eFinBERT: Efficient Financial Sentiment Classification

Aisha Hamad Hassan, Tushar Shinde
MIDAS (Multimedia Intelligence, Data Analysis and compreSsion) Lab
Indian Institute of Technology Madras, Zanzibar, Tanzania
shinde@iitmz.ac.in

Abstract

Transformer-based models, such as FinBERT, have achieved state-of-the-art performance in financial sentiment analysis, a critical task for understanding market trends and investor sentiment. However, their high computational and memory requirements present significant challenges for deployment in resource-constrained edge environments. In this work, we investigate post-training model compression techniques, specifically layer-wise fixed-bit quantization (ranging from 8-bit to 1-bit) and unstructured magnitude-based pruning, to reduce model size and inference latency while maintaining task performance. Using the Financial PhraseBank dataset, we perform a detailed layer sensitivity analysis to identify quantization bottlenecks and prune-tolerant layers. We introduce a sensitivity radar plot to visualize the impact of bit-width reduction on layer-wise accuracy, providing an interpretable framework for mixed-precision optimization. Furthermore, we demonstrate that selectively applying lower bit-widths to robust layers (e.g., Layers 5 and 7) enables targeted compression with minimal accuracy loss. Our results show that up to 90% parameter reduction is achievable with less than 2% absolute accuracy degradation compared to the full-precision model, underscoring the potential for efficient deployment of financial NLP models in low-power environments. This work provides a scalable and effective approach to optimizing transformer models for real-world applications in financial analysis and beyond.

1 Introduction and Related Work

Transformer-based models have revolutionized natural language processing (NLP) by effectively capturing long-range dependencies and complex semantic patterns. FinBERT [Araci, 2019], a BERT-based model fine-tuned specifically for financial sentiment analysis, has demonstrated impressive performance. However, its high computational complexity and memory requirements pose significant challenges for deployment in resource-constrained environments, hindering real-time financial applications.

To address these challenges, model compression techniques such as quantization and pruning have gained significant attention. Quantization reduces model size by lowering the precision of weights and activations, ranging from 32-bit to 8-bit or lower, thus accelerating inference and reducing memory footprint [Cheng et al., 2017, Wang et al., 2024, Xu and McAuley, 2023, Shinde, 2025]. Post-training quantization (PTQ) methods, such as GPTQ [Frantar et al., 2022] and ZeroQuant [Yao et al., 2022], offer efficient compression without the need for retraining the model. Layer-wise and mixed-precision quantization adapt bit-widths based on layer sensitivity, optimizing the trade-off between model accuracy and size [Liu et al., 2024, Ashkboos et al., 2024, Shinde, 2024].

Pruning, another key compression technique, removes redundant weights or model components such as neurons or attention heads [Gao et al., 2020, Chin et al., 2020]. Unstructured pruning typically uses magnitude-based thresholds, while structured methods consider global ranking and latency constraints [Singh et al., 2020]. Furthermore, pruning can improve model robustness and fairness,

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

which is particularly critical for handling the noisy and often ambiguous nature of financial data [Xu and Hu, 2022].

While model compression has made significant strides for general-purpose large language models (LLMs), domain-specific models like FinBERT remain underexplored Sharma et al. [2025]. Recent studies have highlighted the potential of adaptive compression techniques that incorporate layer sensitivity analysis to further optimize model efficiency and accuracy [Liu et al., 2024, Ashkboos et al., 2024].

In this work, we explore post-training quantization (ranging from 8-bit to 1-bit) and unstructured pruning on FinBERT using the Financial PhraseBank dataset. We introduce a sensitivity radar plot to visualize the impact of varying bit-widths on individual layers, providing a novel framework for designing mixed-precision strategies. Our approach reduces model size by up to 90% with less than 2% accuracy degradation, making it feasible to deploy FinBERT efficiently in low-resource financial applications.

2 Method

We apply two widely-used model compression techniques to FinBERT: post-training quantization and unstructured pruning. These methods are chosen for their effectiveness in reducing model complexity while preserving inference performance. Both techniques are critical for optimizing transformer-based models like FinBERT, which are computationally expensive, particularly in resource-constrained environments.

Quantization. Quantization compresses a neural network by mapping full-precision weights and activations to lower-precision representations, thus reducing memory usage and speeding up inference. Let $W \in \mathbb{R}^n$ denote a vector of floating-point weights. Uniform k-bit quantization maps W to a discrete set $\mathcal{Q} = \{q_1, \ldots, q_{2^k}\}$ using a scaling function:

$$\hat{w}_i = \text{round}\left(\frac{w_i - \alpha}{\Delta}\right)\Delta + \alpha, \quad \forall i \in \{1, \ldots, n\}, \tag{1}$$

where $\alpha = \min(W)$, $\beta = \max(W)$, and $\Delta = \frac{\beta - \alpha}{2^k - 1}$ is the quantization step size. The quantization process reduces the precision of model weights and activations, which effectively reduces the model size and computational cost during inference. In this study, we focus on post-training quantization (PTQ), where quantization is applied to a pretrained model without retraining. We evaluate both uniform and per-layer quantization, analyzing the sensitivity of different layers across the model to different bit-widths, allowing us to determine the optimal compression for each layer.

