Layer-wise Analysis of FinBERT Embeddings for Enhanced Financial Sentiment

Alice Xiao[1], Aimee Yu[1], Grace Yuan[1], Celinda Zhu[1]
[1]Massachusetts Institute of Technology
{alicex31, aimeeyu, yuangc, celindaz}@mit.edu

Abstract

Domain-specific language models have shown promise in capturing the nuanced sentiment signals in financial text. While exisiting approaches commonly rely on final-layer embeddings, recent findings in natural language processing suggest that intermediate transformer layers encode distinct and potentially valuable semantic information. In this study, we conduct a comprehensie, layer-wise analysis of FinBERT and BERT embeddings for financial sentiment classification. We visualize layer representations via UMAP and evaluate their utility across mutiple downstream models, including Logistic Regression, Feedforward Neural Networks, and LSTMs. Our results show that FinBERT consistently outperforms general-purpose BERT, with intermediate layers often providing clearer sentiment separability and stronger predictive performance than the final layer. These findings demonstrate that selectively using embeddings from specific FinBERT layers may enhance model interpretability and efficiency, ultimately improving sentiment analysis tools within finance domain.

1 Introduction

In recent years, accurate financial sentiment analysis has emerged as a critical tool for informing investment decisions, shaping trading strategies, and guiding risk management in the financial markets. The ability to extract nuanced sentiment from news articles, analyst reports, and social media posts can offer decision-makers valuable insights into market sentiment and investor behavior. Within this domain, transformer-based language models have had a transformative impact, enabling the capture of subtle linguistic patterns that traditional methods often overlook.

FinBERT, a domain-specific adaptation of BERT trained on financial texts, has gained particular prominence due to its strong performance and specialized understanding of finance-related language. Most downstream applications of FinBERT in the literature rely heavily on final-layer embeddings, drawn from the last transformer block, to encode semantic information for classification tasks. While this approach is convenient and often effective, it may leave unexplored insights. Recent NLP research suggests that lower and intermediate model layers primarily capture grammatical and structural aspects of language, while layers closer to the output focus on more abstract, task-oriented representations. Building on these findings, the intermediate layers of FinBERT can thus encode a rich spectrum of both semantic and syntactic features.

This observation raises an important question: Are we overlooking beneficial information by only using the final-layer embeddings of FinBERT for financial sentiment classification? If intermediate layers offer valuable cues that influence downstream performance, identifying these layers could yield performance gains or help reduce computational complexity by allowing the use of shallower representations.

With this motivation, our work aims to conduct a systematic layer-wise examination of FinBERT and BERT embeddings for financial sentiment prediction tasks. We consider multiple classification models—ranging from simple, interpretable methods like Logistic Regression to more complex neural architectures such as Feedforward Neural Networks (FFNN) and Long Short-Term Memory networks (LSTM). Through this analysis, we seek to determine which layers are most informative and to develop a more nuanced understanding of how these transformer-based models process domain-specific financial language. Ultimately, this study would contribute to ongoing efforts to optimize model efficiency, interpretability, and performance in financial sentiment analysis.

2 Related Work

The emergence of transformer-based language models has substantially advanced natural language processing in various domains. BERT, introduced by Devlin et al. (2019) [1], demonstrated that deep bi-directional transformers trained on large corpora can effectively capture context-dependent semantic and syntactic features, paving the way for numerous domain-specific adaptations. In the financial sector, FinBERT (Araci, 2019 [2]), fine-tuned with financial texts from BERT, has become a cornerstone model for sentiment analysis, offering representations that align closely with the linguistic and conceptual nuances of financial text. Subsequent work by Yang et al. (2020) [3] and others built on FinBERT's final-layer embeddings to achieve strong results in tasks such as stock sentiment classification.

While leveraging final-layer embeddings from FinBERT is common, recent research on general-purpose BERT has revealed that different layers encode distinct types of linguistic information. Lower layers often focus on syntactic structures, while higher layers encode increasingly abstract semantic and task-relevant features (Jawahar et al., 2019 [4]; Tenney et al., 2019 [5]). To examine layer representations, visualization techniques like UMAP and t-SNE have been used, revealing clustering patterns and complexity of embeddings across layers, providing more insights(Coenen et al., 2019 [6]; Liu et al., 2020 [7]). While the value of examining internal representations in general NLP is recognized, the role of intermediate FinBERT layers in financial sentiment tasks remains underexplored.

