Financial Sentiment Analysis using FinBERT Model

Arun Rimal, n264s646
School of Computing
College of Engineering
axrimal1@shockers.wichita.edu

Rajeet Chaudhary, j992y875
School of Computing
College of Engineering
rxchaudhary5@shockers.wichita.edu

***Abstract*—This paper presents a financial sentiment analysis framework using FinBERT, a pre-trained language model tailored to financial texts. We compare its performance with traditional machine learning and deep learning models including SVM, LSTM, and DistilBERT. By applying domain-specific fine-tuning and advanced training strategies, our approach outperforms baseline methods on benchmark datasets. The results demonstrate that FinBERT offers superior accuracy and robustness in classifying financial sentiment, making it a reliable tool for real-world economic text interpretation.**

I. Introduction

In the modern financial ecosystem, sentiment extracted from unstructured text sources—such as news articles, earnings reports, and social media—plays a critical role in shaping investor behavior and market dynamics. Accurately interpreting this sentiment can offer a competitive edge to financial analysts, traders, and decision-makers. However, financial language is complex, domain-specific, and often ambiguous, making it difficult for traditional natural language processing (NLP) models to perform effectively.

Recent advances in transformer-based models, particularly Bidirectional Encoder Representations from Transformers (BERT), have revolutionized NLP by enabling deep contextual understanding with minimal task-specific architecture. Building upon BERT, FinBERT is a domain-adapted variant pre-trained on financial corpora, designed specifically to tackle sentiment analysis tasks in finance. Prior research has shown that FinBERT significantly improves classification performance on benchmark datasets such as the Financial PhraseBank.

In this paper, we fine-tune FinBERT for financial sentiment classification and benchmark its performance against traditional models such as Support Vector Machines (SVM), Long Short-Term Memory networks (LSTM), and lighter transformer variants like DistilBERT. We also implement advanced training strategies—such as discriminative fine-tuning, gradual unfreezing, and focal loss—to enhance performance under class imbalance and limited data conditions. Our experiments demonstrate that FinBERT not only outperforms baseline models but also achieves state-of-the-art accuracy and robustness in classifying financial sentiments into positive, negative, and neutral categories.

II. Related Work

Sentiment analysis in financial texts has become a prominent area of research due to its applications in forecasting, algorithmic trading, and investor sentiment evaluation. Early studies employed traditional machine learning models such as Naive Bayes and Support Vector Machines, using TF-IDF and manually engineered features for sentiment classification [1]. Although efficient, these approaches lacked the ability to capture contextual and semantic nuances in financial language.

The emergence of deep learning models, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, allowed better handling of word sequences and temporal dependencies [2]. However, these models still faced challenges with long-range dependencies and lacked scalability for large corpora.

The introduction of transformer-based architectures, most notably BERT (Bidirectional Encoder Representations from Transformers), revolutionized natural language processing by introducing self-attention mechanisms that provide deep contextual understanding [3]. FinBERT, a domain-specific variant of BERT, was pre-trained on financial texts and achieved state-of-the-art performance on sentiment classification tasks involving financial news and reports [4]. Studies using the Financial PhraseBank dataset demonstrated that fine-tuning FinBERT significantly improves classification accuracy and generalization in finance-specific tasks [5].

Additionally, distilled transformer models like DistilBERT provide a lightweight alternative to BERT, maintaining competitive accuracy with reduced computational cost [6]. Enhanced fine-tuning methods, including focal loss, gradual unfreezing, and discriminative learning rates, have been explored to address challenges such as data imbalance and overfitting in sentiment classification [7].

In this work, we build upon these advances by comparing classical (Naive Bayes), sequential (LSTM), and transformer-based (DistilBERT, FinBERT) models, and introduce optimization techniques to further enhance FinBERT's effectiveness on imbalanced financial sentiment datasets.

III. Proposed Approach

We adopted and fine-tuned FinBERT, a domain-specific BERT model pre-trained on financial corpora, for sentiment classification in financial text. To improve performance and handle class imbalance, we implemented several enhancement techniques and benchmarked FinBERT against other models like LSTM, Naive Bayes, and DistilBERT.

A. Motivation

Financial sentiment has become a key driver of market behavior, influencing investment decisions and risk assessments. With the rapid increase in unstructured financial data from news articles, corporate reports, and social media, there is a growing need for automated systems that can accurately extract and interpret sentiment. Gauging market sentiment in real time offers strategic advantages, allowing investors to respond to shifting trends and uncertainty with data-backed insights. Moreover, sentiment analysis has broad applications in business operations, including risk management, credit scoring, fraud detection, and algorithmic trading. However, generic NLP models often underperform in this domain due to their lack of exposure to financial terminology and context. This motivates the use of FinBERT—a BERT-based language model further pre-trained on financial corpora—to bridge the gap between general-purpose language understanding and domain-specific sentiment detection. By fine-tuning FinBERT and incorporating additional training enhancements, our approach aims to provide a robust and accurate solution for financial sentiment classification.

