Dean&Francis

ISSN 2959-6130

Financial Sentiment Analysis with Large Language Models

Xinyu Cheng

Department of Mechanical Engineering and Material Science, Yale University, New Haven, US
*Corresponding author: xinyu.cheng@yale.edu

Abstract:

Financial sentiment analysis is vital for applications such as market prediction and risk management. While domain-specific models like (Financial Bidirectional Encoder Representations from Transformers) FinBERT are widely used, their limited scalability constrains performance across diverse financial texts. This paper investigates the effectiveness of large language models (LLMs) with parameter-efficient fine-tuning strategies. We fine-tune Llama-3.1-8B and Owen-3-8B using LoRA and QLoRA, and evaluate them on Financial PhraseBank and FiOA-SA datasets. Experiments show that LLMs consistently outperform FinBERT, achieving up to 88.9% accuracy on PhraseBank and 81.7% accuracy with 0.74 macro-Fl on FiQA-SA LoRA yields stronger performance, especially on minority classes, while QLoRA maintains comparable accuracy with significantly reduced memory cost. Moreover, Qwen-3 outperforms Llama-3.1 on noisy microblogs, benefiting from its Mixture-of-Experts (MoE) architecture, which enhances efficiency and diversity through conditional computation. These findings confirm that parameter-efficient fine-tuned LLMs provide both accuracy and efficiency, and represent strong alternatives to domain-specific models in financial sentiment analysis.

Keywords: Financial Sentiment Analysis; Large Language Models (LLMs); Parameter-Efficient Fine-Tuning.

1. Introduction

In financial markets, asset prices often reflect macroeconomic expectations, and investors quickly adjust their positions when new information emerges. Therefore, financial texts such as news reports, analyst reports, and company announcements contain important sentiment signals that affect market volatility. With the rapid expansion of such texts, manual interpretation has become difficult to meet demand, and automated sentiment analysis has emerged as the times require.

The core task of financial sentiment analysis is to classify text as positive, negative, or neutral. However, this process is limited by two aspects: first, there is a large amount of proprietary vocabulary and obscure expressions in the financial field, making it difficult for general models to accurately understand semantics; second, labeled data is scarce because

sentiment annotation requires professional financial knowledge. Traditional dictionary methods (such as the Loughran-McDonald dictionary) are interpretable, but difficult to capture contextual semantics; while deep learning models can model complex semantic relationships, they rely on large-scale labeled corpora [1].

The existing research has developed from the dictionary method to deep learning and the transformer model. FinBERT has achieved 84% accuracy and 0.83 F1 value by continuing to pre-train Bert on the financial corpus and fine-tuning it on the financial Phrasebank, which is more than 10% higher than the traditional method [1]. Zhang et al. further showed that small-scale Large Language Models (LLMs) (such as opt-1.3b and pythia-1.4b) can also surpass FinBERT after full parameter fine-tuning, but the calculation cost is high [2]. These studies not only highlight the potential of large models in financial text analysis but also reveal the necessity of efficient fine-tuning methods.

This paper studies the performance of Llama3.1 and Qwen3 in financial sentiment analysis based on LoRA and QLoRA, aiming to verify their feasibility and effectiveness under limited computing resources, and to explore whether large models can achieve excellent performance based on small-scale high-quality data sets in the absence of additional field pretreatment training [3, 4]. The experimental results show that neutral emotion is the most balanced among the models, especially in the recall rate and F1 value.

2. Methodology

In this section, we describe the methods designed to conduct the experiments. First, we list the foundation models that are used as a part of this project. Then, two new dataset collections are introduced, one with data based on documents and the second with instructions. We also give details of designing a data augmentation strategy for the instructions, as well as the description of the training process carried out to fine-tune the foundation models.

