Towards Efficient FinBERT via Quantization and Coreset for Financial Sentiment Analysis

Avinash Kumar Sharma, Aisha Hamad Hassan, Tushar Shinde
Indian Institute of Technology Madras, Zanzibar

Abstract

Real-time financial sentiment classification from social media is critical for applications in algorithmic trading, risk assessment, and market surveillance. However, deploying large-scale models like FinBERT on edge devices remains impractical due to their high memory and compute demands. Meanwhile, financial text poses unique challenges such as class imbalance, noisy syntax, and temporal drift. We propose a unified framework that jointly applies coreset selection and post-training quantization to achieve scalable and efficient financial NLP. Our method reduces training data by up to 90% through coreset selection and compresses model size by up to 4× via 8-bit quantization, while preserving over 90% of the original classification accuracy on benchmark financial sentiment datasets. This demonstrates the viability of deploying domain-specific NLP models in constrained environments, offering a principled solution for low-latency, resource-efficient financial text processing.

1 Introduction and Related Work

Financial sentiment analysis is a cornerstone of modern quantitative finance, enabling predictive insights for algorithmic trading, risk modeling, and market surveillance (Smailović et al., 2014; Cortis et al., 2017; Du et al., 2024). Unlike general sentiment tasks, financial sentiment exhibits domain-specific characteristics: formalized language, market-sensitive expressions, and nonlinear or latent connections to asset prices. As such, this task poses unique linguistic and deployment challenges.

We highlight three core bottlenecks that hinder scalable financial sentiment classification. First, the task is challenged by domain specificity and the need for interpretability. Financial sentiment is often subtle, expressed through technical vocabulary and domain-specific idioms. As a result, model predictions must not only be accurate but also interpretable to satisfy both regulatory standards and institutional trust requirements (Wang et al., 2025). Second, deployment environments often impose stringent resource constraints. Financial sentiment models are expected to operate in latency-sensitive and memory-limited contexts such as mobile trading applications or edge-based market monitoring systems, where large-scale transformer models become impractical (Shinde et al., 2025). Third, the availability of high-quality labeled data is limited. Financial text corpora are typically sparse, exhibit temporal non-stationarity, and are expensive to annotate, which restricts the scalability of supervised learning approaches (Wang et al., 2025).

Domain Adaptation in Financial NLP. To address linguistic specificity, domain-adapted language models such as FinBERT (Araci, 2019) and BloombergGPT (Wu et al., 2024) have emerged. These models are pre-trained on financial corpora, achieving superior performance on downstream tasks including entity recognition, sentiment tagging, and financial QA (Yang et al., 2020). However, their training and fine-tuning require substantial computational resources and data access, making them inaccessible to many institutions.

Model Compression Techniques. Model compression offers a viable route to scalable deployment. Quantization reduces memory and compute requirements by constraining weights and activations to lower-bit representations (Jacob et al., 2018; Shinde, a; Shinde and Tukaram Naik, 2024), while pruning removes redundant parameters with minimal accuracy loss (Han et al., 2015). Let $W \in R^{d\times h}$ be the weight matrix of a transformer layer; quantization maps W to a lower precision space $\hat{W} = Q(W)$ such that:

$$\hat{W}_{ij} = \text{round}\left(\frac{W_{ij} - \mu}{\Delta}\right)\Delta + \mu, \qquad (1)$$

where Δ is the quantization step. Recent work incorporates mixed-precision and layer-wise adaptive schemes for optimal compression without quality degradation (Sun et al., 2022; Kuzmin et al., 2023).

Coreset Selection for Data Efficiency. To mitigate data scarcity, coreset selection identifies informative training subsets that retain performance while reducing training cost. Margin-based (Sener and Savarese, 2017), information-theoretic (Bachem et al., 2017), and forgetting-based strategies (Toneva et al., 2018; Shinde, b) prioritize samples that influence decision boundaries. Given training data $\mathcal{D}$, the coreset $\mathcal{C} \subset \mathcal{D}$ is chosen such that:

$$E_{(x,y)\in\mathcal{D}}\left[\ell(f_\theta(x), y)\right] \approx E_{(x,y)\in\mathcal{C}}\left[\ell(f_\theta(x), y)\right] \tag{2}$$

where ℓ is the task loss and f_θ is the model. Recent advances in zero-shot or training-free adaptation offer alternatives to full fine-tuning. Proxy tuning (Liu et al., 2024) enables logit-space adaptation by aligning decision boundaries between expert and base models. This is especially useful in financial domains, where annotation is costly and model update cycles must be rapid. Such methods enable lightweight personalization without retraining. While model compression and data-efficient learning have each advanced separately, their synergy in financial NLP remains underexplored. This work introduces a unified framework that combines coreset selection with quantization for real-time, resource-aware financial sentiment classification.

