparameterized. This research provides significant insights into optimizing the model's architecture, which is crucial for our project on AI-generated text detection. By identifying and potentially disabling less critical attention heads, we could reduce the computational load and improve the efficiency of our AI detection system. However, the labor-intensive manual inspection of attention maps may not be scalable for larger models or extensive datasets.

Karimi et al. (2021) introduced adversarial training into the BERT model for Aspect-Based Sentiment Analysis (ABSA), enhancing model robustness by generating perturbed examples similar to real-world data. This technique is highly relevant to our project, highlighting the potential benefits of adversarial training in improving model robustness against attempts at generating deceptive text. However, the complexity and computational demands of generating and incorporating adversarial examples require careful tuning of hyperparameters.

Finally, Jiang et al. (2020) introduced SMART, a robust and efficient fine-tuning framework for pre-trained language models, addressing overfitting and aggressive updating issues common in transfer learning. The framework combines Smoothness-Inducing Adversarial Regularization (SMART) and Bregman Proximal Point Optimization (Bregman PPO), enhancing generalization and stabilizing

Figure 1: general approach

the fine-tuning process. This approach is particularly relevant to our project, providing techniques to enhance model robustness and prevent overfitting. However, the implementation complexity and computational demands pose challenges for broader application.

Comparative studies have consistently shown that advanced NLP models leveraging BERT and its variants outperform traditional machine learning techniques in both sentiment analysis and AI-generated text detection tasks. The state-of-the-art is characterized by using advanced deep learning architectures that leverage adversarial training, hybrid models, and optimized fine-tuning frameworks. These methods achieve remarkable accuracy and efficiency, reducing the need for manual intervention and improving the reliability of automated systems. Despite these advancements, challenges remain in handling computational complexity and ensuring robustness against adversarial attacks.

By integrating insights from these various approaches, our AI-generated text detection project aims to develop a robust and efficient system capable of accurately identifying AI-generated content across diverse contexts and applications.

4 Approach

In this project, we aimed to enhance the performance of the BERT modelDevlin et al. (2019) on downstream tasks of sentiment analysis, paraphrase detection, and semantic textual similarity (STS) tasks through additional pretraining and multitask fine-tuning. We implemented a miniaturized version of BERT(minBERT), consisting of 12 transformer layers, 12 attention heads, and a hidden size of 768. The reduced size of the BERT model allows for efficient experimentation while maintaining the core capabilities of the BERT architecture. Our baseline model was the vanilla minBERT with single-task fine-tuning on each downstream task individually and no additional pretraining.

4.1 Additional Pretraining

To improve the model's understanding of language, we performed additional pretraining on domain-specific data. Inspired by the methodology in "How to Fine-Tune BERT for Text Classification?" by Sun et al.Sun et al. (2020), we explored two approaches: within-task pretraining, where BERT is further pretrained on the training data of the target task to adapt better to the specific task domain, and in-domain pretraining, where BERT is further pretrained on a large corpus from the same domain as the target task, ensuring the model captures domain-specific nuances.

For the Semantic Textual Similarity (STS) task, we used the Sentences Involving Compositional Knowledge(SICK) datasetMarelli et al. (2014) for in-domain pretraining, optimizing the mean squared error (MSE) loss between predicted and actual similarity scores. For the Paraphrase Detection task, the model was pretrained on a Quora datasetIyer et al. for within-task pretraining, learning to distinguish between paraphrase and non-paraphrase sentence pairs using binary cross-entropy loss. For Sentiment Analysis, we leveraged the Stanford Sentiment Treebank (SST) datasetSocher et al. (2013) and Cornell movie review sentiment scale datasetsPang and Lee (2005) to pretrain the model both within-task and in-domain to classify sentences into five sentiment categories using cross-entropy loss.

4.2 Multitask Fine-tuning

After additional pretraining, we fine-tuned the BERT model on three downstream tasks simultaneously: sentiment analysis, paraphrase detection, and semantic textual similarity. This multitask approach leverages shared representations to improve generalization across tasks. Inspired by "MTRec: Multi-Task Learning over BERT for News Recommendation" by Bi et al.Bi et al. (2022), we implemented a multitask learning framework that fine-tunes minBERT on multiple downstream tasks simultaneously. To handle gradient conflicts among different tasks, we employed a gradient surgery technique Bi et al. (2022), which modifies the gradients to reduce conflicts and improve training efficiency. Gradient Surgery(GS) projects the gradient of the i-th task g_i onto the normal plane of another conflicting task's gradient g_j.

$$g_i = g_i - \frac{(g_j \cdot g_i)}{||g_j||^2} \cdot g_j \tag{1}$$

To better boost the performance of main task, the modified gradient surgery combined the two auxiliary task into g_{aux}, with λ set to 0.3. Then Applies Gradient Surgery between the main task and the merged auxiliary task gradients.

