Assignment 3 Report: Fine-tuning BERT & FinBERT for Financial Sentiment Analysis

A0253408H Daniel Christopher Chan

Dataset and Motivation

The chosen dataset was the Financial PhraseBank dataset (Takala, 2022), a widely used benchmark corpus containing English sentences extracted from financial news articles regarding listed companies in OMX Helsinki. Each sentence is annotated by researchers and master's students at Aalto University School of Business, with positive, negative, or neutral labels. Specifically, the Sentences_75Agree.txt subset, consisting of 3,453 sentences where at least 75% of annotators agreed on the label, was chosen for this assignment. Thus, it serves as an expertly collated dataset ideal for the purposes of evaluating financial sentiment classification. Its moderate size and domain-specific nature also make it suitable for the fine-tuning experiments in this assignment.

The dataset exhibits a notable class imbalance, with approximately 62% neutral, 26% positive, and 12% negative sentences. This distribution reflects the realistic prevalence of factual reporting in financial news, which would be largely neutral in sentiment. The data is split into 80% training (2,762 sentences), 10% validation (345 sentences), and 10% test (346 sentences) to enable proper model selection and unbiased evaluation. Train, validation and test splits followed a largely similar distribution of neutral, positive and negative sentences to the overall dataset.

This choice of Financial PhraseBank directly aligns with the broader goals of the group project proposal submitted by my team for this course, where we aimed to explore financial news sentiment analysis by evaluating the performance of FinBERT against prompting large general-purpose language models in financial sentiment classification. Additionally, we plan to explore their applicability in an Asian financial news context, along with their practical applicability in providing meaningful signals of market movements.

Models and Fine-Tuning Strategies

Two pretrained models were explored: BERT-base-uncased, a standard version of the general-purpose BERT (Bidirectional Encoder Representations from Transformers) model, and FinBERT, a financial-domain variant further pretrained on large-scale financial text such as corporate filings and news articles (Huang et al., 2023). BERT is a Transformer-based language model that learns bidirectional contextual representations during pretraining, enabling it to capture nuanced semantic dependencies in text (Devlin et al., 2019). For this

assignment, BERT-base serves as a baseline, while FinBERT represents a domain-adapted model expected to perform better on financial sentiment classification. Baseline evaluations were conducted to assess the zero-shot performance of both models on the held-out Financial PhraseBank test set prior to any fine-tuning.

For fine-tuning, two strategies were explored. First, full fine-tuning was applied to both BERT and FinBERT, which updates all model parameters on the Financial PhraseBank training split, representing the standard but computationally expensive approach. Second, Low-Rank Adaptation (LoRA) was applied to FinBERT as a parameter-efficient fine-tuning (PEFT) method. LoRA injects small trainable matrices into each Transformer layer while keeping the original weights frozen, greatly reducing the number of trainable parameters (often by over 10,000×) with minimal performance loss (Hu et al., 2022). LoRA represents the weight updates as a low-rank decomposition $\Delta W = BA$, where $A \in \mathbb{R}^{r \times k}$and $B \in \mathbb{R}^{d \times r}$are the trainable matrices and r is the rank of the decomposition. A rank-ablation study (r = 4, 8, 16) was conducted to examine trade-offs between update capacity and efficiency.

While Huang et al. (2023) reported FinBERT's performance fully fine-tuned on the Financial PhraseBank dataset, they did not explore PEFT methods such as LoRA, which were introduced only more recently in the literature. It is important to note that although some Financial PhraseBank sentences were included in the pretraining text of FinBERT, Huang et al. (2023) conducted robustness tests confirming that this does not "introduce bias that favors FinBERT in sentiment classification because the pretraining sample does not include sentiment labels" (p. 814).

In this study, LoRA was applied only to FinBERT, as it is already a domain-adapted derivative of BERT. Applying PEFT to FinBERT thus provides a more meaningful evaluation of efficient adaptation for a specialized model, while BERT and its fully fine-tuned variant serve as general-purpose baselines to contextualize FinBERT's gains.