2 Methodology

This section presents our integrated framework for efficient financial sentiment classification using the Twitter Financial News Sentiment dataset. The framework combines coreset selection and quantization-based model compression to address both computational cost and memory footprint in resource-constrained scenarios, such as mobile and edge deployment environments.

2.1 Framework Overview

Our framework comprises three sequential components: class-balancing preprocessing, coreset-based sample selection, and adaptive quantization-aware model compression. Initially, we mitigate class imbalance inherent in financial sentiment data by preserving all samples from minority classes: Bearish (1,442) and Bullish (1,923), and reducing the majority Neutral class to match the size of the largest minority class, resulting in a balanced dataset with 5,288 training samples distributed equally across the three sentiment classes. This ensures unbiased model training while retaining representative information from all categories.

Subsequently, we employ coreset selection to extract informative subsets from the balanced dataset, evaluating coreset fractions in the set $\{1.0, 0.5, 0.25, 0.1, 0.05\}$. These fractions are used to create reduced yet representative training sets, thereby enabling systematic analysis of data efficiency and computational cost reduction.

2.2 Fine-tuning Procedure

We fine-tune the pretrained FinBERT model on the selected coreset using hyperparameters optimized for the financial social media domain. Validation is conducted at the end of each epoch to monitor overfitting and performance stability.

2.3 Quantization-Aware Model Compression

To facilitate efficient inference, we compress the fine-tuned FinBERT model through post-training quantization. We evaluate uniform symmetric quantization across bit-widths $b \in \{8, 7, 6, 5, 4, 3, 2, 1\}$ to study the trade-off between model size, accuracy, and compute efficiency. For a given weight tensor w_i in layer i, we compute the quantization scale as:

$$\text{scale}_i = \frac{\max(w_i) - \min(w_i)}{2^b - 1} \tag{3}$$

The weights are quantized using the following transformation:

$$w_i^{(q)} = \text{round}\left(\frac{w_i - \min(w_i)}{\text{scale}_i}\right) \cdot \text{scale}_i + \min(w_i) \tag{4}$$

This linear mapping ensures that weights are projected into a discrete set of 2^b values, reducing memory requirements and enabling faster inference on low-power devices. We evaluate performance degradation due to quantization at each bit level, measuring accuracy, precision, recall, and F1-score on the test set.

2.4 Joint Evaluation with Coreset Selection

We perform a comprehensive evaluation of the combined effects of coreset size and quantization bit-width. For each coreset fraction, we apply quantization at all bit-widths from 8 to 1, creating a grid of models. Each model is evaluated for classification performance and compression efficiency. This systematic design allows us to jointly analyze

Table 1: Distribution of sentiment classes across train, validation, and test sets.

Class	Train	Validation	Test
Bullish	1,670	358	358
Bearish	1,253	269	269
Neutral	5,118	1,097	1,097

the impact of training data reduction and quantization granularity, providing insights into the optimal trade-off between accuracy and computational efficiency for real-world financial NLP deployment.

3 Experimental Setup

Dataset Description. We conduct our experiments using the publicly available Twitter Financial News Sentiment dataset (Zeroshot, 2022), which consists of 11,932 annotated tweets related to financial news and market discourse. Each tweet is categorized into one of three sentiment classes: Bullish (positive market outlook), Bearish (negative market outlook), and Neutral (no clear directional sentiment). The dataset exhibits a pronounced class imbalance, with approximately 65% of the samples labeled as Neutral, 20% as Bullish, and 15% as Bearish.

To ensure robust model learning and unbiased evaluation, we adopt a stratified data split strategy, reserving 70% of the dataset for training, 15% for validation, and 15% for testing. The class distributions are preserved across the splits. Table 1 summarizes the class distribution in each subset.

Experimental Configuration. All models were implemented using the PyTorch framework, and experiments were conducted on the Kaggle cloud platform equipped with an NVIDIA Tesla P100 GPU (16GB). To ensure reproducibility, we fixed the random seed across runs and used deterministic training settings wherever supported.