In addition, while final-layer embeddings are commonly employed, several studies have noted that penultimate (second-to-last) or other intermediate layers can sometimes yield equal or even superior performance for certain downstream tasks. For instance, Liu et al. (2019) [8] and Merchant et al. (2020) [9] reported that representations from layers just below the final layer often strike an effective balance between raw syntactic cues and the overly task-specific features that can characterize the final layer. Such insights provide a precedent and theoretical grounding for examining whether intermediate FinBERT layers might similarly enhance performance in financial sentiment classification.

In parallel, various downstream models, ranging from simpler classifiers like Logistic Regression to more complex architectures such as FFNNs and LSTMs, have incorporated BERT-derived embeddings for sentiment-related tasks (Li et al., 2019 [10]; Xu et al., 2021 [11]). However, these efforts generally rely on embeddings from a single layer. Thus, the literature lacks a systematic, layer-wise evaluation of FinBERT for financial sentiment tasks—an analysis that promises to clarify which layers provide the most informative representations and potentially guide higher performance, more efficient and interpretable model deployment.

3 Methodology

We selected the **Labeled Financial PhraseBank (Hugging Face)** as our dataset, comprising annotated financial news sentences classified into positive, negative, or neutral sentiments. For our analysis, we chose the **"All Agree"** subset, to ensure the highest level of annotation consistency and minimize ambiguity in sentiment classification (Takala, 2014 [12]).

3.1 Embedding Analysis via UMAP Projection

To visualize the layer embeddings, we applied the **Uniform Manifold Approximation and Projection (UMAP)** algorithm (McInnes et al., 2018 [13]). UMAP was chosen for its ability to preserve both local and global relationships in high-dimensional data while projecting it into a two-dimensional space. Preserving these relationships is crucial for financial sentiment analysis, as it helps reveal the nuanced contextual structures within the embeddings.

The resulting clusters were examined to evaluate the models' abilities to capture sentiment-specific features at various depths. This analysis provided qualitative insights into the embeddings of financial texts by BERT and FinBERT, setting a foundation for our subsequent classification tasks.

3.2 Sentiment Analysis with BERT and FinBERT

- **Extract Embeddings:** We modified the output of BERT and FinBERT to obtain embeddings from every hidden layer, aiming to learn the contextual information encoded at each level.

- **Use as Input Features:** These embeddings were then fed as input features into downstream neural network models, designed to predict sentiment labels.

To evaluate the sentiment classification performance across different layers, we utilized three models individually: Logistic Regression (LR), Feedforward Neural Network (FFNN), and Long Short-Term Memory (LSTM).

- **Logistic Regression (LR):** The embeddings were treated as fixed input features for Logistic Regression. This simple model established a baseline to assess the utility of embeddings from each layer.

- **Feedforward Neural Network (FFNN):** This model was employed because it can capture the non-linear relationships within the embeddings. We choose the Adam optimizer and Cross-Entropy Loss for training.

- **Long Short-Term Memory (LSTM):** The LSTM model, known for its ability to capture sequential dependencies, explored the temporal patterns embedded in the contextual representations. Similar to FFNN, we choose the Adam optimizer and Cross-Entropy Loss.

Each model was trained and evaluated separately on embeddings extracted from every layer of BERT and FinBERT, allowing us to systematically compare their performance across layers and architectures.

3.3 Evaluation

To assess the performance of our sentiment classification models, we employed standard evaluation metrics widely used in NLP tasks – **Accuracy** and **F1-score**. These metrics provided a comprehensive view of the models' abilities to correctly classify sentiments.

The models were benchmarked on embeddings extracted from every layer of BERT and FinBERT, enabling a detailed comparison of their utility in sentiment classification tasks. This analysis highlighted which layers provided optimal representations for financial text.

The dataset was divided into **80% for training and 20% for testing**, ensuring the evaluation was conducted on unseen data to prevent overfitting.