Experimental Setup

All experiments were conducted on Google Colab utilising its T4 GPU. Models were implemented in PyTorch using the Hugging Face Transformers and PEFT libraries. Both BERT-base-uncased and FinBERT used their respective pretrained tokenizers to ensure vocabulary consistency with pretraining. Each sentence was tokenized with max_length = 128, padding = "max_length", and truncation = True to standardize input sequences while preserving sufficient contextual coverage. All experiments used identical hyperparameters to

enable fair comparison across models: learning rate = 2×10^{-5}, batch size = 16, epochs = 4, weight decay = 0.01, and early stopping (patience = 2) to prevent overfitting. These settings follow standard practice for BERT-based sequence classification and balanced stability with convergence speed.

For FinBERT, label indices needed to be remapped to match its original configuration (0 = positive, 1 = negative, 2 = neutral), since the Financial PhraseBank dataset encodes them as (0 = negative, 1 = neutral, 2 = positive).

Fine-tuning and evaluation were performed using the Trainer API with identical training/validation splits. Each model optimized cross-entropy loss for three-class sentiment classification (positive, neutral, negative) through a linear classification head added to the final hidden state. Model efficiency was assessed by recording training time, GPU memory usage, and the proportion of trainable parameters. Performance evaluation used identical metrics across all models to ensure fair comparison on the held-out test dataset. Accuracy, macro- and weighted-F1, precision, and recall were computed using Hugging Face's evaluate library.

During training, the macro-F1 score on the validation set was used as the metric for model selection, as it provides a balanced measure of performance across all classes despite the dataset's class imbalance. This ensures that the model's ability to correctly classify minority sentiments (e.g., negative sentences) is not overshadowed by the dominant neutral class. The best-performing checkpoint was automatically loaded at the end of training.

Results

Overall Results. *Table 1* summarizes the performance and computational efficiency of all evaluated models.

Zero-shot Evaluation (BERT and FinBERT). As expected, FinBERT substantially outperformed the general-purpose BERT-base model in zero-shot evaluation (macro-F1 = 0.9401 vs 0.1255), demonstrating the strong benefit of domain-adaptive pretraining on financial text. This highlights that even without fine-tuning, FinBERT is already well-aligned with financial sentiment classification, while generic BERT struggles to generalize to this domain.

Full fine-tuning (BERT and FinBERT). After full fine-tuning (full FT), BERT-base (full FT) attained a much-improved test macro-F1 of 0.9002 and accuracy of 0.9191. However, FinBERT (full FT) reported scores of 0.9400 and 0.9509 respectively, nearly identical to its results prior to fine-tuning. This suggests that FinBERT's domain-adaptive

pretraining already captured much of the sentiment structure present in the Financial PhraseBank dataset, leaving limited room for improvement through task-specific tuning. In contrast, the general-purpose BERT model required substantial parameter updates to adapt to financial language, reflecting its weaker initial alignment with the target domain.

LoRA (FinBERT). Interestingly, for FinBERT, the best configuration of parameter-efficient fine-tuning (PEFT) using LoRA (r = 8) achieved slightly higher performance (macro-F1 = 0.9474, accuracy = 0.9595) than full fine-tuning while updating less than 1 % of parameters and cutting training time from 465.7117 s to 179.4156 s (≈ 61 % faster). These underscore the practicality of LoRA for efficient adaptation of large Transformer models to domain-specific sentiment tasks, achieving a strong balance between accuracy and efficiency that is especially valuable in resource-constrained environments.

LoRA Rank Ablations (FinBERT). Increasing the LoRA rank from r = 4 to r = 8 yielded modest gains in performance (macro-F1 = 0.9445 → 0.9474) at a moderate increase in training time (129.9347 s → 179.4156 s) and memory usage (5023.0487 MB → 5466.5083 MB). However, further increasing the rank to r = 16 did not improve results (macro-F1 = 0.9445), and incurred higher memory costs (peak = 5914.8175 MB) despite similar training time (179.9530 s). This suggests that beyond a certain point, increasing rank yields diminishing or unstable returns, indicating that LoRA performance does not scale linearly with rank. In this task, r = 8 provided the best balance between accuracy and efficiency, achieving near-optimal performance while maintaining low computational cost.