2.1 Foundation Models

2.1.1 FinBERT

FinBERT is a domain-specific adaptation of BERT for financial natural language processing tasks, particularly sentiment analysis, developed through a two-step process that first involved further pretreatment training BERT on a large corpus of financial texts, such as the Reuters TRC2 financial subset to capture domain-specific terminology, and then fine-tuning the model for sentiment classification using the Financial PhraseBank dataset [5]. By leveraging transfer learning in this way, FinBERT significantly outperformed traditional machine learning methods and earlier neural architectures, establishing itself as one of the most widely used baselines in financial sentiment analysis.

Despite its strong performance, FinBERT is limited by its reliance on the BERT-base architecture, which was trained on relatively smaller and older corpora compared to modern large language models. Moreover, its design was narrowly optimized for sentiment classification, making it less flexible for broader financial NLP tasks [6].

2.1.2 Llama 3.1

Large Language Model Meta AI (Llama) is a family of open-source foundation models proposed by Meta AI, designed to provide efficient and high-performing alternatives to proprietary LLMs [7]. The latest version, Llama 3, further improves model capacity by scaling up training data and parameters, incorporating high-quality multilingual corpora, and enhancing instruction-following capabilities. Compared to earlier open-source models such as OPT or Pythia, Llama achieves significantly stronger performance across a wide range of natural language understanding and generation tasks, even with relatively smaller parameter counts [8] [9].

The paper adopts Llama 3.1 as one of the backbone models for financial sentiment analysis. Its improved pretreatment training corpus and stronger alignment with instruction-following tasks make it highly suitable for low-resource domains such as finance, where labeled data is scarce. Furthermore, the availability of different parameter sizes provides flexibility for experimenting under varying computational constraints.

2.1.3 Qwen 3

Qwen is a family of open-source large language models released by Alibaba Cloud, designed with strong multilingual support and scalability for real-world applications [10]. The most recent version, Qwen 3, advances previous generations by significantly expanding its training corpus, improving cross-lingual reasoning, and enhancing its performance on instruction-following and dialogue-oriented tasks. Unlike earlier open-source models, Qwen emphasizes robustness across multiple languages, making it particularly suitable for tasks that require domain adaptation beyond English.

The research employs Qwen 3 as another backbone model for financial sentiment analysis. Its large-scale multilingual pretreatment training allows it to capture nuanced financial terminology, while its strong performance in zero-shot and few-shot settings helps mitigate the scarcity of annotated financial datasets. Together with Llama 3.1, Qwen 3 provides a complementary foundation for exploring the effectiveness of parameter-efficient fine-tuning methods in the financial domain.

ISSN 2959-6130

2.2 Parameter-Efficient Fine-tuning Methods

2.2.1 LoRA

Low-Rank Adaptation (LoRA) is one of the most widely used parameter-efficient fine-tuning (PEFT) methods for adapting large language models to downstream tasks [11]. Instead of updating all model parameters during fine-tuning, LoRA freezes the original pre-trained weights and introduces a set of low-rank trainable matrices into the Transformer architecture. Specifically, the adaptation is applied to weight matrices in attention layers, where the update matrix is factorized into two smaller matrices with a low-rank decomposition. This drastically reduces the number of trainable parameters while retaining the expressive power of the model.

With LoRA, we can significantly lower the memory footprint and computational cost, making fine-tuning feasible on limited hardware. Also, multiple LoRA modules can be trained for different tasks and easily combined without retraining the full model. Empirical results show that LoRA achieves performance close to, or even surpassing, full-parameter fine-tuning on various NLP tasks. Therefore, in this work, we apply LoRA to LLM for financial sentiment analysis, aiming to achieve strong performance under resource constraints while maintaining flexibility for future extensions.

2.2.2 QLoRA

Quantized Low-Rank Adaptation (QLoRA) is an extension of LoRA designed to further improve the efficiency of fine-tuning large language models [12]. QLoRA combines low-rank adaptation with quantization techniques, where the base model is quantized into 4-bit precision using double quantization and paged optimizers. This approach substantially reduces GPU memory consumption while maintaining model accuracy.

