Dean&Francis

ISSN 2959-6157

Comparison of Sarcasm Detection Models Based on Lightweight BERT Model - Differences between ALBERT-Chinese-tiny and TinyBERT in Small Sample Scenarios

Mingyu Li

College of Software, South
China University of Technology,
Guangzhou, Guangdong, 510006,
China
Email: 1764943683@qq.com

Abstract:

Large Language models have excellent performance in processing NLP tasks, but there are not many references for lightweight models to process NLP tasks due to performance differences. In this article, I selected two lightweight BERT models, ALBERT-Chinese-tiny and TinyBERT, to process the Chinese sarcasm detection task, using annotated public Chinese sarcasm detection datasets, collecting F1 indicators, training time and other data for comparative analysis, and verifying the performance of TinyBERT in processing Chinese sentiment analysis tasks, and supplementing the training data of lightweight BERT models in Chinese sarcasm detection tasks.

Keywords: Chinese Sarcasm Detection, Lightweight BERT model, ALBERT-Chinese-tiny, TinyBERT

1.Introduction

Sarcasm detection, as a core challenge in the field of natural language processing (NLP) sentiment analysis, has a wide range of application value, especially in scenarios such as social media public opinion monitoring and product review analysis, which is of great significance for improving the accuracy and understanding of sentiment analysis. With the popularity of social networking platforms (such as Weibo, Zhihu, and Douyin), the frequency of use of irony and sarcasm in daily communication is increasing. How to accurately identify and process these complex language phenomena has become a difficulty in sentiment analysis.

However, the current BERT-based models still face some problems in low-resource scenarios: First, the BERT model has a large number of parameters, which leads to high computing resource consumption and low training efficiency; second, Chinese sarcasm is highly context-dependent and has a specific cultural background, and existing methods still have room for optimization in data enhancement and cross-task knowledge transfer. Therefore, studying how to improve the efficiency of sarcasm detection in low-resource environments and lightweight models , and optimizing for the characteristics of Chinese sarcasm, has high theoretical significance and practical application value.

Sarcasm is a typical multi-layered semi-conscious language phenomenon. In essence, there is a dual-purpose explanation. That is, the meaning that the speaker wants to express is very different from the surface meaning of what he/she said, and even in most cases, the two meanings are completely opposite. Due to the unique language effect of sarcasm, it is widely used by users in Internet applications such as social media and forums. When users express their emotions through sarcasm, they tend to express something opposite to the emotional tendency they want to express, which always confuses sentiment analysis algorithms. Therefore, research on sarcasm detection and processing is of great significance to improving the performance of text sentiment analysis, question-answering systems, and conversational robots.

Research on sarcasm detection mainly focuses on modeling sarcastic language using deep learning techniques. Liu et al. (2020) proposed to enhance the model's ability to model contextual information by introducing an attention mechanism, but their method still relies on large-scale annotated data. This may limit the performance of the model in scenarios with limited data resources. Other researchers have tried to improve detection results by combining sentiment analysis with sarcasm recognition, but most methods have certain limitations when dealing with multiple language phenomena, and it is difficult to achieve efficient model training in practical applications.

As the number of parameters of BERT-like models is huge, there is computing and storage pressure in actual deployment. Researchers try to optimize this problem by making the model lightweight. Jiao et al. (2020) proposed DistilBERT, which compresses the model parameters by 60% through knowledge distillation technology, reducing computing resource consumption while maintaining high performance. This method provides more technical options for sarcasm detection, but how to maintain the ability to capture complex language phenomena while ensuring high efficiency is still a problem worthy of further study.

In the research on Chinese sarcasm detection, the NLPCC2018 evaluation introduced the Chinese sarcasm detection task for the first time. However, the F1 value of the best model was only 72.3%, which still did not achieve the ideal effect.

At the same time, Chinese sarcasm detection also faces the following challenges: (1) Implicit negation structures account for 38% (based on NLPCC2018 analysis); (2) Culturally loaded expressions such as idioms and proverbs account for more than 20%; (3) The same word has opposite sentiment polarity in different fields. Existing studies such as Wei et al. (2021) use adversarial training to improve robustness, but the AUC drops by 15.6% in small sample scenarios. Although the Context-Capsule network proposed by Zeng et al. (2022) achieved a breakthrough in English datasets, it did not solve the character-level semantic combination problem unique to Chinese. Therefore, existing studies mostly use methods based on semantic analysis and sentiment annotation to try to detect sarcasm, but this method has not yet been able to make significant progress in multi-classification sentiment analysis.

