Comparative Evaluation of Neural Architectures for Sentiment Analysis on Financial News: From MLPs to Transformers

Josh Le Grice - 720017170

1 Introduction

Traditional financial models tend to focus on quantitative inputs such as stock prices and macroeconomic indicators. However, these inputs rarely capture the full picture of market behaviour [1]. In light of this, there has been growing interest in integrating sentiment analysis to account for the emotional and psychological factors that influence investor decisions [1]. In previous research, social media has received the majority of discussion within this area, with news content receiving less attention despite its proven potential [2].

Traditional machine learning (ML) methods provide a foundation for sentiment analysis but they may underperform in more complex domains like finance, due to the need for domain-specific feature engineering, limiting their ability to understand complex or subtle sentiments that are common in financial news [3]. Manual feature engineering can potentially introduce bias, leading to reduced generalisability on unseen datasets, a critical consideration when making financial market predictions. To combat this, deep learning methods such as RNNs, LSTMs and Transformers offer a more accurate alternative, stemming from their ability to capture sequential patterns, making them better suited for analysing sentiment in financial news [3].

For example, FinBERT, developed by Huang et al. [4], is a finance-specific language model trained on financial documents and news. It outperforms general-purpose BERT and other ML methods, e.g. Naive Bayes and LSTMs, in extracting sentiment from financial text, highlighting its advantages in domain-specific deep learning models [4]. However, while FinBERT provides accurate embeddings, the choice of neural architecture used as a classifier on top of these embeddings may impact performance [5]. These trade-offs between complexity and accuracy are scarcely researched in the finance domain.

This project compares multiple neural architectures when classifying sentiment in financial news headlines, investigating whether architectural complexity leads to performance gains. While sentiment predictions could inform market movements, evaluating this link is beyond the scope of this project, and will be discussed in future work.

2 Dataset Description

This project uses the *English Financial News* dataset [6], which contains approximately 27,000 financial news headlines collected from Twitter and Yahoo Finance. Each headline is labelled as *positive*, *negative*, or *neutral* based on its perceived sentiment.

The dataset was selected due to its large and diverse corpus of real-world financial language, which clearly shows the types of news content that influence investor sentiment and market movements.

Given the unstructured nature of textual data, preprocessing was required before training the models. The following steps were applied:

- Removal of URLs, stock tickers, special symbols, and unnecessary whitespace.
- Tokenisation and vectorisation of each sentence to convert text into numerical vectors suitable for neural network input.

The dataset was split into training (80%) and testing (20%) sets to enable model evaluation. While well-suited for sentiment classification tasks, the dataset contains a class imbalance, with *neutral* sentiment being most common (62.5%). This imbalance was addressed during model training through stratified sampling and weighted metrics to mitigate bias towards the majority *neutral* class.

3 Methods

This project treats financial news sentiment prediction as a three-class classification problem, labelling each headline as *positive*, *negative*, or *neutral*. A baseline TF-IDF + Logistic Regression model was compared with three neural architectures: Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), and Transformer, trained on FinBERT embeddings from Prosus AI [7].

The baseline uses TF-IDF embeddings [8] and Logistic Regression classification [9], which is commonly included in similar research [10], allowing for a comparison between FinBERT-based models and traditional feature engineering approaches. For neural architectures, each headline was encoded using pre-trained FinBERT embeddings [4]. As stated in Section 1, FinBERT typically outperforms general-purpose models in financial text tasks due to its domain-specific training, making it a strong foundation for this sentiment classification task [4]. These embeddings were then passed through three classification architectures: an MLP with ReLU activation [11] as a neural baseline, a bidirectional LSTM to model sequential relationships [12], and a Transformer with self-attention [13] as a state-of-the-art architecture.

Due to computational constraints, all neural models were trained only once, on a T4 GPU, with a single set of hyperparameters, on 80% of the data and evaluated on the remaining 20% using various performance metrics. Full details on training parameters, optimisation, regularisation, and architectures are provided in Appendix A (Table 2), along with training dynamics which can be found in Appendix C (Figures 2, 3, and 4).

4 Results

Results indicate that all FinBERT-based neural architectures significantly outperformed the TF-IDF + Logistic Regression baseline, highlighting the importance of high-quality embeddings in financial sentiment analysis. Furthermore, increasing architectural complexity led to classification improvements, albeit with diminishing returns.