Similar to LoRA, QLoRA freezes the original pre-trained model weights and introduces trainable low-rank adapters. However, by storing the base model in 4-bit precision, QLoRA allows fine-tuning of very large models (e.g., 65B parameters) on a single GPU with limited memory. Experimental results show that QLoRA not only matches the performance of full-precision fine-tuning but also improves efficiency by enabling training under much lower resource requirements. In this work, the combination of quantization and low-rank adaptation makes QLoRA particularly suitable for financial NLP, where resource efficiency is critical and training often needs to be performed with constrained labeled data.

3. Experimental Setup

3.1 Model Evaluation

In the task of financial sentiment classification, we use four commonly used indicators: accuracy, precision, recall, and F1 score to comprehensively evaluate the performance of the model in terms of overall correctness, error types, and the balance of positive and negative categories, so as to ensure the robustness and fairness of the results [13]. In order to verify whether the general large-scale model can replace the special model in the financial field, this study selected two types of models for comparison in the experiment. On the one hand, we have fine-tuned LoRA and QLoRA for Llama-3.1-8b and Qwen-3-8b, which represent the cutting-edge level of current general-purpose LLM; On the other hand, we introduced FinBERT (base/large), a representative model in the financial field, and used full parameter fine-tuning as the baseline for comparison. This design can comprehensively examine the performance of the general model and domain-specific model in financial sentiment analysis, and compare the performance and efficiency differences between the efficient parameter tuning method and the traditional full parameter tuning method.

3.2 Datasets

This paper evaluated all models on two widely used financial sentiment corpora. These datasets differ in domain granularity (news vs. microblogs) and annotation style, which allows us to assess in-domain and cross-domain robustness.

3.2.1 Financial PhraseBank

The primary dataset used for financial sentiment analysis in this study is the Financial PhraseBank [14]. It contains 4,845 English sentences randomly selected from financial news articles in the LexisNexis database. Each sentence was annotated by 16 experts with backgrounds in finance and business, who were asked to assign a sentiment label (positive, negative, or neutral) based on how the information might affect the stock price of the company mentioned. In our experiments, we utilize the 50% agreement subset, which includes sentences where at least half of the annotators agreed on the sentiment label. Following this selection, the dataset is split into 70% training, 10% validation, and 20% test sets. This ensures a balanced evaluation protocol.

3.2.2 FiQA-SA

In addition to the Financial PhraseBank, we also employ the FiQA Sentiment Analysis (FiQA-SA) dataset, which was introduced as part of the FiQA 2018 Challenge on Financial Opinion Mining and Question Answering [15].

The dataset consists of 1,174 English financial news headlines and microblog posts (e.g., tweets), each annotated with a continuous sentiment score ranging from 1 (most negative) to +1 (most positive). Unlike Financial PhraseBank, which provides discrete sentiment classes, FiQA-SA enables evaluation in a regression setting, capturing more fine-grained sentiment variations.

3.3 Training Details

All experiments were conducted using PyTorch with the HuggingFace Transformers and PEFT libraries [16]. Model optimization was performed with the Adam W optimizer, using a fixed learning rate of 2×10-4. The training batch size was set to 8, and evaluation was carried out with a batch size of 16. To balance efficiency and GPU memory constraints, we applied gradient accumulation with an accumulation step of 2. A warm-up ratio of 0.1 was employed to stabilize optimization during the early training phase. Gradient check pointing was optionally enabled to further reduce memory usage during training. Each input sequence was truncated or padded to a maximum length of 256 tokens. Training was performed for 3 epochs, with early stopping based on the validation macro-Fl score to mitigate overfitting. Mixed precision was used to accelerate training and reduce memory consumption. For all models, a task-specific classification head was added on top of the final hidden representation to perform downstream classification.