Figure 1 visualizes these trends, showing how $r = 8$ achieves the highest F1 score relative to both trainable parameters and training time. The left plot illustrates that while higher ranks increase parameter count, performance plateaus beyond a moderate rank. The right plot similarly shows that gains in F1 saturate despite longer training durations, reinforcing that $r = 8$ is the optimal configuration for this task.

model	trainable_ params	trainable_ ratio	train_ time_s	best_ val_f1	test_ accuracy	test_f1_ macro	peak_ memory_mb
BERT-base (zero-shot)	109,484,547	1.0000	0.0000	0.0000	0.2312	0.1255	0.0000
BERT-base (full FT)	109,484,547	1.0000	370.4627	0.9237	0.9191	0.9002	3093.2828
FinBERT (baseline)	109,484,547	1.0000	0.0000	0.0000	0.9509	0.9401	0.0000
FinBERT (full FT)	109,484,547	1.0000	465.7117	0.9613	0.9509	0.9400	4846.1481
FinBERT LoRA r=4	149763	0.0014	129.9347	0.9597	0.9566	0.9445	5023.0487
FinBERT LoRA r=8	297219	0.0027	179.4156	0.9613	0.9595	0.9474	5466.5083
FinBERT LoRA r=16	592131	0.0054	179.9530	0.9613	0.9566	0.9445	5914.8175

Table 1. Summary of model performance and efficiency across all experiments *(Rounded to 4 dp where appropriate)*

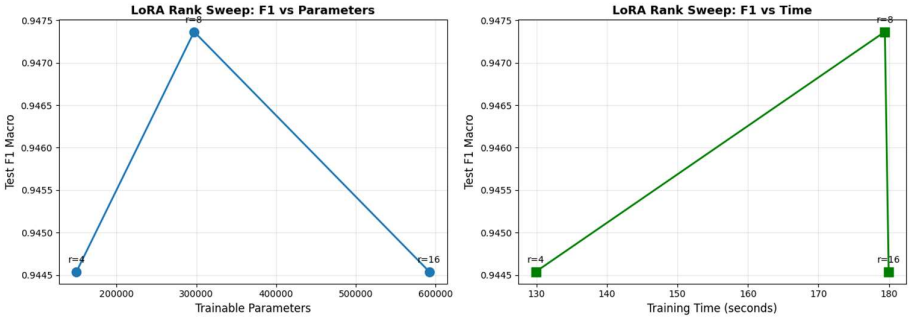

Figure 1. LoRA Rank FinBERT: Test F1 Macro vs. Trainable Parameters (left) vs. Training Time (right).

Confusion Matrix and Per-Class Analysis (FinBERT Full Fine-tuning and LoRA). Both FinBERT fully fine-tuned and LoRA (r=8) models demonstrate strong and consistent performance in classification behaviour across the three sentiment classes, with most predictions concentrated along the diagonal, indicating correct classifications. Notably, the number of neutral sentences incorrectly labelled as positive was reduced in the LoRA model compared to the fully fine-tuned model (1 vs 7), suggesting slightly better calibration between these classes. Overall, both models exhibit nearly identical confusion patterns, showing that LoRA can replicate full fine-tuning performance not only in aggregate metrics but also in per-class predictive behaviour, despite training less than 1% of the parameters.

Further per-class analyses supported these findings, showing that both FinBERT full fine-tuning and FinBERT LoRA (r = 8) achieved uniformly strong precision and recall across all three sentiment classes. Their per-class F1-scores differed by less than 0.03 on average, confirming that LoRA closely replicates full fine-tuning performance, even on minority classes.

