Sentiment analysis of Chinese corpus (Weibo) based on BERT

Qiuyi Feng

Nov 29, 2024

Abstract

This project aims to perform sentiment analysis on Weibo's daily top 60 trending topics by collecting and analyzing the top 100 tagged posts per trend. We will compare the results from a Transformer- based sentiment analysis model with Weibo's inbuilt sentiment ratings to evaluate the model's accuracy and effectiveness. The comparison will highlight areas where the model outperforms or aligns with Weibo's assessments.

Introduction

Sentiment analysis for English and other Latin-based languages has advanced significantly, with Transformer models like BERT achieving exceptional accuracy. However, for Chinese—a language with a distinct structure—building high-accuracy models remains challenging, requiring innovative approaches.

To capture real-time public sentiment, we used data from Weibo, one of China's largest social media platforms with over 582 million monthly active users as of 2024 [1]. While Weibo provides sentiment ratings, independent models offer a valuable way to validate and enhance sentiment analysis.

In this project, we leveraged Transformer-based models, experimenting with various architectures and training strategies. These included freezing the pre-trained BERT model while modifying the linear layer, adding Transformer or LSTM layers, and a novel two-step approach where BERT is first trained and then fine-tuned with its parameters frozen. By comparing these methods, we aim to identify the most effective approach for Chinese sentiment analysis and evaluate its performance against Weibo's native system.

Related Work

The Transformer model, introduced in the seminal paper *"Attention Is All You Need"* by Vaswani et al. [2], revolutionized natural language processing by proposing a novel mechanism called the **Attention mechanism**. Central to this framework is the concept of **multi-head attention**, which splits the attention mechanism into multiple parallel "heads," enabling the model to focus on different aspects of the input sequence simultaneously.

Each attention head computes attention independently, using its own sets of query (Q), key (K), and value (V) projections. This design allows each head to capture diverse patterns or relationships within the input sequence. The outputs from all heads are then concatenated and transformed into the final attention output. This multi-head structure enhances the model's ability to represent complex dependencies and significantly improves its performance on a wide range of tasks.

Building upon the Transformer architecture, Devlin et al. introduced the **BERT (Bidirectional Encoder Representations from Transformers)** model in their paper *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"* [3]. BERT revolutionized pre-trained language models by introducing a bidirectional training approach, allowing it to capture both left and right context in text simultaneously. This was achieved through two innovative pre-training tasks: *Masked Language Modeling (MLM)* and *Next Sentence Prediction (NSP)*. BERT demonstrated state-of-the-art performance across numerous NLP tasks, such as question answering, sentiment analysis, and natural language inference, cementing its place as a foundational model in the field.

In this work, we perform fine-tuning using the *bert-base-chinese* model, which is based on the BERT architecture and is available on Hugging Face [3].

Proposed Work

Tokenization and Model Design

To begin, the dataset was tokenized using the same *bert-base-chinese* tokenization method to ensure consistency with the pre-trained model. After tokenization, the dataset was split into training and validation sets. Various fine-tuning and comparison models were then trained on the processed data.

In total, seven models were trained. For baseline comparisons, we utilized simple SVM and Random Forest models. In addition, models based on BERT were trained with various configurations, where additional layers such as linear classifiers, Transformers, and LSTMs were added after the BERT backbone. The classification task aimed to distinguish between different labels using a linear layer as the final output.

Challenges and Solutions

During training, we observed that both the vanilla BERT model and its variants with additional layers such as Transformers and LSTMs exhibited overfitting as the number of epochs increased. Specifically, while the training accuracy reached 100%, the validation accuracy decreased, and the validation loss increased significantly. This is a clear sign of overfitting. Upon investigation, we found that the model had 35,356,166 parameters, whereas the dataset size was comparatively small.

To address overfitting, we decided to freeze the pre-trained BERT model and train only the linear classification layer, which consisted of just 4,614 parameters. While this reduced overfitting and improved validation accuracy, the maximum accuracy achieved was still lower than the best results obtained when the entire model was unfrozen during the initial training phase.

To balance high accuracy and reduced overfitting, we adopted a hybrid approach. First, we trained the model with the pre-trained BERT backbone unfrozen for the initial epoch, followed by freezing the backbone for the subsequent ten epochs. This method yielded the best results, achieving a good balance between accuracy and generalization.

Datasets

Training and Validation Data

The training and validation datasets were obtained from an online source [**?**], which provides a comprehensive collection of labeled Chinese Weibo data. The dataset is divided into two categories:

- **General Weibo Dataset:** Includes 27,768 training samples, 2,000 validation samples, and 5,000 test samples.
- **COVID-19 Weibo Dataset:** Includes 8,606 training samples, 2,000 validation samples, and 3,000 test samples.

In total, the dataset comprises 45,374 comments. However, the sentiment distribution is not uniform. As shown in Figure 1, the ”angry” sentiment accounts for 30% of the samples, whereas ”fear” has the lowest representation, making up only 4.4%.

Testing Data

For additional testing, we created a custom dataset using a web scraping script to extract data from Weibo’s daily trending topics. Specifically, we extracted the top 60 trending topics each day and retrieved the 100 most popular comments for each topic. This resulted in a test set of 36,019 comments.