The Transformer architecture achieved the highest overall accuracy at 90.3% followed by LSTM (89.7%), MLP (88.2%) and Logistic Regression (85.4%) (Table 1). Macro F1-scores ranged from 0.814 to 0.876, indicating consistent performance gains across all sentiment classes, while ROC-AUC scores (0.942 – 0.969) demonstrated strong classification performance even for the baseline model.

However, neural architectures required 32.6 – 35.2 minutes of training compared to 0.057 minutes for Logistic Regression, a 570 – 620× increase for 2.8% (MLP), 4.3% (LSTM), 4.9% (Transformer) absolute accuracy gain. Confusion Matrices (Figure 1) and detailed per-class metrics (Appendix B, Table 3) show that neutral sentiment classification was strongest across all architectures, while negative sentiment remained most challenging, indicating potential issues caused by the imbalances within the dataset. Finally, qualitative analysis of test predictions revealed consistent struggles with subtle positive sentiment when lacking clear growth indicators, with the models often classifying these headlines into the neutral class (Appendix B, Table 4).

5 Discussion

The improvement in performance from Logistic Regression to the Transformer demonstrates that more complex architectures improve financial sentiment classification. Although the gains between FinBERT-based models were small (MLP to Transformer: 2.1%), suggesting an element of diminishing returns once you have high-quality embeddings these small gains could be important in high-stakes financial settings, where even small improvements can have large financial implications.

The Transformer's self-attention mechanism [13] helped with technical financial language, but whether this justifies the 8% increase in training time when compared to the other neural architectures is unclear, especially when all models already achieved high ROC-AUC scores.

Despite promising results, there were a few limitations that affected this research. The dataset was skewed towards neutral sentiment (62.5%), causing all models to perform better on the neutral class (F1: 0.898 - 0.931) than negative classes (F1: 0.733 - 0.830). Additionally, keeping FinBERT embeddings frozen saved model training time but prevented dataset-specific fine-tuning. Furthermore, test examples showed that all models struggled with implicit positive sentiment, where headlines described a company's strengths rather than clear growth language. This suggests that capturing subtle financial sentiment remains a challenge regardless of architectural choice.

When deploying these in practice, simpler models might be preferable if frequent retraining is needed, while the Transformer's accuracy could justify its complexity for strategic decision-making. Furthermore, in high-stakes financial applications, the interpretability of simpler models like Logistic Regression or Tree-based methods might be more desirable, introducing the need for SHAP or LIME analyses on neural models [14], [15].

6 Conclusions

This project compared four approaches to sentiment classification in financial news headlines, showing that architectural complexity provides measurable performance improvements. Transformers showed the best performance; however, gains over simpler FinBERT-based neural models were small. The results suggest that high-quality embeddings may matter more than classifier architectural complexity, with all neural models clearly outperforming the baseline. However, statistical significance testing is required to provide definitive conclusions.

Future work should address the class imbalance through data augmentation and explore whether FinBERT embeddings improve performance enough to justify the increased training cost, as well as using more extensive model optimisation and discussing statistical significance. Additionally, future research could explore the

practical application of these models by evaluating if these sentiment predictions can inform market movements more accurately than traditional quantitative models.