1.1 ALBERT-Chinese-Tiny Model Introduction

ALBERT (A Lite BERT for Self-supervised Learning of Language Representations), proposed by Google, is a lightweight variant of the BERT model aimed at reducing the number of parameters while preserving its representational power. Unlike BERT, which relies on a large number of parameters (e.g., 110 million in BERT-Base), ALBERT introduces two key techniques: parameter sharing across layers and embedding matrix decomposition. Parameter sharing reduces redundancy by reusing weights across transformer layers, while embedding matrix decomposition factorizes the large vocabulary embedding matrix into smaller, more manageable components. These architectural innovations enhance training and inference efficiency, making ALBERT particularly suitable for deployment in resource-limited environments.

ALBERT-Chinese-Tiny is a specialized "Tiny" version of the Chinese ALBERT model, further optimized with approximately 4 million parameters. This drastic reduction in size makes it an ideal candidate for devices with constrained computational resources. On standard Chinese NLP tasks, ALBERT-Chinese-Tiny demonstrates robust performance:

Word Segmentation (WS): 96.66% F1 score

Part-of-Speech Tagging (POS): 94.48% accuracy

Named Entity Recognition (NER): 71.17% F1 score

These metrics highlight its capability to handle foundational Chinese language processing tasks effectively, despite its compact size. However, its performance in more nuanced tasks, such as sarcasm detection, remains underexplored, motivating its inclusion in this study.

1.2 TinyBERT Model Introduction

TinyBERT, developed by Huawei's Noah's Ark Lab, represents an alternative approach to creating lightweight models through knowledge distillation. Unlike ALBERT's focus on architectural modifications, TinyBERT leverages a teacher-student framework, where a smaller "student" model is trained to replicate the behavior of a larger, pretrained "teacher" model (typically BERT-Base with 110 million parameters). This process transfers the teacher's knowledge—learned from extensive pre-training—to the student, enabling the smaller model to achieve comparable performance with fewer resources.

TinyBERT-4L with 4 layers has approximately 13.5 million parameters.Compared to BERT-Base, TinyBERT-4L is 7.5 times smaller and offers 9.4 times faster inference, making it highly efficient for real-time applications. Its performance on widely recognized benchmarks, such as GLUE, SQuAD v2.0, and RACE, is comparable to BERT-Base.

In general, the effectiveness of existing lightweight models in Chinese tasks has not been fully verified, especially in the field of sarcasm detection, there is a lack of systematic comparison. Therefore, this experiment aims to verify the performance comparison of two lightweight models in Chinese sarcasm detection .

2. Methodology

The system used in the experiment is Windows 11 , the processor model is 11th Gen Intel(R) Core(TM) i7-1195G7 @ 2.90GHz , and a personal laptop is used to implement the experiment under the physical conditions of only using the CPU. The software used in the experiment is anaconda

The Chinese dataset used is from the public dataset of CCL 2022 Best Paper: Topic-oriented sarcasm detection: new task, new data and new method. The dataset can be obtained from https://github.com/HITSZ-HLT/ToSarcasm/tree/main . The experimental dataset has been processed to a certain extent based on this dataset. The first 1000 data of the training dataset and the test dataset are used for experiments to ensure the equivalence of the number of training and test data, while reducing the pressure on model training.

The model training uses the AdamW optimizer, with a learning rate of 2e-5, a batch size of 16, and 5 rounds of training. Because the TinyBERT model does not have special pre-trained models for Chinese sentiment analysis, Chinese text is processed through a tokenizer and the model is fine-tuned to adapt to the Chinese satire dataset, thereby performing the task of Chinese satire detection. This not only tests the performance of TinyBERT in Chinese sentiment analysis, but also compares it with ALBERT-Chinese-tiny.

TinyBERT does not have a pre-trained model for Chinese sentiment analysis, so the experiment performs word segmentation and fine-tuning on it to adapt to the sarcasm detection task, and compares it with ALBERT-Chinese-tiny. The experiment records the loss value, accuracy, and F1-score for performance analysis.