$$g_{aux} = \lambda(g_{paraphrase} + g_{STS}) \tag{2}$$

The overall loss function for multitask fine-tuning combined the losses from the main tasks and the auxiliary tasks.

$$L_{MT} = L_{sentiment} + L_{paraphrase} + L_{semantic} \tag{3}$$

The combined loss function is designed to simultaneously optimize the main task of Sentiment Analysis along with two auxiliary tasks: Paraphrase Detection and Semantic Textual Similarity. This multi-task learning approach helps the model capture a more comprehensive representation of the datasets by leveraging additional information.

4.3 Model Architecture and Training Procedure

The core of our model is a multitask BERT model, which we designed to handle three distinct NLP tasks: sentiment analysis, paraphrase detection, and semantic textual similarity. The primary components of our model include:

The BERT embedding layer, which provides rich contextual embeddings for input sentences, generated using the pre-trained minBERT model. To prevent overfitting, a dropout layer with a dropout probability of 0.4(default value is 0.3) is applied to the pooled output from BERT before it is passed to the task-specific heads.

For the task-specific heads, we implemented three different layers: the sentiment classification head, which is a linear layer that maps BERT embeddings to five sentiment classes, ranging from negative to positive(score from 0 to 4); the paraphrase detection head, which is a linear layer that takes concatenated embeddings of two input sentences and produces a single logit indicating whether the sentences are paraphrases(binary); and the semantic textual similarity head, which is a linear layer that takes concatenated embeddings of two input sentences and outputs a similarity score(from 0 to 5).

Our training process consisted of several critical steps designed to optimize performance across all tasks while managing potential conflicts between them. Data preparation involved creating data loaders for each task, ensuring proper batching and shuffling to facilitate efficient training.

In the multitask training loop, for each batch, the model computed predictions for sentiment classification, paraphrase detection, and STS tasks. We calculated the loss for each task separately: cross-entropy loss for sentiment classification, binary cross-entropy loss for paraphrase detection, and MSE loss for STS. Gradients were computed for each task, and gradient surgery was applied to resolve conflicts, projecting the gradients to reduce interference and improve learning. Model parameters were updated using the AdamW optimizer with a learning rate of 1×10^{-5} and a weight decay of 0.01. After each epoch, the model's performance was evaluated on the development sets for all tasks, and the model with the highest development accuracy was saved as a checkpoint for further testing and evaluation.

4.4 Data Handling and Preprocessing

Our data handling and preprocessing are facilitated by several custom Dataset classes, tailored to the specific needs of each task.

The SentenceClassificationDataset handles data for sentiment classification, including tokenization and padding. The SentencePairDataset is used for paraphrase detection and STS tasks, managing tokenization and padding of sentence pairs. The SICKDataset is modified from SentencePairDataset class to adapt to different structure of datasets. It manages the SICK dataset for the STS task, handling tokenization and loading of sentence pairs. The SentenceClassificationTestDataset and SentencePairTestDataset are similar to their training counterparts but are adapted for test data without labels. The implementation of these classes ensures efficient data loading and preprocessing, crucial for training our multitask BERT model.

By leveraging advanced techniques such as additional pretraining, gradient surgery, and careful management of hyperparameters, our approach aims to achieve robust performance across multiple NLP tasks. The integration of these components in our multitask BERT model allows for efficient and effective multitask learning, enhancing the model's capability to generalize across diverse linguistic tasks.

5 Experiments

5.1 Data

- Stanford Sentiment Treebank (SST) DatasetSocher et al. (2013): The SST consists of 11,855 single sentences from movie reviews, labeled as negative, somewhat negative, neutral, somewhat positive, or positive. This dataset was used for the sentiment analysis task.
- Cornell Movie Review DatasetPang and Lee (2005): The Cornell Movie Review Dataset consists of 5,009 single sentences from movie reviews, labeled from 0 (most negative) to 4 (most positive) stars. This dataset was used for the sentiment analysis task.
- Quora DatasetIyer et al.: The Quora dataset contains 404,298 question pairs with labels indicating whether the pairs are paraphrases of each other. This dataset was used for the paraphrase detection task.
- SemEval STS Benchmark DatasetAgirre et al. (2013): The SemEval STS Benchmark dataset consists of 8,628 different sentence pairs with scaled similarity scores ranging from 0 (unrelated) to 5 (equivalent meaning). This dataset was used for the semantic textual similarity task.
- Sentences Involving Compositional Knowledge(SICK) DatasetMarelli et al. (2014): The SICK dataset is built starting from two existing paraphrase sets: the 8K ImageFlickr data set and the SEMEVAL-2012 Semantic Textual Similarity Video Descriptions dataset. Each sentence pair is annotated for relatedness in meaning and for the entailment relation between the two elements. It consists of 9,930 different sentence pairs with scaled similarity scores ranging from 0 (unrelated) to 5 (equivalent meaning). This dataset was used for the semantic textual similarity task.