Figures & Tables

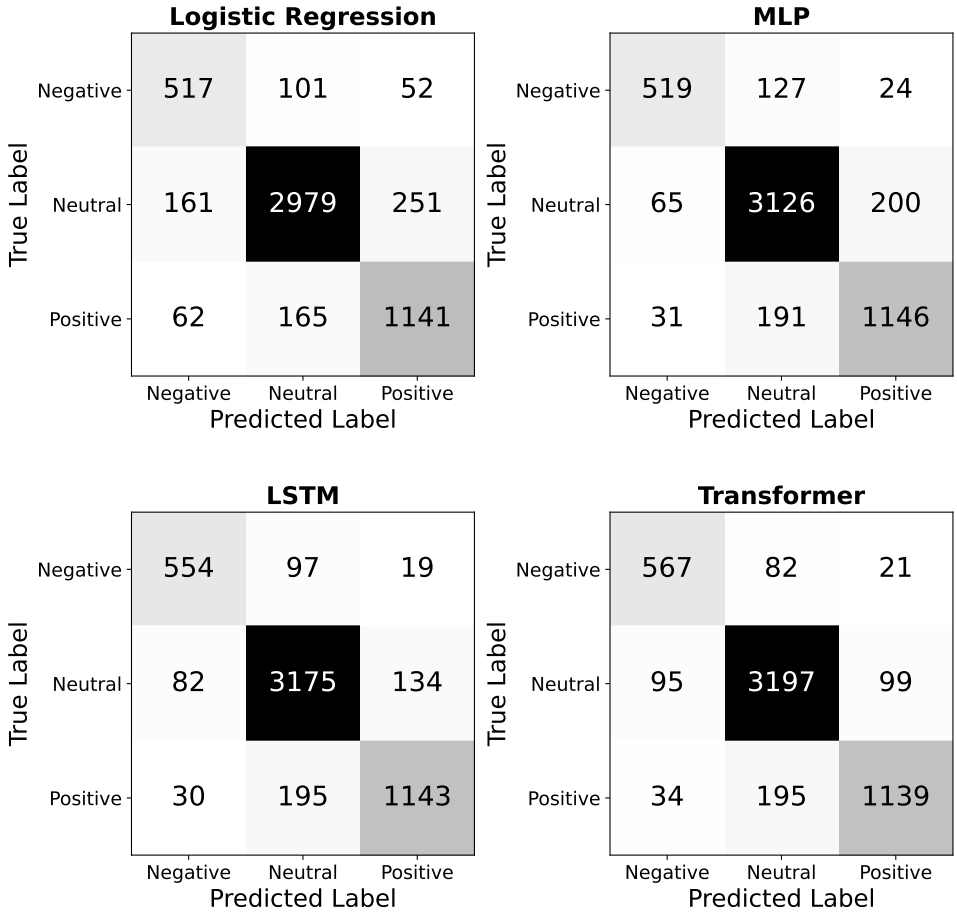

Figure 1: Confusion Matrices for All Models (Logistic Regression, MLP, LSTM, and Transformer) This figure shows the classification performance across all four models on the test set. Darker cells indicate higher prediction frequencies. All models performed best on the neutral sentiment, with negative sentiment classification shown to be most difficult. The Transformer achieved the highest true positive values indicating higher overall performance, while the Logistic Regression baseline shows more misclassifications.

Table 1: Summary of Model Performance Across All Architectures

Metric	Logistic Regression	MLP	LSTM	Transformer
Accuracy	0.854	0.882	0.897	0.903
Precision (Weighted)	0.859	0.882	0.897	0.903
Recall (Weighted)	0.854	0.882	0.897	0.903
F1 (Weighted)	0.856	0.882	0.897	0.903
Precision (Macro)	0.802	0.863	0.877	0.880
Recall (Macro)	0.828	0.845	0.866	0.874
F1 (Macro)	0.814	0.853	0.871	0.876
ROC-AUC (Macro)	0.942	0.962	0.966	0.969
Loss	–	0.542	0.500	0.493
Training Time (min)	0.057	32.578	32.602	35.226

Table 1 shows the summary of model performance across all architectures evaluated on the test set. The Transformer demonstrates the highest performance across all classes, followed by the LSTM and MLP. All FinBERT based models clearly outperform the baseline. Training times for neural architectures were 570 – 620x longer than the baseline, representing a large computational trade-off for 4.9% accuracy increase. ROC-AUC scores above 0.94 across all models indicate strong classification ability.

References

[1] A. Atkins, M. Niranjan, and E. Gerding, "Financial news predicts stock market volatility better than close price," *J. Finance Data Sci.*, vol. 4, no. 2, pp. 120–137, 2018, doi: https://doi.org/10.1016/j.jfds.2018.02.002.

[2] M. N. Ashtiani and B. Raahemi, "News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review," *Expert Syst. Appl.*, vol. 217, no. 119509, p. 119509, 2023, doi: https://doi.org/10.1016/j.eswa.2023.119509.

[3] H. Bashiri and H. Naderi, "Comprehensive review and comparative analysis of transformer models in sentiment analysis," *Knowl. Inf. Syst.*, vol. 66, no. 12, pp. 7305–7361, 2024, doi: https://doi.org/10.1007/s10115-024-02214-3.

[4] A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text," *Contemp. Acc. Res.*, vol. 40, no. 2, pp. 806–841, 2023, doi: https://doi.org/10.1111/1911-3846.12832.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018, doi = https://doi.org/10.48550/arxiv.1810.04805.