For LoRA settings, the adapter rank was set to 16, with a scaling factor of 32 and a dropout rate of 0.05. For QLoRA, we employed NF4 quantization to further reduce memory usage. All experiments were trained on 4 NVIDIA RTX 3090 GPUs with 24GB memory each in parallel. To ensure reproducibility, each experiment was repeated with five different random seeds, and the reported results represent the average performance across runs.

3.4 Results

The results of FinBERT, Llama3.1, and Qwen 3 using two parameter adjustment methods on the classification tasks of Financial PhraseBank and FiQA data sets are shown in Tables 1 and 2.

Table 1. Performance comparison of models with different fine-tuning methods

Models	Fine-tuning Methods	Accuracy	precision	recall	F1-score
FinBERT	Full fine-tuning	0.8400	0.8100	0.8600	0.8300
Llama	LoRA	0.8887	0.8946	0.8690	0.8811
	QLoRA	0.8845	0.8861	0.8546	0.8686
Qwen	LoRA	0.8845	0.8963	0.8628	0.8783
	QLoRA	0.8845	0.8927	0.8568	0.8733

Table 2. Performance comparison of Llama 3.1, Qwen 3, and FinBERT under different fine-tuning methods

Models	Fine-tuning Methods	Accuracy	precision	recall	F1-score
FinBERT	Full fine-tuning	0.56	0.52	0.58	0.50
Llama 3.1	LoRA	0.7617	0.6196	0.6131	0.6160
	QLoRA	0.7872	0.7033	0.6536	0.6696
Qwen 3	LoRA	0.8170	0.7442	0.7368	0.7403
	QLoRA	0.7617	0.6667	0.6818	0.6737

The experimental results on the financial Phrasebank and FiQA-SA datasets show that the large language model is significantly better than the domain-specific FinBERT on the whole. In the Phrasebank dataset, the accuracy of Llama-3.1 and Qwen-3 reached about 88–89%, and the macro average F1 score was balanced, while the result of FinBERT was significantly lower. Further comparison shows that LoRA fine-tuning is slightly better than QLoRA in macro average F1, but the gap between them is no more than one percentage point, indicating that although quantization has a small loss in a few types of recalls, the overall performance is still stable and has higher computational efficiency. Category-level analysis shows that neutral emotions are the most easily identified, while negative emotions are weak due to uneven data distribution.

In the FiQA-SA dataset, the gap between the general large model and FinBERT is more prominent. The accuracy of FinBERT is only 56.2%, and the macro average F1 is 0.50, indicating that its generalization ability is limited when dealing with noise and unstructured text. In contrast, the accuracy of Llama-3.1 can be improved to 76–78%, while the LoRA version of Qwen-3 has an accuracy of 81.7%

ISSN 2959-6130

and a macro average F1 of 0.74, which is more than 20% higher than that of FinBERT. Although the performance of QLoRA is slightly lower, it is still significantly better than the baseline model, reflecting the application potential of efficient parameter tuning in resource-constrained scenarios.

To sum up, the experimental results fully illustrate the advantages of the general large-scale model in financial sentiment analysis. It can not only maintain high accuracy in structured financial news, but also show strong robustness in noisy social media texts. At the same time, LoRA and QLoRA have their own advantages in accuracy and efficiency, providing a variety of options for subsequent applications.

4. Discussion

The results of this study reveal the advantages and disadvantages of the parameter-efficient fine-tuning large language model in financial sentiment analysis. On the one hand, Llama-3.1 and Qwen-3 are significantly better than the domain-specific baseline model in terms of accuracy and robustness, which verifies the applicability of the general large model in specific domain tasks [17] [18]. On the other hand, there are differences in the performance of different emotion categories: neutral texts are easier to identify, while negative emotions are weaker due to uneven data distribution, which is consistent with the findings in existing studies [13] [19].