Pruning. Pruning reduces model size by eliminating weights or units that contribute minimally to the network's output. Formally, let W be the weight tensor of a layer. A pruning function $\mathcal{P}_\tau$ removes weights based on a saliency criterion $s(\cdot)$ and a threshold τ: $\mathcal{P}_\tau(W) = W \odot \mathbb{I}(s(W) > \tau)$, where $\odot$ denotes element-wise multiplication and $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if the condition is true and 0 otherwise. The saliency criterion $s(W)$ is typically based on the magnitude of weights, with smaller weights being considered less important and thus pruned. We adopt magnitude-based unstructured pruning, ensuring compatibility with transformer blocks while minimizing accuracy degradation. This method helps eliminate redundant parameters without compromising model performance, particularly in the case of large-scale transformer models like FinBERT.

These two techniques, are evaluated independently to quantify the trade-off between model compactness and predictive accuracy for domain-specific sentiment classification, which is crucial for deploying models like FinBERT in resource-constrained environments.

3 Experimental Setup

All experiments were conducted on the Kaggle platform using an NVIDIA Tesla P100 GPU, which provided sufficient computational resources for fine-tuning and evaluating large transformer models, such as FinBERT.

Dataset. We use the Financial PhraseBank dataset Malo et al. [2014] augmented with the FiQA (Financial Question Answering) dataset, totaling 5,842 financial sentences labeled into *Positive*, *Negative*, and *Neutral* categories. This diverse dataset was selected for its comprehensive coverage of financial language and real-world sentiment expressions. It is publicly available at `https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis`.

Implementation Details. The FinBERT model, a domain-specific BERT variant pre-trained on financial text, was used for this study. It consists of 110 million parameters, including 12 transformer layers with multi-head self-attention and feed-forward networks. The pretraining on financial text ensures the model's ability to capture domain-specific terminology and sentiment nuances.

Model Compression Configuration. We evaluate four compression techniques: post-training quantization, unstructured pruning, layer-wise quantization, and mixed-precision quantization.

Quantization: Applied post-training, with bit-widths from 16 to 1 for weights and activations. This range was chosen based on preliminary experiments that showed higher bit-widths did not improve performance significantly, while lower bit-widths provided substantial compression with minimal accuracy loss. *Pruning:* Unstructured pruning was applied by progressively removing 10% to 80% of weights, based on magnitude ranking. This technique reduces model size while retaining key information for financial sentiment classification. *Layer-wise Quantization:* Bit-widths were varied from 8 to 1 across transformer layers 0 to 11, allowing fine-grained compression based on layer sensitivity, enabling aggressive compression of less important layers without compromising performance. *Mixed-Precision Quantization:* Applied varying bit-widths (7 to 1) to layers 5 and 7, while keeping other layers (including embeddings and classifiers) at 8-bit precision. This strategy compresses less sensitive layers while preserving performance-critical layers at higher precision.

Training and Evaluation Protocol. The models were fine-tuned using the Adam optimizer with a learning rate of 2×10^{-5}, batch size of 16, and trained for 4 epochs. No retraining was performed post-quantization or pruning. All experiments were repeated with three different random seeds, and mean performance is reported.

Evaluation Metrics. Performance was evaluated using **Accuracy**, defined as the proportion of correctly classified instances. Additionally, we evaluate the **Compression Ratio (CR)**, which is defined as the ratio of the original model size to the compressed model size, and **Sparsity (%)**, which measures the proportion of weights pruned. These metrics provide insights into the trade-offs between model efficiency and performance.

4 Results and Discussions

Quantization Performance. As shown in Table 1, FinBERT maintains high accuracy (75.8%) with 8-bit and 16-bit quantization, demonstrating strong resilience to moderate precision reduction. Performance degradation begins below 8 bits, with accuracy dropping to 74.6% at 7-bit and 72.3% at 6-bit. However, aggressive quantization (5-bit and below) causes a significant performance drop, with accuracy falling to 14.7% at 2-bit and 1-bit. These results emphasize the trade-offs in quantization and highlight the limitations of extreme bit-width reductions for nuanced tasks like financial sentiment analysis.

Pruning Effects. Table 2 shows that unstructured pruning severely impacts model performance, even at low sparsity levels (10-20%). This sensitivity is attributed to the dense interdependencies between layers in transformer models, where removing weights disrupts the model's ability to encode key semantic relationships. Beyond 40% sparsity, accuracy drops drastically, reaching random-guess levels at 70-80%. These findings highlight the need for careful pruning strategies to prevent performance degradation.