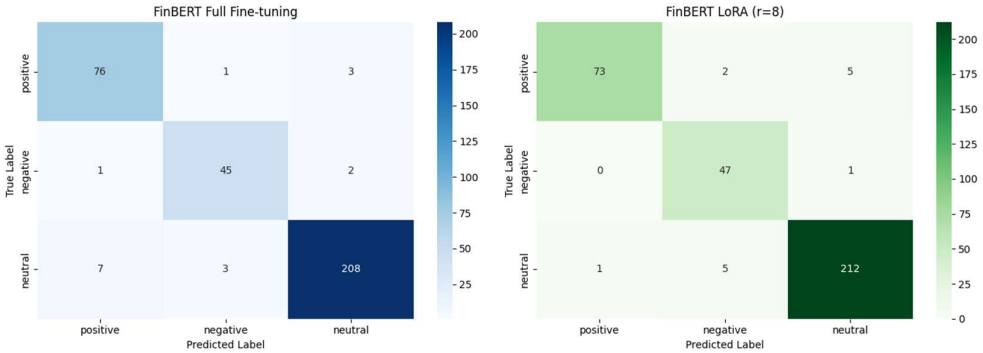

Figure 2. Confusion Matrices of FinBERT Full Fine-tuning and FinBERT LoRA (r=8)

Error Analysis. Out of 346 test samples, the FinBERT full fine-tuning and LoRA (r = 8) models disagreed on only 16 instances (agreement rate = 95.38%). *Table 2* highlights six such

examples — three where LoRA misclassified while full fine-tuning was correct, and three where the reverse occurred. For the cases LoRA misclassified, it may have relied on standout words within longer contexts, such as interpreting “loss” or “attacks” as clear negative cues, or “plan and implement” as positive. In contrast, the fully fine-tuned model appeared to struggle more with numeric, fact-based sentences, occasionally inferring sentiment from neutral financial statements. However, it is reasonable to see how the complexity of some of these sentences could be misclassified (e.g. case 32 is a positive improvement framed in negative language, while cases 4 and 122 contain negative-sounding phrasing).

Case	Excerpt (truncated)	True	Full FT	LoRA
	LoRA wrong; Full FT right			
32	The loss for the third quarter was EUR 0.3 mn smaller than the loss of the second quarter.	positive	positive	negative
68	Consumers have once again started to plan and implement building projects.	neutral	neutral	positive
4	… growing number of targeted malware attacks on individuals, companies, and organizations.	neutral	neutral	negative
	Full FT wrong; LoRA right			
45	The Pension Fund lost about $3.5 million in the Madoff Ponzi scheme.	negative	neutral	negative
54	Sampo Housing Loan Bank priced its EUR 1 bn bond at 99.889 %.	neutral	positive	neutral
122	… will furlough employees for less than 90 days.	neutral	negative	neutral

Table 2. Examples representing disagreements between FinBERT full fine-tuning and LoRA.

Key Takeaways and Limitations

This study shows that domain-adaptive pretraining contributes far more to financial sentiment classification than the fine-tuning strategy itself. As expected, FinBERT substantially outperformed the general-purpose BERT-base model in zero-shot evaluation (macro-F1 = 0.94 vs 0.13). Further fine-tuning revealed that LoRA achieved marginally higher performance compared to full fine-tuning (macro-F1 = 0.947 vs 0.940), while updating less than 1 % of parameters and reducing training time by over 60 %, underscoring the impressive value of parameter-efficient fine-tuning. Performance gains plateaued beyond moderate ranks, indicating that most task-specific knowledge can be captured with few trainable parameters. Error and confusion matrix analyses also showed both methods exhibit similar class-level behaviour.

The experiments remain limited by the dataset’s modest size and focus on English financial news from a single market, which may constrain generalizability. GPU peak memory was measured in Google Colab’s runtime environment and should be treated as an approximate measure. Future work, including the planned course project, could extend this evaluation to more diverse corpora in alternative financial markets and explore additional parameter-efficient fine-tuning methods (e.g., adapters or prompt-tuning).

Acknowledgement

Portions of the coding workflow and report writing were supported by ChatGPT (OpenAI, 2025). All outputs were reviewed, adapted, and verified by the author before inclusion in the final submission.

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research, 40*(2), 806-841.

Takala, P. (2022). *Financial PhraseBank* [Dataset]. Hugging Face. https://huggingface.co/datasets/takala/financial_phrasebank