The implementation was carried out in **Python**. Logistic Regression was implemented using `scikit-learn`, which also supplied metrics like accuracy and F1-score. For FFNN and LSTM, we used `PyTorch`. Dimensionality reduction was performed using the `umap-learn` library to project high-dimensional embeddings into a two-dimensional space.

To handle data processing tasks, we used `pandas` and `numpy` for efficient manipulation of the dataset. Pre-trained models (BERT and FinBERT) and tokenizers were loaded using the `transformers` library from Hugging Face. Additionally, the `datasets` library was used to load and preprocess the Labeled Financial PhraseBank dataset, ensuring smooth integration into the modeling pipeline.

All experiments were conducted on Google Colab with the A100 GPUs. The computational setup facilitated efficient training, especially for LSTM, while FFNN and Logistic Regression required significantly less training time due to their simplicity.

4 Results

4.1 UMAP Visualization of Layer Embeddings

We applied UMAP to project the high-dimensional representations from all hidden layers into a two-dimensional space. Figures 1 and 2 illustrate the resulting visualizations for BERT and FinBERT, respectively, across their hidden layers.

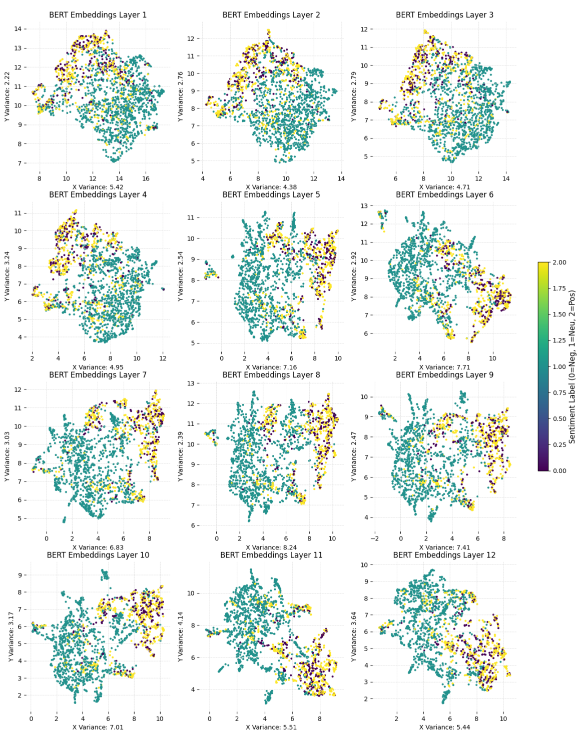

Figure 1: UMAP Projection of BERT Layer-wise Embeddings

The UMAP projection for **BERT** embeddings (Figure 1) shows a progression in clustering quality across the layers. In the earlier layers (e.g., Layers 1 to 4), sentiment-specific clusters (the negative, neutral, and positive) are less defined, since there exists a significant overlap between classes. This reflects that BERT initially focuses on capturing general linguistic features. In the middle layers (e.g., Layers 5 to 8), the clusters become more distinct, indicating that these layers encode richer contextual representations. However, in the later layers (e.g., Layers 9 to 12), the cluster separability begins to degrade, potentially due to task-specific encoding aligned with BERT's pretraining objectives (e.g., masked language modeling) (Jawahar et al., 2019 [14]).

In contrast, the UMAP projection for **FinBERT** embeddings (Figure 2) shows more cohesive clusters across all layers. Even in the earlier layers (e.g., Layers 1 to 3), FinBERT demonstrates better separation between sentiment classes compared to BERT. This trend continues through the middle layers (e.g., Layers 4 to 8), where the sentiment clusters are most distinct, reflecting FinBERT's ability to capture domain-relevant contextual nuances effectively. Unlike BERT, the separability of clusters remains relatively stable in the later layers (e.g., Layers 9 to 12), suggesting that FinBERT retains sentiment-specific information throughout its depth.