[6] L. Calarte, "English_Finance_News," 2023, accessed: 2025-10-10. [Online]. Available: https://huggingface.co/datasets/lukecarlate/english_finance_news

[7] ProsusAI. (2023) ProsusAI/finbert · hugging face. Accessed: 2025-10-18. [Online]. Available: https://huggingface.co/ProsusAI/finbert

[8] L. Havrlant and V. Kreinovich, "A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation)," *International journal of general systems*, vol. 46, no. 1, pp. 27–36, 2017.

[9] C. M. Bishop, *Pattern recognition and machine learning*, ser. Information science and statistics. New York: Springer, 2006.

[10] A. Karanikola, G. Davrazos, C. M. Liapis, and S. Kotsiantis, "Financial sentiment analysis: Classic methods vs. deep learning models," *Intell. Decis. Technol.*, vol. 17, no. 4, pp. 893–915, 2023, doi: https://doi.org/10.3233/idt-230478.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. Alanna Maldonado, 2023.

[12] A. Sharaff, T. R. Chowdhury, and S. Bhandarkar, "LSTM based sentiment analysis of financial news," *SN Comput. Sci.*, vol. 4, no. 5, 2023, doi: https://doi.org/10.1007/s42979-023-02018-2.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, doi: https://doi.org/10.48550/ARXIV.1706.03762.

[14] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, doi = https://doi.org/10.48550/arxiv.1705.07874.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016, doi = https://doi.org/10.48550/arxiv.1602.04938.

Link to Presentation

`https://youtu.be/4NpsymyU5d8`

AI Declaration

In line with the University of Exeter's academic integrity policy, I declare the use of Generative AI (GenAI) as a supporting tool in this report. My use aligns with the AI-supported assessment guidelines for this module.

- **Tools Used:**
  - Gemini
  - Perplexity
- **Purpose of Use:**
  - Research Planning - Provide feedback on the scope of my research question
  - Draft Review - Review and give feedback on drafts including grammatical accuracy, punctuation, and suggesting improvements.

Specific prompts and outputs are listed within these links:

Gemini: `https://gemini.google.com/share/70a841eda8f0`

Perplexity: `https://www.perplexity.ai/search/don-t-delete-lfd-prompts-qicweTQCR2mG8N4xQW4i0A#3`

Appendix A - Extended Methods

Table 2: Summary of Methods and Model Architectures

Component	Description / Configuration
Preprocessing	• Lowercased all text • Removed URLs, stock tickers (e.g., $AAPL), special characters • Removed extra whitespace
Tokenisation and Embeddings	• Used ProsusAI/finbert [7] tokenizer from Hugging Face • Truncated input sequences to a maximum length of 64 tokens • Extracted 768-dimensional [CLS] embeddings from FinBERT
Model Architectures	• **Baseline:** TF-IDF + Logistic Regression • **MLP:** Three fully connected layers ($768 \rightarrow 256 \rightarrow 128$) with ReLU activations, dropout (0.3), and LayerNorm; output fed to a softmax classifier. • **LSTM:** Three-layer bidirectional LSTM (hidden size = 256) with dropout (0.3), followed by LayerNorm, an attention layer, and a final classification layer for 3-class output. • **Transformer:** Four-layer Transformer Encoder with 8 attention heads, Feedforward Dimension of 256, LayerNorm before/after encoding, followed by a two-layer MLP classifier. • All models use FinBERT embeddings as frozen input features.
Training Process	• All models trained on Google Colab's T4 GPU • Loss Function - CrossEntropy Loss with L2 Regularisation and Label Smoothing • Optimiser - AdamW with learning rate of 1×10^{-4} • Gradient clipping to prevent exploding gradients • Early stopping with a patience of 5 epochs, monitoring validation loss • Trained for a maximum of 100 epochs with batch size of 64 • FinBERT parameters frozen to reduce training time and memory usage