B. Model Architecture

This study explores and compares five models for financial sentiment classification: FinBERT (baseline), FinBERT with enhancements, DistilBERT, LSTM, and Naive Bayes. These models span across classical machine learning, RNN-based architectures, and modern transformer-based language models. Below is a detailed breakdown of each architecture.

1) FinBERT (Baseline): FinBERT is a domain-specific variant of BERT (Bidirectional Encoder Representations from Transformers) designed for financial texts. It is based on the BERT-base architecture and pre-trained on financial documents such as analyst reports and market commentary.

- **Embedding Layer:** Combines word embeddings (30,522 tokens × 768), positional embeddings (512 × 768), and segment embeddings, followed by LayerNorm and dropout.
- **Transformer Encoder:** 12 encoder layers with 12 attention heads (64 dimensions per head), hidden size of 768, and a feedforward network of size 768 $\rightarrow$ 3072 $\rightarrow$ 768 with GELU activation.
- **Pooler Layer:** Applies a linear layer (768 $\rightarrow$ 768) followed by `tanh` on the [CLS] token representation.
- **Classification Head:** A dropout layer followed by a linear layer (768 $\rightarrow$ 3) with softmax activation.

Total Parameters: $\sim$109.5 million
Memory Usage: $\sim$563 MB

2) FinBERT with Enhancements: The enhanced FinBERT builds upon the baseline with several key modifications for improved generalization:

- **Focal Loss:** Replaces cross-entropy to emphasize difficult samples, using $\alpha = 1$ and $\gamma = 2$ [7].
- **Gradual Unfreezing:** Layers are unfrozen incrementally over 4 epochs to stabilize training and preserve pre-trained knowledge [8].
- **Discriminative Learning Rates:** Lower layers trained at 1e-5, middle at 2e-5 to 3e-5, and classifier head at 4e-5 to 5e-5.
- **Increased Dropout:** Dropout rate increased from 0.1 to 0.3 in attention and feedforward layers.

These enhancements led to significant improvements in F1 score and MCC, especially for imbalanced classes.

3) DistilBERT: DistilBERT is a distilled, lightweight version of BERT that retains 97% of its performance with 40% fewer parameters. It offers faster inference while maintaining strong accuracy.

- **Transformer Encoder:** 6 encoder layers with 12 attention heads and hidden size of 768.
- **Feedforward Network:** Linear layer with GELU activation and dropout.
- **Classification Head:** Linear projection (768 $\rightarrow$ 3) followed by softmax.

Total Parameters: $\sim$66.9 million
Memory Usage: $\sim$331 MB

4) LSTM: Long Short-Term Memory (LSTM) networks are RNNs capable of learning long-term dependencies. Our implementation is as follows:

- **Embedding Layer:** 100-dimensional trainable word embeddings.
- **LSTM Layers:** Two stacked LSTM layers, each with 512 hidden units.
- **Classifier:** Final hidden state passed to a dense layer (512 $\rightarrow$ 3) with softmax output.

Total Parameters: $\sim$3.3 million
Memory Usage: $\sim$14 MB

5) Naive Bayes: Naive Bayes is a simple yet effective probabilistic classifier based on Bayes' Theorem. It operates on vectorized input features.

- **Feature Extraction:** TF-IDF vectorization with up to 5000 features; preprocessing includes stopword removal and lowercasing.
- **Classification:** Multinomial Naive Bayes calculates posterior probabilities based on word frequency and selects the most probable class.

While efficient, Naive Bayes lacks the contextual understanding needed for nuanced sentiment classification.

To better understand the design and capabilities of each approach, we conducted a detailed architectural comparison across all implemented models. Table I presents a comprehensive comparison of the architectures and core characteristics of all five models used in this study: FinBERT (baseline), FinBERT with enhancements, DistilBERT, LSTM, and Naive Bayes.