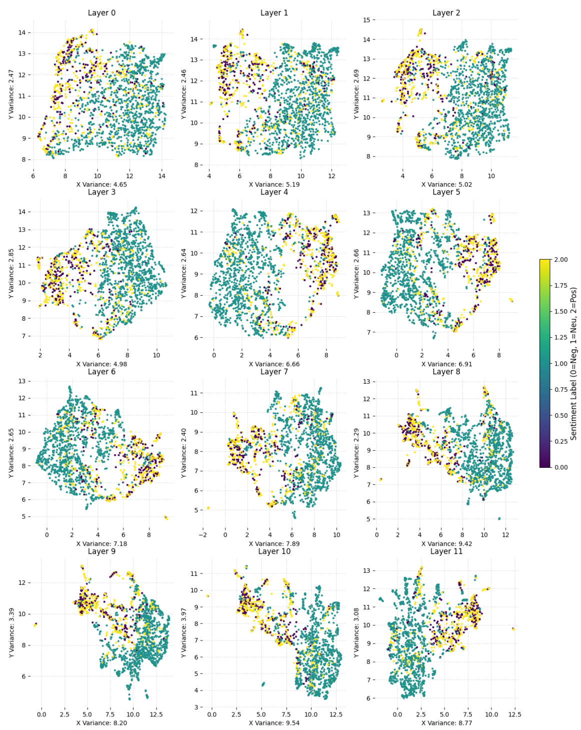

Figure 2: UMAP Projection of FinBERT Layer-wise Embeddings

Overall, the comparison shows that while both models reach their peak separability around layer 9, **FinBERT consistently outperforms BERT in terms of cluster coherence and sentiment-specific separability across all layers**.

4.2 Performance Analysis

In this section, we present a detailed comparison of FinBERT and BERT embeddings across different transformer layers. We train three models - Logistic Regression (LR), Feedforward Neural Network (FFNN), and Long Short-Term Memory (LSTM) networks - on layer-wise embeddings to evaluate their sentiment classification performance. Our objective is to identify where the most informative features for financial sentiment are, and how different downstream models adopt these representations.

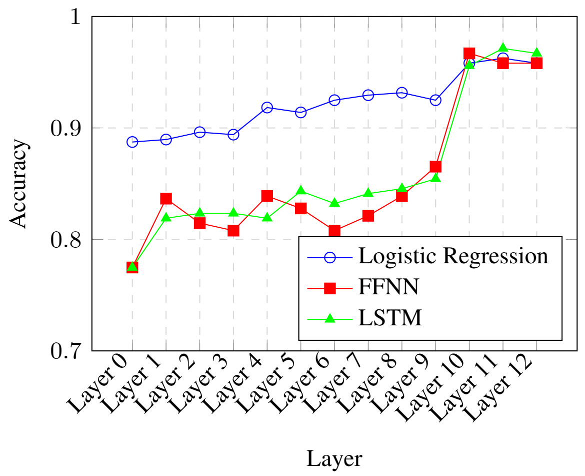

Figure 3: Trends of Accuracy Across FinBERT Layers for Different Models

4.2.1 Layer-wise Accuracy Trends

We begin by examining how accuracy changes as embeddings are extracted from progressively deeper layers of FinBERT. Table 1 and Figure 3 show that all three models generally achieve higher accuracy when using embeddings from later layers of FinBERT. Early layers exhibit lower performance, suggesting that these layers focus more on basic syntactic features rather than the nuanced financial sentiment. As we move toward the upper layers, these representations become richer and more aligned with the complexities of financial language.

Interestingly, while the last few layers demonstrate a strong overall performance, the final layer does not consistently outperform the other layers (Layer 10 and 11). For example, in the LSTM model, Layer 11 achives the highest accuracy (97.13%, slightly outperforming both Layer 10 and the final Layer 12. This suggests that the final layer, while capturing task-specific representations, might be more susceptible to overfitting the pretraining objectives.