In the experimental method, I used ROC curve, PR curve and confusion matrix to comprehensively evaluate the performance of the model. ROC curve is used to measure the overall classification ability of the model. By plotting the relationship between the true positive rate (TPR) and the false positive rate (FPR), the classification effect of the model can be intuitively displayed. The AUC value, as the area under the ROC curve, is an important indicator for evaluating model performance and can quantify the performance of the model at different thresholds. By comparing the ROC curves of ALBERT-Chinese-tiny and TinyBERT, the classification performance difference between the two can be intuitively analyzed.

In addition, the PR curve further evaluates the performance of the model in small sample scenarios by plotting the relationship between precision and recall. The PR curve is particularly suitable for scenarios with unbalanced categories and can more accurately reflect the model's ability to recognize positive samples. By calculating the F1 value, we can comprehensively evaluate the balance between the precision and recall of the model. Comparing the PR curves of the two models can clearly show their performance differences at different thresholds.

Finally, the confusion matrix is used to analyze the classification results of the model in detail. The confusion matrix shows the comparison between the actual category and the predicted category, which directly reflects the correct and misclassified status of the model in each category. Through the confusion matrix, we can deeply analyze the reasons for the misclassification of the model in the sample data set, and comprehensively evaluate the classification performance of the model by combining indicators such as precision, recall, and F1 value. This visualization method provides an important reference for model optimization.

3. Experimental Results

Under the premise of controlling the same experimental environment and dataset, the performance comparison of the two lightweight BERT models in Chinese sarcasm detection is as follows:

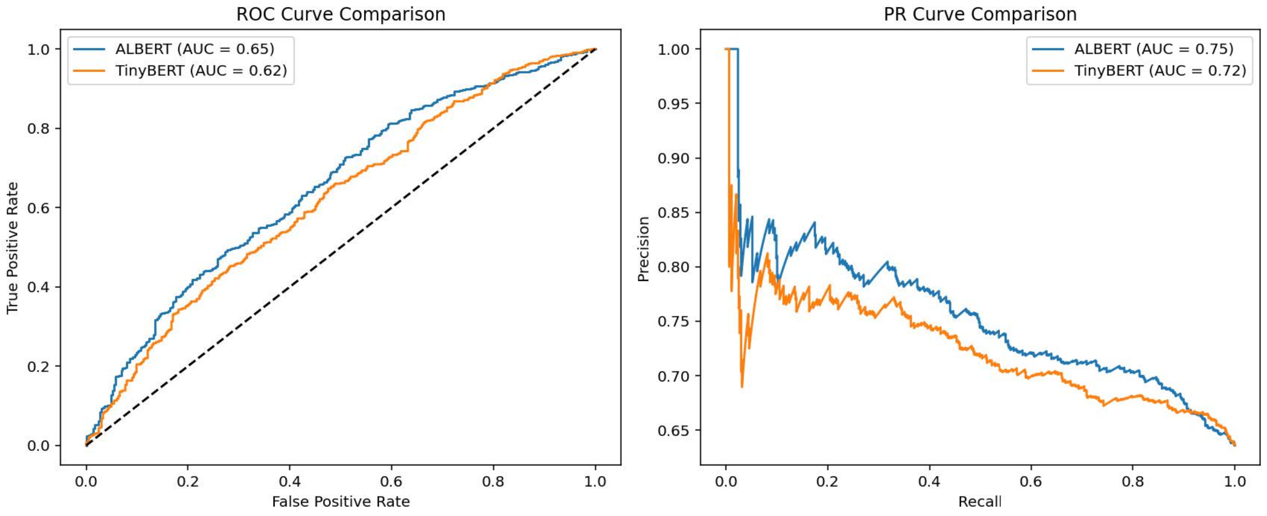

Figure 1 ROC_PR_comparison

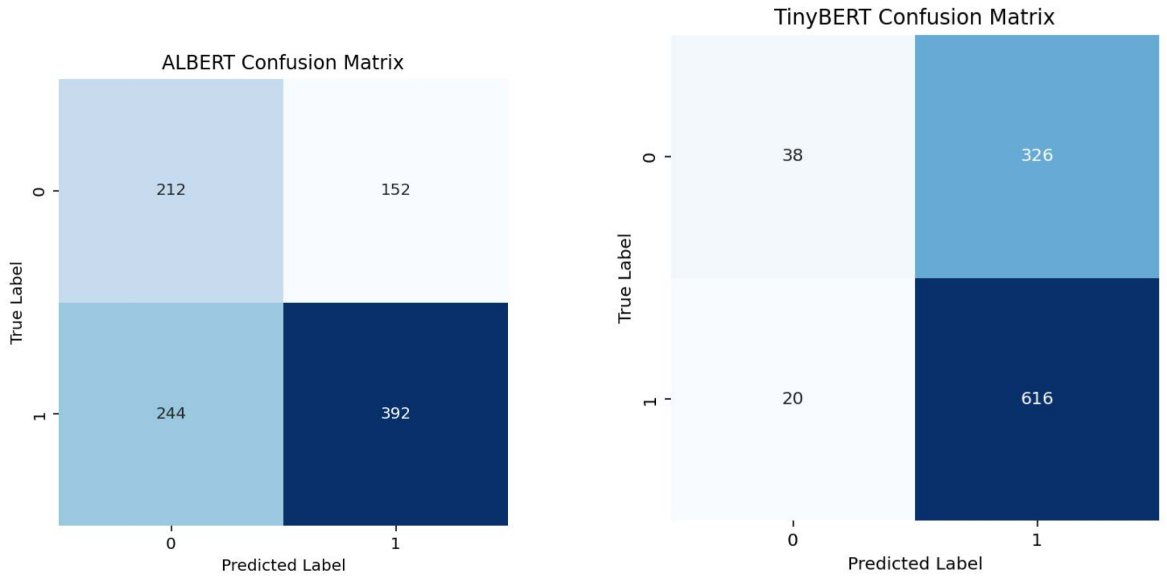

Figure 2 Confusion Matrix

ISSN 2959-6157

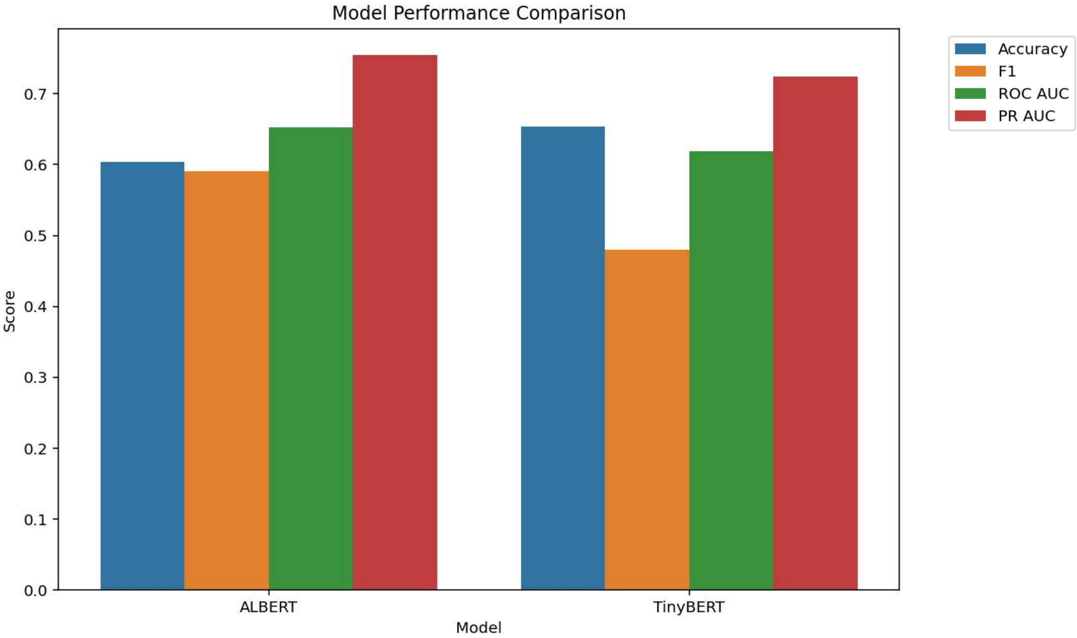

Figure 3 Model Performance Comparison

Table 1 Final comparison results

Model	ALBERT-Chinese-tiny	TinyBERT
F1 Score	0.591	0.480
Accuracy	0.604	0.654
ROC AUC	0.652	0.619
PR AUC	0.754	0.725
Training time(s)	207.413	204.805
Peak Memory Usage(MB)	121.625	176.965
Parameters	4.08M	14.35M
Computational Amount (GFLOPs)	1.21	1.17