5.2 Evaluation method

To evaluate the performance of our model on the three downstream tasks, we employed different metrics tailored to each task:

For Sentiment Analysis and Paraphrase Detection, we used accuracy as the evaluation metric. Accuracy measures the proportion of correct predictions made by the model out of all predictions.

For Semantic Textual Similarity, we used Pearson correlation coefficient and Spearman's rank correlation coefficientAgirre et al. (2013). These metrics evaluate the degree of correlation between the predicted similarity scores and the ground truth similarity scores, ensuring a comprehensive evaluation of model performance.

5.3 Experimental details

For our baseline model, we used the vanilla minBERT, which consists of 12 transformer layers, 12 attention heads, and a hidden size of 768. The additional pretraining involved using the datasets from the three downstream tasks to pretrain the model. The pretraining took approximately 1 hour. The training was conducted on NVIDIA T4 GPUs.

The multitask fine-tuning involved training the model on the sentiment analysis, paraphrase detection, and STS tasks simultaneously. Each epoch of multitask training took approximately 30 minutes. We used the AdamW optimizer with a learning rate of 1×10^{-5} and a weight decay of 0.01. We applied gradient surgery to manage gradient conflicts between tasks, ensuring effective multitask learning, and λ is set to 0.3. Bi et al. (2022)

5.4 Results

We compared the performance of the vanilla minBERT model with our modified BERT model that includes additional pretraining and multitask fine-tuning. The results are summarized in the table below:

Model	SA Accuracy	PD Accuracy	STS Corr
Vanilla minBERT	0.353	0.598	-0.141
Modified BERT	0.530	0.751	0.299

Table 1: Results for baseline model and modified model

The results demonstrate significant improvements in the sentiment analysis, paraphrase detection and STS tasks due to additional pretraining and multitask fine-tuning, indicating that the additional pretraining and multitask fine-tuning enhanced the model's ability to generalize across different tasks. The improvement in sentiment analysis accuracy suggests that the model benefits from shared representations learned during multitask training. And the improvement in the auxiliary tasks suggests additional training can enhance the ability of our model.

6 Analysis

In this section, we conduct a qualitative evaluation of our multitask BERT model, focusing on its performance in sentiment analysis, paraphrase detection, and semantic textual similarity tasks. We delve into the model's behavior by examining selected examples and performing error analysis.

For sentiment analysis, the model correctly classifies the sentences with words that clearly shows its polarity, like "greatest", "awful", or "fantasy". However, for a sentence such as "It helps that Lil Bow Wow ... tones down his pint-sized gangsta act to play someone who resembles a real kid," the model incorrectly classifies it as neutral rather than somewhat positive. This error suggests that the model has difficulty with sentences containing uncommon expressions. Also, when the model process the expression with mixed sentiments, it will struggles with these kinds of sentences, for example, initially negative but overall positive.

For paraphrase detection, the model's successfully shows its improved capability to understand different phrasings conveying the same meaning. However, the model still struggle with related but not equivalent questions.

For semantic textual similarity, for sentence with similar structure, the model can effectively recognizes synonyms and similar semantic content. On the other hand, although we add an additional