Layer	Logistic Regression	FFNN	LSTM
Layer 0	0.8874	0.7748	0.7748
Layer 1	0.8896	0.8366	0.8190
Layer 2	0.8962	0.8146	0.8234
Layer 3	0.8940	0.8079	0.8234
Layer 4	0.9183	0.8389	0.8190
Layer 5	0.9139	0.8278	0.8433
Layer 6	0.9249	0.8079	0.8322
Layer 7	0.9294	0.8212	0.8411
Layer 8	0.9316	0.8389	0.8455
Layer 9	0.9249	0.8653	0.8543
Layer 10	0.9581	**0.9669**	0.9558
Layer 11	**0.9625**	0.9581	**0.9713**
Layer 12	0.9581	0.9581	0.9669

Table 1: Model Accuracy Comparison Across FinBERT Layers

4.2.2 Comparison of FinBERT and BERT

To assess the benefits of domain-specific training, we compare FinBERT with BERT on Logistic Regression (Figure 4), FFNN (Figure 5), and LSTM models (Figure 6). We found that FinBERT consistently outperforms BERT across layers and model architectures. This improvement is most pronounced in the upper layers, implying that FinBERT's domain adaption allows it to capture sector specific sentiment sigmores more effectively than a general-purpose model like BERT. Notably, both models show upward trends in accuracy as we ascend through the layers, but FinBERT's performance gains are more substantial and more stable.

4.2.3 Model-Specific Observations

1. **Logistic Regression:**
   The LR model, as shown in Table 1 and Figure 4, benefits from increasing layer depth but exhibits a relatively gradual performance improvement. Being a linear model, LR can utilize the improved linear separability of sentiment signals at higher layers. However, it lacks the complexity to fully exploit the subtle non-linear features embedded in deeper FinBERT layers.

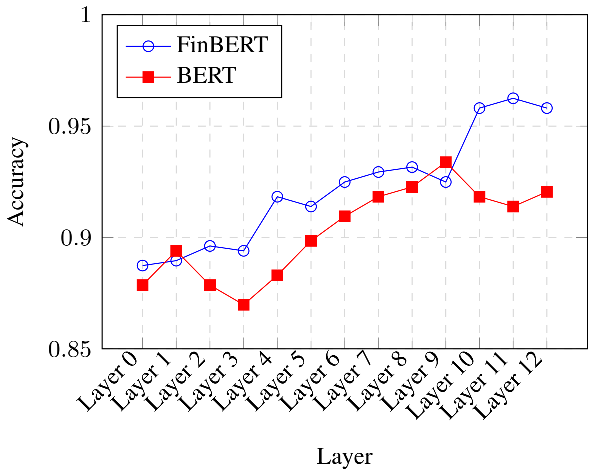

Figure 4: Comparison of BERT and FinBERT Accuracy on Logistic Regression Across Layers

2. **Feedforward Neural Network (FFNN):**
   The FFNN results in Figure 5 and Table 1 reveal a more pronounced gain in accuracy at the top layers of FinBERT. Non-linear transformations in the FFNN can make use of the richer semantic representations from upper layers, translating into marked improvements in accuracy, particularly beyond Layer 8. This pattern suggests that more advanced hidden representations, likely encoding domain-specific nuances, are crucial for improved sentiment classification.

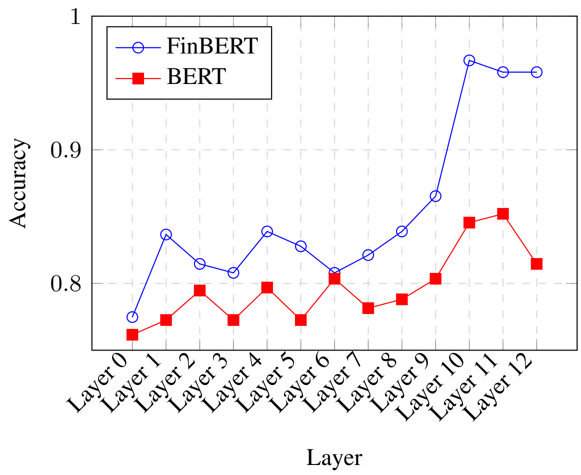

Figure 5: Comparison of BERT and FinBERT Accuracy on FFNN Across Layers

3. **Long Short-Term Memory (LSTM) Networks:**

The LSTM results (Figure 6) show the most dramatic improvements at higher layers of FinBERT. The LSTM architecture excels at modeling sequential dependencies and capturing contextual information. When provided with embeddings that already encode sophisticated semantic and sentiment-related cues, the LSTM can integrate these signals more effectively than simpler models. As a result, the LSTM achieves its highest accuracy in the penultimate and final layers, surpassing both LR and FFNN. This outcome underscores how model architectures that leverage temporal and contextual cues benefit disproportionately from rich, domain-specific representations.