C. Dataset

For this study, we utilized the Financial PhraseBank, a widely-used benchmark dataset curated by Malo et al. [9], which contains 4,845 financial sentences labeled by domain experts. Each sentence is assigned a sentiment label—positive,

TABLE I
DETAILED COMPARISON OF MODEL ARCHITECTURES

Feature	FinBERT without enhancemnet (Baseline)	FinBERT (Enhanced)	DistilBERT	LSTM	Naive Bayes
Base Model	BERT-base (12L)	BERT-base (12L)	DistilBERT (6L)	BiLSTM (2L)	Traditional ML
Embedding	30k vocab, 768d	30k vocab, 768d	30k vocab, 768d	10k vocab, 100d	TF-IDF (max 5k)
Dropout (Embed)	0.1	0.3	0.1	0.5	N/A
Encoder Layers	12 Transformer	12 Transformer + Dropout	6 Transformer	2 BiLSTM	N/A
Dropout (Other)	0.1	0.3	0.1 / 0.2	0.5	N/A
Hidden Size	768	768	768	256×2 (bi)	N/A
Classifier Head	Linear(768→3)	Linear(768→3)	Linear(768→3)	Linear(512→3)	Probabilistic
Enhancements	None	Focal Loss, Unfreezing, LR	None	None	None
Parameters	∼110M	∼110M	∼66M	∼3.3M	-
Training Speed	Slow	Slower	Fast	Fastest	Very Fast
Use Case	Financial sentiment	Imbalanced data	General NLP	Lightweight tasks	Low-resource inference

neutral, or negative—based on the consensus of human annotators . To ensure high label reliability, we used the subset where annotators had 100% agreement, thereby avoiding subjective inconsistencies. The dataset consists of short, headline-style financial statements extracted from real-world news articles, press releases, and reports which we can see in Figure 2. One notable challenge of the dataset is class imbalance, with a significantly higher proportion of neutral sentiments compared to positive and negative labels which we can see in Figure 1. This imbalance poses difficulties for standard classifiers and necessitates techniques to mitigate biased learning during model training.

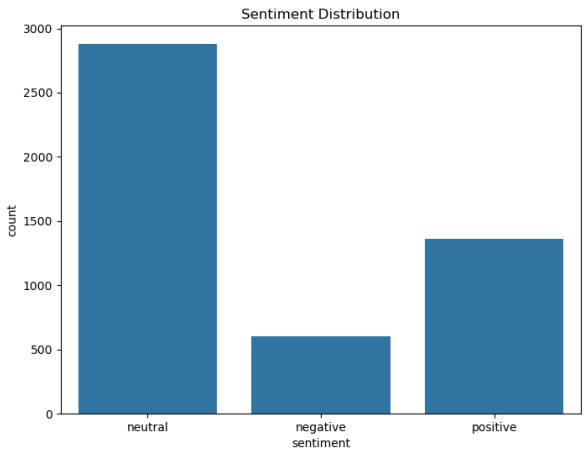

Fig. 1. Dataset

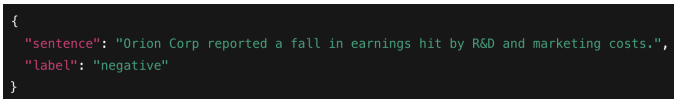

Fig. 2. Sentence Format along with the label

D. Training Procedure

The training process for our enhanced FinBERT model involved fine-tuning the pre-trained ProsusAI/FinBERT on the Financial PhraseBank dataset. The dataset was split into 80% for training and 20% for validation. To address class imbalance and improve learning efficiency, we implemented several advanced strategies during training.

We used the AdamW optimizer with weight decay regularization and applied a linear learning rate scheduler with warm-up steps. Discriminative learning rates were assigned to different layers—lower rates for earlier layers and higher rates for the classifier head—to preserve pre-trained knowledge while adapting effectively to the new task.

Gradual unfreezing [8] was applied over four epochs to stabilize the training process. In the first epoch, only the classification head was trained. In subsequent epochs, layers were progressively unfrozen: the last encoder layer in epoch two, the last two layers in epoch three, and all layers by epoch four. This approach helped prevent catastrophic forgetting and improved generalization.

Focal loss [7] was used as the loss function to address the imbalance between sentiment classes by focusing more on hard-to-classify examples. Dropout regularization (set to 0.3) was applied in both attention and feedforward layers to reduce overfitting.

The model was trained using a batch size of 16 for fifty epochs on a GPU-enabled environment. Throughout the process, we monitored training loss, validation accuracy, F1 score, and MCC to ensure consistent performance improvements.

To handle class imbalance, we incorporated focal loss, which assigns greater weight to misclassified or underrepresented classes. This process was repeated for baseline models (LSTM, Naive Bayes, DistilBERT, FinBERT without enhancements) to enable a consistent comparison. The combination of these training strategies contributed significantly to the performance gains observed in the enhanced FinBERT model.