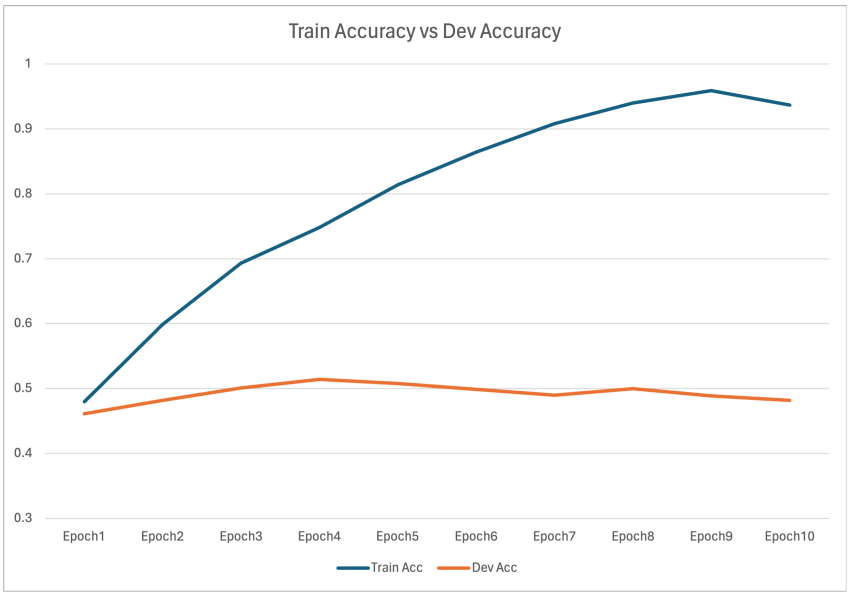

Figure 2: Train accuracy and Dev accuracy

pretraining with SICK dataset to further pretrain the model to transfer learning, the Pearson Correlation is still relatively low, showing the impact of overfitting and inability to correctly assess dissimilar content.

Figure 2 illustrates the train and dev accuracy over 10 epochs. It is evident that while the training accuracy continuously improves, the development accuracy plateaus and even slightly declines after the fourth epoch. This trend suggests that our model is overfitting to the training data. Despite the high training accuracy, the model's inability to generalize well to the dev set indicates that additional regularization techniques, such as increasing dropout or early stopping, may be necessary to improve generalization performance.

7 Conclusion

In this project, we enhanced the BERT model for sentiment analysis, paraphrase detection, and semantic textual similarity using additional pretraining and multitask fine-tuning. Implementing a miniaturized version of BERT (minBERT) and applying gradient surgery improved the model's ability to generalize across tasks.

Our modified BERT model showed significant performance improvements compared to the baseline, with higher accuracy in sentiment analysis and paraphrase detection, and better correlation in semantic textual similarity. However, the model exhibited overfitting, as indicated by the plateau in development accuracy after the fourth epoch, suggesting the need for further regularization.

Qualitative evaluation revealed the model's strengths in handling clear sentiments and synonymous paraphrases but also highlighted challenges with mixed sentiments, nuanced paraphrases, and dissimilar content in the STS task.

Overall, our work demonstrates the benefits of additional pretraining and multitask fine-tuning for improving BERT's performance on multiple NLP tasks. Future efforts should focus on enhancing regularization, expanding pretraining data, and better handling nuanced inputs to create more robust and versatile models.

8 Ethics Statement

One significant ethical challenge in this project is the potential for the BERT model is the existing biases present in the training data. For instance, when using sentiment analysis to analyze movie

reviews, the model might unfairly classify reviews written by women or people of color as more negative due to historical biases in the language and expressions used in these reviews. This issue arises because language models like BERT are trained on vast datasets that reflect societal prejudices. If not properly addressed, this could lead to biased sentiment classifications that develop stereotypes. To mitigate this risk, it is essential to implement robust bias detection. One possible approach is to employ fairness-aware training techniques, such as re-weighting the training data to reduce bias. For instance, we could use penalties to get points off those biases and then re-weight the data. These measures can help create a more inclusive and unbiased model, reducing the societal risks associated with biased AI systems. Another ethical issue concerns the environmental impact and resource allocation associated with training large-scale models like BERT. Training these models requires significant computational resources, leading to substantial energy consumption. Fine-tuning BERT on extensive datasets for sentiment analysis, paraphrase detection, and semantic textual similarity tasks can result in high energy usage, contributing to the carbon footprint of AI research and deployment. To mitigate this environmental impact, it is important to implement strategies that minimize energy consumption and optimize resource allocation. Taking an example from ourselves, we need around 6 hours to get the result and accuracy for our model and it took days for us to explore extensions and improve our model to boost accuracy. One possible approach is to employ more energy-efficient hardware. Additionally, we could use techniques such as mixed-precision training to optimize the training process. By adopting these strategies, the project can contribute to more sustainable AI practices and reduce its negative impact on the environment.

9 Contribution

Ricky Liu - Ricky Liu mainly contributed by building up the foundational framework of the Bert Model (Part 1) and fine-tuning the model to optimize its performance.

Muran Yu - Muran Yu mainly contributed by finding additional training datasets and fine-tuning the model to optimize its performance.

References

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Qiwei Bi, Jian Li, Lifeng Shang, Xin Jiang, Qun Liu, and Hanfang Yang. 2022. MTRec: Multi-task learning over BERT for news recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2663–2669, Dublin, Ireland. 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.

Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. Bertese: Learning to speak to bert.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using BERT. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196, Turku, Finland. Linköping University Electronic Press.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. First quora dataset release: Question pairs.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. Adversarial training for aspect-based sentiment analysis with bert. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8797–8803.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

Md. Mostafizer Rahman, Ariful Islam Shiplu, Yutaka Watanobe, and Md. Ashad Alam. 2024. Robertabilstm: A context-aware hybrid model for sentiment analysis.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?