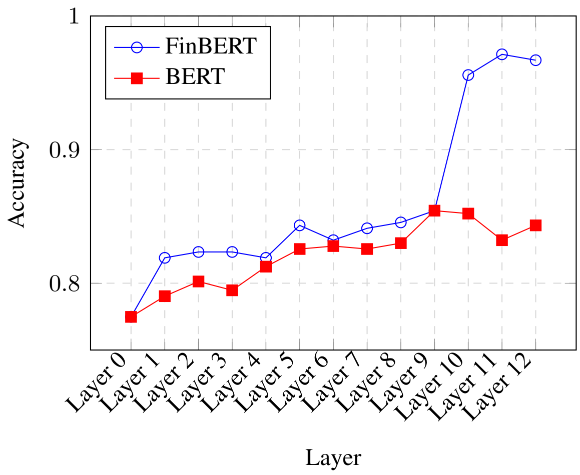

Figure 6: Comparison of BERT and FinBERT Accuracy on LSTM Models Across Layers

5 Conclusion

The layer-wise exploration of FinBERT and BERT embeddings for financial sentiment classification in our study clearly shows that domain-specific models can capture more nuanced and contextually rich financial language patterns or semantic meaning. By examining multiple classification paradigms, from logistic regression to FFNNs and LSTMs, we observed that intermediate FinBERT layers consistently yielded richer and more linearly separable embeddings that directly translated into stronger downstream performance. It has also been established that the last layer (Layer 12) is outperformed by previous layers 10 or 11, which is seen across models and true for both BERT and FinBERT. The rise in accuracy of these layers suggest that it probably encapsulated the critical semantic and contextual signals needed for accurate sentiment classification while BERT failed to do so, validating the intuition that domain adaptation at the embedding level can yield sizable gains in both interpretability and efficiency.

Together, these findings validate the intuition that domain adaptation at the embedding level can yield sizable gains in both interpretability and efficiency. By demonstrating that intermediate layers hold the key to improved performance, this study paves the way for more targeted layer utilization strategies in financial NLP pipelines. As a result, practitioners can more confidently select, tune, and deploy transformer-based models tailored to their domain-specific demands, ultimately fostering more robust and effective financial sentiment analysis solutions.

6 Limitations and Future Work

Although this study provides a clearer understanding of how FinBERT's internal layers contribute to financial sentiment classification, it has certain limitations. First, our findings are focused on one domain and task. While FinBERT is well-suited to financial text, it is not yet clear whether these insights apply to other types of financial NLP tasks, such as risk assessment, event prediction, or credit scoring. Second, we have primarily analyzed embeddings from single layers, without exploring whether combining multiple layers or using more advanced aggregation methods could yield richer, more informative representations.

Future work could address these limitations by expanding the scope of analysis. For example, researchers could evaluate layer-wise embeddings in a broader set of financial NLP tasks, including named entity recognition in corporate filings or identifying market-moving events in news articles. Additionally, applying the same methodology to other domain-specific BERT-based models—such as those adapted for legal or biomedical texts—could help determine how domain adaptation affects the nature of information encoded at different layers. Exploring layer combinations and advanced aggregation techniques may also reveal new opportunities to enhance model performance and interpretability.

7 Impact Statement

The enhanced understanding and interpretability and embeddings in FinBERT in layer-wise analysis have the potential to improve the precision and reliability of financial sentiment analysis, thereby influencing decisions made by investors, analysts and policy makers in relative fields. By identifying which layers produce the most semantically rich features, developers can reduce the model's complexity, thus lowering computing costs while enhancing accessibility. This may democratize access to advanced sentiment modeling, increasing the number of market participants who can use cutting-edge analysis methods. However, this increased accessibility and influence on decision making also raises ethical concerns; for instance, reliance on data-driven language models could inadvertently skew financial forecasts, produce systemic inequalities, or misinterpret semantic signals leading to unintended biases. Therefore, it is important to ensure that these embeddings are not only powerful, but also unbiased and transparent. Failure to address these issues could undermine trust in automated sentiment analysis systems and lead to negative market outcomes.