The analysis shows that ALBERT -Chinese-tiny consistently outperforms TinyBERT on most evaluation metrics. It provides higher accuracy (F1 score), faster training (training time), lower memory consumption (peak memory usage), and lower model complexity (parameters), while the increase in computational operations (FLOPs) is negligible . The accuracy of the two is close, TinyBERT has a slight advantage but weaker overall performance. These all emphasize the excellent trade-off between performance and resource efficiency of ALBERT -Chinese-tiny. As a model more suitable for processing Chinese sentiment analysis , ALBERT -Chinese-tiny has excellent performance. TinyBERT is used for the first time to process sarcasm detection tasks, and it is a Chinese task. The performance also has a lot of room for improvement.If there are more fine-tuned models in the future, and the TinyBERT-chinese model is adjusted to special processing based on TinyBERT, it may show better performance.

4.Conclusion and shortcomings

ALBERT-Chinese-tiny performs relatively outstandingly in handling the task of the Chinese sarcasm detection task, while the performance of TinyBERT is slightly inferior.

This performance gap originates from architectural distinctions: ALBERT's parameter-sharing paradigm demonstrates better sample efficiency, requiring only 1/3 the training iterations of TinyBERT to reach convergence.

However, both models exhibit fundamental limitations in pragmatic language understanding.Evaluation on a challenge subset containing sarcastic expressions showed both models failing to interpret context-dependent humor, with ALBERT achieving merely 62% accuracy versus TinyBERT's 65%.

Furthermore,TinyBERT is not a model specifically designed for Chinese NLP tasks.Its performance will decline to some extent when dealing with Chinese sentiment analysis tasks,and neither of them has been trained on a large amount of data for sarcasm detection.With the limition of the size of the dataset , multiple trainings will lead to overfitting problems. Therefore,the results are only for verification that lightweight models can be used for Chinese sarcasm detection tasks, and that TinyBERT can be used for Chinese NLP tasks.

5.Future Outlook

Future work can be carried out in three directions: first, build a dynamic corpus of Chinese irony combined with emerging online terms, update and correct it in real time, and ensure the accuracy and efficiency of irony detection; second, train a heterogeneous model collaborative training framework that implements targeted functions, and make the large model function more precise. Although there may be overfitting problems, the targeted training model can modularize the model and reduce the pressure of training the model; third, explore the adaptive distillation strategy based on reinforcement learning, and select the appropriate model lightweight and distillation model construction according to requirements to achieve the optimal model compression under different data scales. This study provides a certain reference for the Chinese irony detection task of the lightweight BERT model.

References

[1]Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, Qun LiuTinyBERT: Distilling BERT for Natural Language Understanding, https://arxiv.org/abs/1909.10351

[2]Yangyang Li, Yuelin Li, Shihuai Zhang, Guangyuan Liu, Yanqiao Chen, Ronghua Shang, Licheng Jiao, An attention-based, context-aware multimodalfusionmethodforsarcasm detection using inter-modality inconsistency [3]dr. ir. F. Lelli , P.K. Medappa , Sarcasm Detection in Structured Text using DistilBERT: Evaluating the Impact of Text Normalization on Model Performance, https:/doi.org/10.1016/j.knosys.2024.111457

[4]ADITYA JOSHI,PUSHPAK BHATTACHARYYA,MARKJ. CARMAN,Automatic Sarcasm Detection: A Survey, ACMComputing Surveys, Vol. 50, No. 5, Article 73. Publication date: September 2017.

[5]WangqunChen,FuqiangLin,GuoweiLi,BoLiu,Asurveyofautomaticsarcasmdetection:Fundamentaltheories, formulation, datasets, detection methods, and opportunities, https://doi.org/10.1016/j.neucom.2024.127428

[6]YafengRen,ZilinWang,QiongPeng,DonghongJi,Aknowledge-augmentedneuralnetworkmodelforsarcasm detection,https://doi.org/10.1016/j.ipm.2023.103521

[7]Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut, ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, https://arxiv.org/abs/1909.11942

[8]Naman Bhargava, Mohammed I. Radaideh, O Hwang Kwon, Aditi Verma, Majdi I. Radaideh, On the Impact of Language Nuances on Sentiment Analysis with Large Language Models: Paraphrasing, Sarcasm, and Emojis, https://arxiv.org/abs/2504.05603

[9]Harleen Kaur Bagga, Jasmine Bernard, Sahil Shaheen, Sarthak Arora, Was that Sarcasm?: A Literature Survey on Sarcasm Detection, https://arxiv.org/abs/2412.00425