IV. RESULTS

A. Evaluation Metrics

To evaluate the performance of our models, we employed a comprehensive set of classification metrics: Accuracy, Precision, Recall, F1 Score, and Matthews Correlation Coefficient (MCC). These metrics provide a balanced understanding of the model's behavior, especially in the presence of class imbalance. Accuracy reflects overall correctness, while Precision

and Recall offer insight into performance on specific sentiment classes. F1 Score provides the harmonic mean of Precision and Recall. MCC is particularly useful in imbalanced datasets, as it considers true and false positives and negatives, offering a more informative and reliable score [10].

B. Key Results

Table 1 presents a performance comparison of our enhanced FinBERT model with several baseline models including FinBERT without enhancements, DistilBERT, LSTM, and Naive Bayes. Our enhanced FinBERT model achieved the highest results across all metrics, with an accuracy of 86.48%, F1 score of 86.57%, and MCC of 0.7606. This confirms the effectiveness of our training strategies and the strength of domain-specific pretraining.

In comparison, the baseline FinBERT model also performed strongly but slightly lower, achieving an F1 Score of 85.56%. DistilBERT demonstrated competitive performance with a lightweight architecture, yielding an F1 Score of 83.34%. Naive Bayes achieved moderate performance (F1 Score: 61.01%), while LSTM showed the weakest performance among all models with an F1 Score of 69.18%, highlighting the limitations of sequential models in capturing complex financial language.

TABLE II
MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1 Score	MCC
DistilBERT	0.8338	0.8333	0.8338	0.8334	0.6985
LSTM	0.7038	0.6903	0.7038	0.6918	0.4361
Naive Bayes	0.6718	0.6590	0.6718	0.6101	0.3337
FinBERT (baseline)	0.8555	0.8564	0.8555	0.8556	0.7375
FinBERT (enhanced)	**0.8648**	**0.8679**	**0.8648**	**0.8657**	**0.7606**

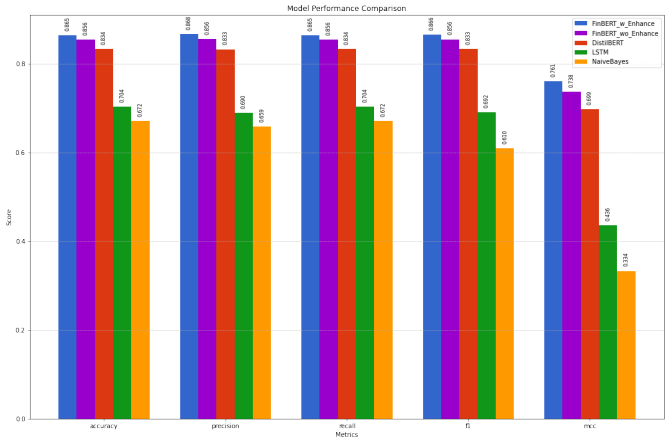

Fig. 3. Model performance comparison across Accuracy, Precision, Recall, F1 Score, and MCC.

Figure 3 illustrates the performance of various models across five key evaluation metrics. The enhanced FinBERT model (FinBERT with Enhancement) consistently outperformed all other models, achieving the highest scores in accuracy, precision, recall, F1 score, and MCC. This demonstrates the effectiveness of the applied fine-tuning techniques, including gradual unfreezing, focal loss, and discriminative learning rates.

While FinBERT without enhancements and DistilBERT also delivered strong results, they fell slightly short compared to the enhanced version, indicating the added value of advanced optimization strategies. On the other hand, LSTM and Naive Bayes showed noticeably lower performance—particularly in F1 score and MCC—highlighting their limitations in capturing the nuanced structure of financial language. Overall, transformer-based models, especially those fine-tuned for domain-specific tasks, proved to be significantly more capable for financial sentiment analysis than traditional approaches.