From a societal perspective, more nuanced financial language-processing models utilizing FinBERT or other fine-tuned versions of LLMs could influence how market-related information is distributed, potentially impacting individual investors' financial well-being and shaping public policy and regulation of the market. As these technologies evolve, their effects will go beyond the trading floor, shaping lending decisions, risk evaluations, and broader economic policies. Newer or more advanced domain-specific models shall not be put forward without implementing guidelines that emphasize fairness, security, and responsible AI practices in financial analytics. Developing robust frameworks, curating balanced datasets, and incorporating human-in-the-loop strategies (such as human annotated-labels or dataset reflecting equal representation) are crucial steps toward ensuring that FinBERT-based sentiment analysis systems serve as reliable and equitable financial tools. Ultimately, the responsible deployment of these models can promote fairer financial ecosystems and more informed decision making, provided that ethical considerations remain integral to their ongoing refinement and use.

References

[1] Jacob Devlin et al. "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 (NAACL-HLT)* (2019). URL: https://arxiv.org/abs/1810.04805.

[2] Didem Araci. "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models". In: *arXiv preprint arXiv:1908.10063* (2019). URL: https://arxiv.org/abs/1908.10063.

[3] Liuqing Yang et al. "FinBERT: A Deep Learning-based Financial Sentiment Classifier". In: *Proceedings of the First Workshop on Financial Technology and Natural Language Processing (FinNLP) at COLING* (2020). URL: https://aclanthology.org/2020.finnlp-1.1.

[4] Ganesh Jawahar, Benoıt Sagot, and Djamé Seddah. "What Does BERT Learn About the Structure of Language?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)* (2019). URL: https://aclanthology.org/P19-1356.

[5] Ian Tenney, Dipanjan Das, and Ellie Pavlick. "BERT Rediscovers the Classical NLP Pipeline". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)* (2019). URL: https://aclanthology.org/P19-1452.

[6] Andy Coenen et al. "Visualizing and Measuring the Geometry of BERT". In: *Advances in Neural Information Processing Systems (NeurIPS)* (2019). URL: https://proceedings.neurips.cc/paper/2019/hash/f2c9371b52f1e9c65f0fd6f617f3b74f-Abstract.html.

[7] Sijie Liu et al. "Visual Exploration of Semantic Relationships in Neural Word Embeddings". In: *IEEE Transactions on Visualization and Computer Graphics* (2020). URL: https://doi.org/10.1109/TVCG.2020.3030439.

[8] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv preprint arXiv:1907.11692* (2019). URL: https://arxiv.org/abs/1907.11692.

[9] Nazneen Fatema Merchant, Anupama Sastry, and Steve Tratz. "Is the Last Layer Really All You Need? Revisiting Transformer-based Models and Their Performance". In: *Findings of the Association for Computational Linguistics: EMNLP 2020* (2020). URL: https://aclanthology.org/2020.findings-emnlp.270.

[10] Changjin Li, Housheng Wu, and Jiasi Wang. "Exploiting BERT for Stock Price Prediction". In: *2019 IEEE Symposium Series on Computational Intelligence (SSCI)* (2019). URL: https://doi.org/10.1109/SSCI44817.2019.9002992.

[11] Yuan Xu et al. "Stock Movement Prediction with Sentiment Embedding and Domain Adaptation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* (2021). URL: https://ojs.aaai.org/index.php/AAAI/article/view/17291.

[12] P. Takala. *Financial PhraseBank: Sentences for Financial Sentiment Analysis*. URL: https://huggingface.co/datasets/takala/financial_phrasebank/blob/main/README.md.

[13] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[14] Ganesh Jawahar, Benoıt Sagot, and Djamé Seddah. "What Does BERT Learn about the Structure of Language?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3651–3657. DOI: 10.18653/v1/P19-1356. URL: https://aclanthology.org/P19-1356.