At the optimization level, LoRA is outstanding in maintaining fine-grained semantic features, while QLoRA significantly reduces resource consumption with similar accuracy, showing high application value. Future improvements can consider hybrid methods, such as combining LoRA with domain adaptive strategies, or introducing sparse activation structure to improve efficiency, which is consistent with the research view of Bai et al [10].

It should be pointed out that this study also has some limitations in the experimental design. First of all, the data set used is limited in scale, and it is difficult to fully cover the diversity of financial texts. Secondly, the experiment is only based on English data and does not involve multilingual financial sentiment analysis, which may affect the universality of the conclusion.

Future research can further improve the performance of the model by expanding the data scale, introducing a multilingual corpus, and exploring hybrid fine-tuning methods. For example, the research of Zhang et al. shows that the accuracy of the model can be significantly improved by introducing more user-generated financial texts [9]. Similarly, combining domain adaptive pretreatment training or retrieval to enhance the structure may become a potential direction to improve the robustness of the model to noisy financial texts.

5. Conclusion

This paper systematically evaluates the performance of LoRA and QLoRA on Llama-3.1 and Qwen-3. The experimental results show that the general large-scale model is better than FinBERT on the whole, with the highest accuracy of 81.7% and the macro average F1 of 0.74 on the FiQA-SA dataset. LoRA has a slight advantage in F1 score, while QLoRA significantly reduces the memory requirements while maintaining close accuracy, which is more suitable for resource-constrained scenarios. At the same time, Qwen-3 performs better than Llama-3.1 on noisy data, which reflects that the differences in the pretreatment training corpus and structure design of the model family will affect the cross-domain adaptability. In general, this study confirms the feasibility and advantages of the large-scale model with efficient parameter tuning in financial sentiment analysis, which can not only ensure accuracy and efficiency but also provide new enlightenment for the follow-up study of financial natural language processing.

References

[1] Araci D T. FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063, 2019.

[2] Inserte P, Rodriguez P, Nakhlé M, Qader R, Caillaut G, Liu J. Large language model adaptation for financial sentiment analysis. FinNLP-2, 2023.

[3] Meta AI. Llama 3.1: Advancing open foundation models. Technical Report, Meta AI, 2024.

[4] Alibaba Cloud. Qwen 3.0: Scaling multilingual open LMs. Technical Report, Alibaba Group, 2024.

[5] Devlin J, Chang M W, Lee K, Toutanova K. 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, 1: 4171–4186.

[6] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 2017: 5998–6008.

[7] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava A, Bhosale S, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[8] Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, Dewan C, et al. OPT: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068, 2022.

[9] Biderman S, Schoelkopf H, Anthony Q, Bradley H, Ohlson N, Black S. Pythia: A suite for analyzing large language models

across training and scaling. Proceedings of the 40th International Conference on Machine Learning (ICML 2023), PMLR 202, 2023: 4370–4385.

[10] Bai J, Dai Z, Dong L, Zhang W, Zhang X, Zhang S, Huang S, et al. Qwen technical report: Open foundation and chat models by Alibaba Cloud. arXiv preprint arXiv:2309.16609, 2023.

[11] Hu E J, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[12] Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient finetuning of quantized LLMs. arXiv preprint arXiv:2305.14314, 2023.

[13] Manning C D, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press, 2008.

[14] Malo P, Simha A, Korhonen P, Wallenius J, Takala P. Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the Association for Information Science and Technology, 2014, 65(4): 782–796.

[15] Maia M, Handschuh S, Freitas A, Davis B, McDermott R, Zarrouk M, Balahur A. WWW'18 open challenge: Financial opinion mining and question answering. Companion Proceedings of The Web Conference 2018 (WWW'18 Companion), 2018: 1941–1942.

[16] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Rush A M. Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020: 38–45.

[17] Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient finetuning of quantized LLMs. arXiv preprint arXiv:2305.14314, 2023.

[18] Hu E J, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[19] Kraus M, Feuerriegel S. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. Expert Systems with Applications, 2017, 118: 65–79.