Appendix B - Extended Results

Table 3: Detailed Per-Class and Overall Performance Metrics Across All Models

Metric	Logistic Regression	MLP	LSTM	Transformer
Accuracy	0.854	0.882	0.897	0.903
Precision (Negative)	0.699	0.844	0.832	0.815
Recall (Negative)	0.772	0.775	0.827	0.846
F1 (Negative)	0.733	0.808	0.829	0.830
Precision (Neutral)	0.918	0.908	0.916	0.920
Recall (Neutral)	0.879	0.922	0.936	0.943
F1 (Neutral)	0.898	0.915	0.926	0.931
Precision (Positive)	0.790	0.836	0.882	0.905
Recall (Positive)	0.834	0.838	0.836	0.833
F1 (Positive)	0.812	0.837	0.858	0.867
Precision (Macro)	0.802	0.863	0.877	0.880
Recall (Macro)	0.828	0.845	0.866	0.874
F1 (Macro)	0.814	0.853	0.871	0.876
Precision (Weighted)	0.859	0.882	0.897	0.903
Recall (Weighted)	0.854	0.882	0.897	0.903
F1 (Weighted)	0.856	0.882	0.897	0.903
ROC-AUC (Negative)	0.952	0.972	0.972	0.974
ROC-AUC (Neutral)	0.935	0.954	0.961	0.965
ROC-AUC (Positive)	0.939	0.960	0.965	0.968
ROC-AUC (Macro)	0.942	0.962	0.966	0.969
Loss	–	0.542	0.500	0.493
Training Time (min)	0.057	32.578	32.602	35.226

Table 4: Prediction Examples By Model

Model	Type	True	Pred	Text Examples
LR	Correct	Negative	Negative	Finnish technology company Raute Corporation issued on Tuesday a profit warning for the financial year 2008 . | Tullow Oil diminishing reserves | Samsung currently occupies third place and lost ground during the quarter , dropping by 1.8 to an 11.1 share overall .
LR	Correct	Neutral	Neutral	Nike sets strategy in South America | The start of the negotiations , relating to Glaston 's efficiency program , was announced in October . | Aldata said that there are still a number of operational aspects to be defined between it and Microsoft and further details of the product and market initiatives resulting from this agreement will be available at a later date .
LR	Correct	Positive	Positive	Ramirent is a leading company in machinery and equipment rentals for construction and industry . | In January , traffic , measured in revenue passenger kilometres RPK , went up by 3.2 and capacity , measured in available seat kilometres ASK , rose by 12.2 . | The acquisition of Elektros Pavara completes KONE 's market expansion strategy in the Baltic Countries .
LR	Wrong	Negative	Neutral	BXC Form S3 BlueLinx Holdings Inc.
LR	Wrong	Negative	Positive	Profit before taxes was EUR 4.0 mn , down from EUR 4.9 mn . | According to Scanfil , demand for telecommunications network products has fluctuated significantly in the third quarter of 2006 , and the situation is expected to remain unstable for the rest of the year .
LR	Wrong	Neutral	Negative	2020 the year investors fall in love with stocks again? | Volatility still remains near alltime lows across some major currency pairs, including the euro versus the dollar,??
LR	Wrong	Neutral	Positive	The goal is significant expansion in Finland and in the northern Baltic region .
LR	Wrong	Positive	Negative	The growth of net sales in the first half of 2008 has been 28 compared with the first half of 2007 .
LR	Wrong	Positive	Neutral	The expansion includes the doubling of the floor space and the addition of more lifting capacity and production equipment . | Through this transaction we are able to participate in developing the industry .
LSTM	Correct	Negative	Negative	Finnish technology company Raute Corporation issued on Tuesday a profit warning for the financial year 2008 . | Tullow Oil diminishing reserves | Samsung currently occupies third place and lost ground during the quarter , dropping by 1.8 to an 11.1 share overall .
LSTM	Correct	Neutral	Neutral	Nike sets strategy in South America | The start of the negotiations , relating to Glaston 's efficiency program , was announced in October . | 2020 the year investors fall in love with stocks again?