We fine-tuned the pretrained `FinBERT` model using the AdamW optimizer with a learning rate of 2×10^{-5}, batch size of 16, and linear learning rate warmup. Each model was trained for 3 epochs. Early stopping based on validation F1-score was used to prevent overfitting. Quantization and compression experiments were performed post-training.

Evaluation Protocol. We evaluate the models using standard classification metrics: Accuracy, Precision, Recall, and F1-score. In addition to classification performance, we measure the model's memory efficiency using the Compression Ratio (CR):

Table 2: Fine-tuning progression on balanced validation set.

Stage	Accuracy	F1 (Macro)	F1 (Weighted)
Pre-trained Baseline	0.326	0.284	0.309
Epoch 1	0.594	0.549	0.577
Epoch 2	0.773	0.765	0.772
Epoch 3 (Final)	0.834	0.833	0.835

$$\text{CR} = \frac{\text{Original Model Size (in MB)}}{\text{Compressed Model Size (in MB)}} \quad (5)$$

This metric quantifies the storage reduction achieved by quantization-based model compression. We also report inference time per batch to assess real-time deployment feasibility.

4 Results and Analysis

4.1 Fine-tuning Performance Analysis

We first evaluate the impact of domain-specific fine-tuning on the pre-trained `FinBERT` model using the balanced validation set. As shown in Table 2, accuracy increases from 32.6% (pre-trained) to 83.4% after three epochs of fine-tuning. This represents an absolute improvement of +50.8 percentage points, accompanied by similar gains in both macro and weighted F1-scores. These results confirm the substantial benefit of adapting language models to domain-specific financial discourse, particularly in the context of noisy, sentiment-rich social media text.

4.2 Coreset Selection Efficiency

We next assess the effect of coreset selection by training on reduced fractions $\{1.0, 0.5, 0.25, 0.1, 0.05\}$ of the full balanced training dataset. Table 3 reports performance metrics and training speedups. Remarkably, using only 10% of training data retains 90% of the full-model accuracy (75.1% vs. 83.4%) while yielding an 8.3× reduction in training time. This highlights the potential of representative subset selection in high-dimensional, redundant financial language datasets.

4.3 Quantization Performance

We evaluate post-training quantization of the fine-tuned model across bit-widths $b \in \{8, 7, ..., 1\}$. Table 4 summarizes the accuracy, macro F1-score, and corresponding compression ratios. Notably, 6-bit quantization maintains 97.4% of the full-precision accuracy (81.2% vs. 83.4%) with a 1.3×

Table 3: Coreset selection performance across data fractions.

Fraction	Samples	Accuracy	F1 (Macro)	Speedup
100%	1,297	0.834	0.833	1.0×
50%	648	0.825	0.822	1.9×
25%	324	0.798	0.795	3.6×
10%	129	0.751	0.748	8.3×
5%	64	0.687	0.684	14.3×

Table 4: Quantization impact on model accuracy and compression. CR - Compression Ratio

Bit-Width	Accuracy	F1 (Macro)	CR
8-bit	0.834	0.833	1.0×
7-bit	0.829	0.826	1.1×
6-bit	0.812	0.809	1.3×
5-bit	0.785	0.782	1.6×
4-bit	0.743	0.740	2.0×
3-bit	0.687	0.684	2.7×
2-bit	0.542	0.539	4.0×
1-bit	0.334	0.331	8.0×

model size reduction, representing an effective trade-off between model compactness and predictive performance.

4.4 Integrated Efficiency Trade-offs

We investigate the combined effect of coreset selection and quantization to identify optimal operating points for deployment. Table 5 presents the resulting accuracy and efficiency gain for selected configurations. The results suggest a Pareto frontier: configurations offering strong accuracy-efficiency trade-offs for specific deployment scenarios such as mobile inference or low-latency market monitoring.

5 Discussion

Data Efficiency in Financial NLP. Our findings indicate that financial sentiment classification benefits significantly from intelligent data reduction. As little as 10% of the original training data achieves over 75% accuracy, supporting the hypothesis that social-financial discourse contains high levels of redundancy. This has practical implications for reducing annotation costs, accelerating model development cycles, and supporting rapid model deployment in emerging financial events.

Quantization for Deployment-Grade Models. Among the evaluated bit-widths, 6-bit quantization offers a desirable trade-off, preserving over 97% of full-precision model performance. This balance is crucial in financial contexts, where inference latency and memory constraints are critical, yet even minor accuracy degradation may result in measurable trading loss or poor risk signal estimation.