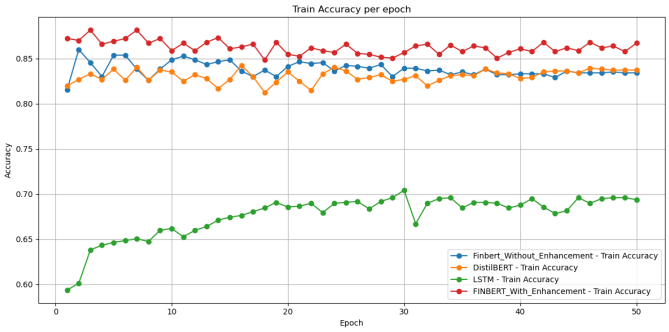

Fig. 4. Training Accuracy per Model over 50 epochs

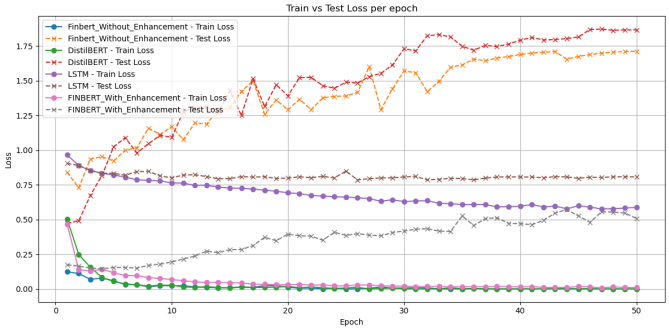

Fig. 5. Train vs. Test Loss per Model over 50 epochs

Alongside standard evaluation metrics, we extended the training over 50 epochs and monitored accuracy and loss progression to assess model stability and generalization over time. Figure 4 displays training accuracy across all models throughout the 50 epochs. FinBERT with enhancements consistently maintained the highest accuracy (ranging around 87–88%), while baseline FinBERT and DistilBERT followed closely behind, fluctuating between 82–85%. LSTM showed a gradual learning curve, improving from 59% to just under 70%, but still lagging behind transformer-based models.

Figure 5 presents a detailed comparison of training and test loss across epochs. Enhanced FinBERT displayed the lowest and most stable loss values for both training and testing, reflecting strong generalization. In contrast, DistilBERT and

baseline FinBERT exhibited increasing test losses after early epochs—an indication of overfitting. LSTM and Naive Bayes showed higher, flatter loss curves, highlighting their limited capacity to capture semantic nuance in financial language. These trends emphasize the benefit of advanced fine-tuning strategies applied in the enhanced FinBERT model.

C. Limitations

Despite the strong performance of the enhanced FinBERT model, several limitations remain. First, the dataset used—Financial PhraseBank which consisted of primarily short, sentence-level statements. This limits the model's exposure to the complexity and context present in longer financial documents such as earnings reports or regulatory filings. As a result, its generalization to more complex texts may be constrained.

Second, the model's performance might degrade when applied to significantly different financial text sources, such as informal language found in social media or conversational tone in earnings call transcripts. This domain shift could lead to reduced accuracy without further domain-specific fine-tuning.

Third, while FinBERT excels in predictive accuracy, it lacks interpretability. In high-stakes environments like finance, explainability is crucial for building trust and ensuring regulatory compliance. The black-box nature of transformer models presents a challenge in this regard.

Lastly, training the enhanced FinBERT model demands considerable computational resources. Due to time and resource constraints, the study was limited to the Financial PhraseBank dataset. Incorporating additional datasets in future work could improve the model's robustness and adaptability across diverse financial contexts.

V. Conclusion

In this paper, we presented a comprehensive approach to financial sentiment classification using FinBERT, a transformer-based language model pre-trained on financial corpora. By implementing advanced fine-tuning strategies such as gradual unfreezing, discriminative learning rates, and focal loss, we significantly improved model performance in terms of accuracy, F1 score, and Matthews Correlation Coefficient. Experimental results demonstrated that our enhanced FinBERT model outperformed baseline models including LSTM, Naive Bayes, and DistilBERT, achieving an F1 score of 86.57% and an MCC of 0.7606. These findings validate the effectiveness of domain-specific language modeling and targeted optimization techniques for sentiment analysis in financial texts. Future work may explore document-level classification, multi-label sentiment analysis, and deployment in real-time financial applications.

VI. Acknowledgment

The authors would like to express their sincere gratitude to Dr. Lokesh Das for his valuable guidance and support throughout this project. We also acknowledge the open-source implementation of FinBERT by Dogu Araci, which served as the foundation for our model architecture, and the Financial PhraseBank dataset curated by Malo et al., which was instrumental in training and evaluating our models.

References

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. EMNLP*, 2002.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

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

[5] P. Malo, A. Sinha, P. Takala, P. Korhonen, and J. Wallenius, "Good debt or bad debt: Detecting semantic orientations in economic texts," *JASIST*, vol. 65, no. 4, pp. 782–796, 2014.

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[7] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. ICCV*, 2017.

[8] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. ACL*, 2018.

[9] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," *arXiv preprint arXiv:1908.10063*, 2019. [Online]. Available: https://arxiv.org/pdf/1908.10063v1

[10] B. Agarwal and N. Mittal, *Machine Learning Approach for Sentiment Analysis*. Cham: Springer, 2016, pp. 21–45.