Layer-wise Quantization Trends. Figure 1 illustrates that FinBERT's sensitivity to quantization varies across layers. Layers 10 and 11 remain robust even at extreme quantization (1-bit), while earlier layers like 5-7 experience significant accuracy loss. This suggests that lower-bit precision is more disruptive to early layers responsible for syntactic and semantic encoding. Mixed-precision strategies, which allocate higher precision to sensitive layers and lower precision to more resilient layers, can optimize performance while reducing model size.

Table 1: Accuracy after Quantization to Various Bit-Widths

Bit-Width	Acc. (%)
32-bit	75.8
16-bit	75.8
8-bit	75.8
7-bit	74.6
6-bit	72.3
5-bit	53.6
4-bit	49.8
3-bit	49.4
2-bit	14.7
1-bit	14.7

Table 2: Accuracy with Varying Sparsity Levels after Unstructured Pruning

Sparsity (%)	Acc. (%)
0	75.8
10	50.3
20	53.4
30	52.7
40	37.6
50	34.9
60	34.7
70	16.9
80	28.5

Table 3: Mixed-Precision Quantization of Layers 5 and 7 with Accuracy and CR

L5 & L7 Bit-width	Acc. (%)	Size (MB)	Comp. Ratio
32	**75.80**	417	1×
8	**75.80**	104.25	4.00×
7	75.61	102.72	4.07×
6	75.33	101.03	4.13×
5	72.80	99.34	4.20×
4	54.01	97.65	4.28×
3	38.77	95.96	4.35×
2	14.72	94.27	4.43×
1	38.86	92.58	4.51×

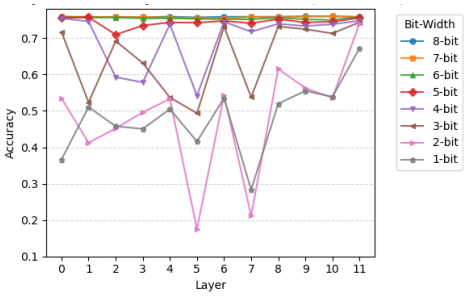

Figure 1: Accuracy of FinBERT under Layer-wise Quantization.

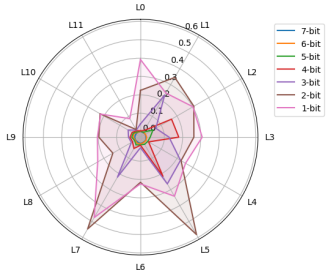

Figure 2: Layer-wise Bit-Width Sensitivity of FinBERT.

Mixed-Precision Quantization. We further explored mixed-precision quantization, applying lower bit-widths to sensitive layers (Layers 5 and 7) while keeping the rest of the model at 8-bit precision. Table 3 demonstrates that 7-bit and 6-bit quantization achieves nearly the same accuracy as the 8-bit baseline (75.61% and 75.33%) while reducing model size by over 4×. Even at 5-bit, the model maintains competitive accuracy (72.80%). However, below 5-bit, performance significantly drops. These results reinforce that mixed-precision quantization can provide substantial compression with minimal accuracy loss, making it a promising approach for resource-constrained environments. Our experiments reveal that FinBERT is robust to global quantization down to 6–7 bits, but extreme quantization (below 4 bits) degrades performance significantly. Selective layer-wise quantization, where sensitive layers retain higher precision, mitigates this degradation. Mixed-precision quantization further reduces model size while preserving accuracy. In contrast, unstructured pruning, especially at moderate to high sparsity, proved more disruptive than quantization. These findings emphasize the importance of layer-sensitive, adaptive compression strategies that balance model size and performance, offering a scalable solution for deploying transformer models in low-resource settings. In Appendix A, we present a detailed comparison of state-of-the-art models for financial sentiment analysis on the Financial PhraseBank dataset.

5 Conclusion and Future Work

We present an empirical study on post-training quantization and pruning applied to FinBERT for financial sentiment analysis. Our results demonstrate that FinBERT retains high accuracy even with 8-bit quantization, while layer-wise sensitivity shows significant variation across different layers. Leveraging this observation, we propose a mixed-precision strategy, assigning lower bit-widths to robust layers (e.g., Layers 5 and 7), achieving substantial model compression with minimal accuracy loss. In contrast, unstructured pruning results in significant performance degradation, even at moderate sparsity levels. This highlights the critical importance of model-aware compression strategies, where the compression method is tailored to the specific characteristics of each layer. Future work will explore structured pruning techniques, such as attention head and neuron pruning, along with quantization-aware training and adaptive bit-width allocation based on layer sensitivity. These methods aim to further optimize transformer models for deployment in resource-constrained environments, ensuring that they maintain high accuracy while minimizing model size and computational overhead.