To evaluate the sentiment of these comments, we leveraged Weibo’s built-in sentiment analysis system, which provided sentiment labels and proportions for each trending topic. After preprocessing the scraped data, the final test set was used for further validation of our model’s performance.

Evaluation

0.1 Validation Set Evaluation

First, in my validation dataset, I choose to use metrics such as accuracy and loss for visualization.

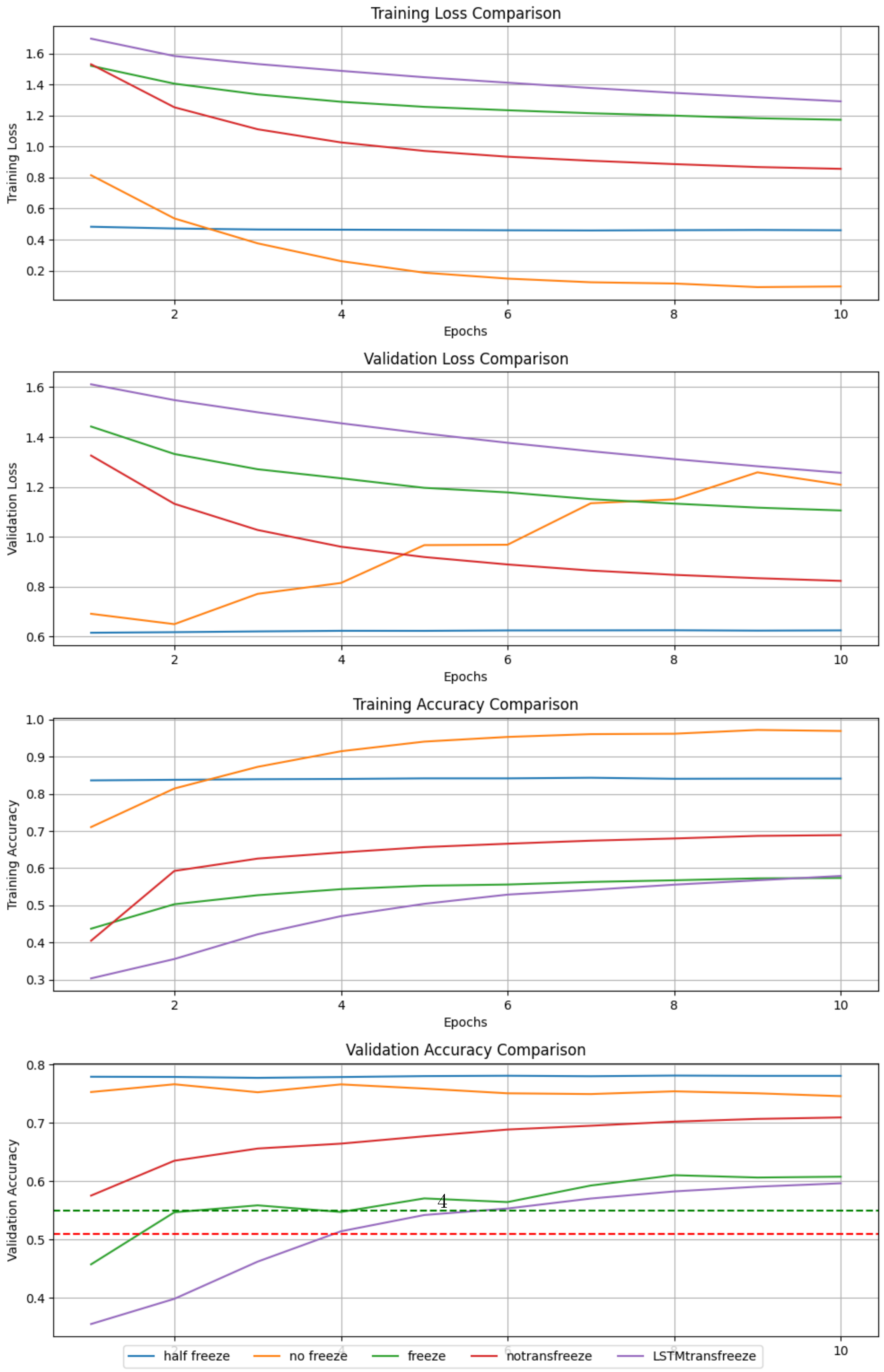

Figure 1: Training and validation accuracy and Loss

The performance of various methods across different training and validation metrics is compared in the plots. The "half freeze" method consistently achieves the lowest training and validation loss and the highest accuracy, outperforming all other methods. The "freeze" method shows stable performance but is not as effective as "half freeze." The "no freeze" method (red) performs the worst, with higher loss and lower accuracy, especially in validation. The "notransfreeze" and "LSTM-transfreeze" methods show moderate performance, improving over epochs, but neither outperforms "half freeze" or "freeze."

In the final validation accuracy, there are two additional models, which I included for comparison: the SVM and Random Forest models. Their accuracies are 0.51 and 0.55, respectively, showing a significant gap compared to the Transformer model.