Table 5: Combined coreset selection and quantization results.

Configuration	Fraction	Bit-Width	Accuracy	Efficiency Gain
Baseline	100%	8	0.834	1.0×
High Efficiency	25%	6	0.776	3.1×
Balanced	50%	6	0.809	2.6×
Quality Focused	100%	6	0.812	1.3×
Maximum Compression	10%	4	0.658	5.0×

Real-World Deployment Implications. The proposed integrated framework offers significant advantages for a range of financial deployment scenarios. In mobile trading applications, low-latency sentiment inference is critical for responsive user experience, while in edge-based market monitoring systems, the reduced model size alleviates bandwidth and storage constraints. High-frequency trading (HFT) systems can benefit from real-time sentiment feeds with minimized inference delay, ensuring rapid decision-making under stringent timing requirements. Additionally, cloud-based financial analytics platforms can leverage the compressed models to lower infrastructure costs while maintaining robust sentiment classification capabilities. These deployment contexts all benefit from the dual advantage of reduced memory footprint and faster inference, without a substantial loss in accuracy.

Limitations. This study is focused on Twitter-based financial sentiment. Extension to other modalities, such as earnings call transcripts, SEC filings, or institutional reports, remains to be explored. Furthermore, quantization is simulated in software; actual deployment on hardware (e.g., INT4 inference on FPGAs or mobile NPUs) may exhibit different characteristics. Finally, we do not currently evaluate fairness or robustness under domain shift, which are important concerns in financial NLP.

6 Conclusion and Future Work

We propose an efficient framework for financial sentiment analysis that combines coreset selection with systematic model quantization. Our experiments show that training on only 10% of the data selected via coreset methods preserves approximately 90% of the original model's accuracy. Coupled with 6-bit quantization, this yields a model that is significantly smaller and faster, yet remains competitive in classification performance. The approach reduces training data by up to 90% and achieves about a 4× compression factor in model size, while maintaining accuracy within 10% of the full-data, full-precision FinBERT baseline on the Twitter Financial News Sentiment dataset. This demonstrates the practical applicability of transformer models in latency-critical financial settings such as real-time trading and mobile applications.

Future work will explore extending this framework to other financial text sources, and integrating additional compression techniques like structured pruning, low-rank factorization, and knowledge distillation.

References

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

Olivier Bachem, Mario Lucic, and Andreas Krause. 2017. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 519–535.

Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56(9):1–42.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713.

Andrey Kuzmin, Markus Nagel, Mart Van Baalen, Arash Behboodi, and Tijmen Blankevoort. 2023. Pruning vs quantization: Which is better? *Advances in neural information processing systems*, 36:62414–62427.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A Smith. 2024. Tuning language models by proxy. *arXiv preprint arXiv:2401.08565*.

Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

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

Tushar Shinde. b. High-performance lightweight vision models for land cover classification with coresets and compression. In *TerraBytes-ICML 2025 workshop*.

Tushar Shinde, Avinash Kumar Sharma, Shivam Bhardwaj, and Ahmed Silima Vuai. 2025. Navigating coreset selection and model compression for efficient maritime image classification. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1608–1616.

Tushar Shinde and Sukanya Tukaram Naik. 2024. Adaptive quantization of deep neural networks via layer importance estimation. In *International Conference on Computer Vision and Image Processing*, pages 220–233. Springer.

Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. 2014. Stream-based active learning for sentiment analysis in the financial domain. *Information sciences*, 285:181–203.

Mengshu Sun, Zhengang Li, Alec Lu, Yanyu Li, Sung-En Chang, Xiaolong Ma, Xue Lin, and Zhenman Fang. 2022. Film-qnn: Efficient fpga acceleration of deep neural networks with intra-layer, mixed-precision quantization. In *Proceedings of the 2022 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pages 134–145.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2018. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*.

Yuxiang Wang, Yuchi Wang, Yi Liu, Ruihan Bao, Keiko Harimoto, and Xu Sun. 2025. Proxy tuning for financial sentiment analysis: Overcoming data scarcity and computational barriers. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 169–174.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2024. Bloomberggpt: A large language model for finance, 2023. *URL https://arxiv. org/abs/2303.17564*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Zeroshot. 2022. Twitter financial news sentiment. `https://huggingface.co/datasets/zeroshot/twitter-financial-news-sentiment`. Dataset available on Hugging Face.