A Existing SOTA Methods Comparison

In this section, we compare the performance of various state-of-the-art (SOTA) models on the Financial PhraseBank dataset, a widely-used benchmark for financial sentiment analysis. Transformer-based models, particularly FinBERT [Araci, 2019], have generally outperformed traditional machine learning techniques due to their ability to capture complex semantic patterns in financial text. FinBERT, a BERT-based model fine-tuned for financial sentiment analysis, achieves an accuracy of 75.8%, demonstrating strong performance on this dataset. However, despite its strong performance, its large model size (110M parameters) can limit its deployment in resource-constrained environments.

In contrast, statistical methods like LPS [Malo et al., 2014] and HSC [Krishnamoorthy, 2018] offer a more lightweight alternative, achieving accuracies of 71.0% with parameter sizes of just 1.5M and 1.8M, respectively. While these models are significantly smaller, they fall short of the accuracy achieved by transformer-based approaches. BloombergGPT [Wu et al., 2023], a large language model designed specifically for finance, achieves an accuracy of 51.0% on the Financial PhraseBank dataset. Despite its large model size (1000M parameters), it does not perform as well as smaller, domain-specific models, underscoring the importance of tailoring models for specific tasks rather than relying on general-purpose architectures. Additionally, LSTM models combined with pretrained word embeddings (ELMo) [Araci, 2019] demonstrate a competitive performance of 75.0% accuracy, with a smaller model size of 45M parameters. Although LSTM-based models offer good performance, they are still outperformed by transformer models such as FinBERT.

This comparison highlights the dominance of transformer-based models like FinBERT, which provide superior accuracy at the cost of larger model sizes. However, for applications requiring resource-efficient models, statistical methods like LPS and HSC offer a viable alternative, albeit with slightly lower accuracy.

Table 4: SOTA Comparison for Financial Sentiment Analysis on Financial PhraseBank Dataset

Model	Approach	Accuracy (%)	# Parameters (in Million)
FinBERT [Araci, 2019]	Transformer-based (BERT variant)	75.8	110
LPS [Malo et al., 2014]	Statistical-based Method	71.0	1.5
HSC [Krishnamoorthy, 2018]	Statistical-based Method	71.0	1.8
BloombergGPT [Wu et al., 2023]	Large Language Model for Finance	51.0	1000
LSTM with ELMo [Araci, 2019]	LSTM + Pretrained Word Embeddings (ELMo)	75.0	45

References

Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

Saleh Ashkboos, Maximilian L Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. Slicegpt: Compress large language models by deleting rows and columns. *arXiv preprint arXiv:2401.15024*, 2024.

Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017.

Ting-Wu Chin, Ruizhou Ding, Cha Zhang, and Diana Marculescu. Towards efficient model compression via learned global ranking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1518–1528, 2020.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Shangqian Gao, Feihu Huang, Jian Pei, and Heng Huang. Discrete model compression with resource constraint for deep neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1899–1908, 2020.

Srikumar Krishnamoorthy. Sentiment analysis of financial news articles using performance indicators. *Knowledge and Information Systems*, 56(2):373–394, 2018.

Songwei Liu, Chao Zeng, Lianqiang Li, Chenqian Yan, Lean Fu, Xing Mei, and Fangmin Chen. Foldgpt: Simple and effective large language model compression scheme. *arXiv preprint arXiv:2407.00928*, 2024.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

Avinash Kumar Sharma, Aisha Hamad Hassan, and Tushar Shinde. Towards efficient finbert via quantization and coreset for financial sentiment analysis. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 70–74, 2025.

Tushar Shinde. Adaptive quantization and pruning of deep neural networks via layer importance estimation. In *Workshop on Machine Learning and Compression, NeurIPS 2024*, 2024.

Tushar Shinde. Towards optimal layer ordering for efficient model compression via pruning and quantization. In *2025 25th International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2025.

Pravendra Singh, Vinay Kumar Verma, Piyush Rai, and Vinay Namboodiri. Leveraging filter correlations for deep model compression. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, pages 835–844, 2020.

Wenxiao Wang, Wei Chen, Yicong Luo, Yongliu Long, Zhengkai Lin, Liye Zhang, Binbin Lin, Deng Cai, and Xiaofei He. Model compression and efficient inference for large language models: A survey. *arXiv preprint arXiv:2402.09748*, 2024.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Canwen Xu and Julian McAuley. A survey on model compression and acceleration for pretrained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10566–10575, 2023.

Guangxuan Xu and Qingyuan Hu. Can model compression improve nlp fairness. *arXiv preprint arXiv:2201.08542*, 2022.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. *Advances in neural information processing systems*, 35:27168–27183, 2022.