0.2 Testing Set Evaluation

In the final testing phase, because our sentiment analysis table is organized by trending topics, we cannot directly categorize the comments as we did in the validation phase. Therefore, I first grouped the classification results obtained from other models by different dates and trending topics, and calculated the percentage of each sentiment. Then, I computed the MAE and MSE between these percentages and the actual data as metrics. Lastly, since each group contains six types of sentiment, resembling a six-dimensional vector, I also performed a cosine similarity comparison to obtain the final results.

Model	**MSE**	**MAE**	**MSE Rank**	**MAE Rank**
half_freeze_proportion	0.049397	0.160145	1.0	1.0
nofree_proportion	0.052746	0.163268	2.0	2.0
notransfree_proportion	0.055019	0.170411	3.0	3.0
LSTMfreeze_proportion	0.062971	0.186200	4.0	4.0
freeze_proportion	0.065944	0.187948	5.0	5.0
SVM_proportion	0.121391	0.234045	6.0	6.0
RF_proportion	0.176604	0.289264	7.0	7.0

Table 1: Comparison of model performance (MSE and MAE)

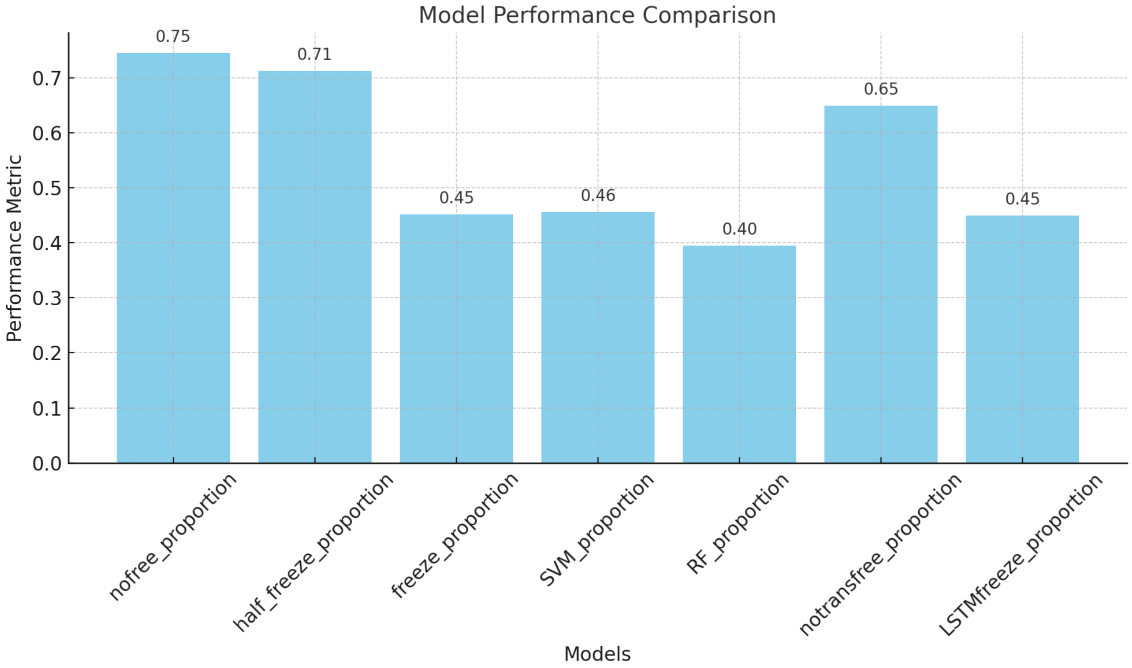

Figure 2: Cosine Similarity

We can observe that in the comparison of MAE and MSE, the results are almost identical to those of our validation phase, with the half-freeze model performing the best. However, in the cosine similarity results, the no-freeze model achieved the highest similarity.

Timeline

- **Weeks 1-2**: Literature review, data acquisition setup, and Transformer model selection.
- **Weeks 3-5**: Data collection and initial processing, finding some pre-trained model.
- **Weeks 6-7**: Model training and sentiment analysis on collected posts.
- **Weeks 8-9**: Comparative analysis between model-generated sentiments and Weibo's sentiment tags.
- **Weeks 10-11**: Fine-tuning the Transformer model and improving evaluation metrics.
- **Week 12**: Final report preparation and documentation.

Conclusion

As expected, the half-freeze model achieved the best performance in both the validation and testing phases. This is because, during the initial training, it adjusted parameters based on Weibo comments without overfitting. After freezing the model, it continued to avoid overfitting, striking a balance between preventing underfitting and avoiding overfitting.

However, there are still issues that require further exploration, such as why the cosine similarity is not as good as the no-freeze model. The dataset is too small, with fewer than 10,000 data points, while the number of parameters has already reached 1 million, making it highly susceptible to overfitting. This is an area we will focus on in our future work.

References

[1] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention is All You Need.* Advances in Neural Information Processing Systems, 2017.

[2] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

[3] This work performs fine-tuning using the *bert-base-chinese* model available on Hugging Face at https://huggingface.co/google-bert/bert-base-chinese.