Continued on next page

Table 4 continued from previous page

Model	Type	True	Pred	Text Examples
LSTM	Correct	Positive	Positive	In January , traffic , measured in revenue passenger kilometres RPK , went up by 3.2 and capacity , measured in available seat kilometres ASK , rose by 12.2 . | The acquisition of Elektros Pavara completes KONE 's market expansion strategy in the Baltic Countries . | Adjusted for changes in the Group structure , the Division 's net sales increased by 1.7 .
LSTM	Wrong	Negative	Neutral	BXC Form S3 BlueLinx Holdings Inc. | The largest hedge fund in the world has reportedly staked more than 1 billion that global equity markets will fall?? | The offer , deemed too low by Finnlines ' board , stands until 4 pm tomorrow .
LSTM	Wrong	Neutral	Negative	Volatility still remains near alltime lows across some major currency pairs, including the euro versus the dollar,??
LSTM	Wrong	Neutral	Positive	The goal is significant expansion in Finland and in the northern Baltic region . | SINT familiar name. Loading zone. Worth a trade again. There is a gap to fill.
LSTM	Wrong	Positive	Neutral	Ramirent is a leading company in machinery and equipment rentals for construction and industry . | ECONX December University of Michigan Consumer Sentiment Prelim 99.2 vs 96.5 consensus?? | BAM 3 Reasons Brookfield Is A MustOwn DividendGrowth Stock. Get more info on Seeking Alpha??
MLP	Correct	Negative	Negative	Finnish technology company Raute Corporation issued on Tuesday a profit warning for the financial year 2008 . | Tullow Oil diminishing reserves | Samsung currently occupies third place and lost ground during the quarter , dropping by 1.8 to an 11.1 share overall .
MLP	Correct	Neutral	Neutral	Nike sets strategy in South America | The start of the negotiations , relating to Glaston 's efficiency program , was announced in October . | 2020 the year investors fall in love with stocks again?
MLP	Correct	Positive	Positive	In January , traffic , measured in revenue passenger kilometres RPK , went up by 3.2 and capacity , measured in available seat kilometres ASK , rose by 12.2 . | The acquisition of Elektros Pavara completes KONE 's market expansion strategy in the Baltic Countries . | Adjusted for changes in the Group structure , the Division 's net sales increased by 1.7 .
MLP	Wrong	Negative	Neutral	BXC Form S3 BlueLinx Holdings Inc. | The largest hedge fund in the world has reportedly staked more than 1 billion that global equity markets will fall?? | Stocks face 50 odds of correction in 2020, Vanguard? Davis says

Continued on next page

Table 4 continued from previous page

Model	Type	True	Pred	Text Examples
MLP	Wrong	Neutral	Positive	The goal is significant expansion in Finland and in the northern Baltic region . | Solaris Outlines Broad Porphyry Deposit in First FollowUp Holes at Warintza West Discovery VANCOUVER, British Columbia, Oct. 13, 2022 Solaris Resources Inc. ??laris??or ??e Company?? is pleased to report assay results from the first series of holes following up on the discovery of Warintza West within its Warintza Project ??rintza??or ??e Project?? in southeastern Ecuador. Highlights are listed below, with a corresponding image in Figure 1 and detailed results in Tables 12. | Deluxe Reports Fourth Quarter 2019 Results and Record Full Year Revenue
MLP	Wrong	Positive	Neutral	Ramirent is a leading company in machinery and equipment rentals for construction and industry . | ECONX December University of Michigan Consumer Sentiment Prelim 99.2 vs 96.5 consensus?? | BAM 3 Reasons Brookfield Is A MustOwn DividendGrowth Stock. Get more info on Seeking Alpha??
Transformer	Correct	Negative	Negative	Finnish technology company Raute Corporation issued on Tuesday a profit warning for the financial year 2008 . | Samsung currently occupies third place and lost ground during the quarter , dropping by 1.8 to an 11.1 share overall . | OPEC slashed forecasts for global oil demand as the coronavirus hits fuel use in China
Transformer	Correct	Neutral	Neutral	Nike sets strategy in South America | The start of the negotiations , relating to Glaston 's efficiency program , was announced in October . | 2020 the year investors fall in love with stocks again?
Transformer	Correct	Positive	Positive	In January , traffic , measured in revenue passenger kilometres RPK , went up by 3.2 and capacity , measured in available seat kilometres ASK , rose by 12.2 . | The acquisition of Elektros Pavara completes KONE 's market expansion strategy in the Baltic Countries . | Adjusted for changes in the Group structure , the Division 's net sales increased by 1.7 .
Transformer	Wrong	Negative	Neutral	Tullow Oil diminishing reserves | BXC Form S3 BlueLinx Holdings Inc. | The offer , deemed too low by Finnlines ' board , stands until 4 pm tomorrow .
Transformer	Wrong	Neutral	Negative	Volatility still remains near alltime lows across some major currency pairs, including the euro versus the dollar,??
Transformer	Wrong	Neutral	Positive	The goal is significant expansion in Finland and in the northern Baltic region . | SINT familiar name. Loading zone. Worth a trade again. There is a gap to fill.
Transformer	Wrong	Positive	Neutral	Ramirent is a leading company in machinery and equipment rentals for construction and industry . | ECONX December University of Michigan Consumer Sentiment Prelim 99.2 vs 96.5 consensus?? | TENB Cowen sees upside in TENB's Indegy purchase

Appendix C - Training Dynamics

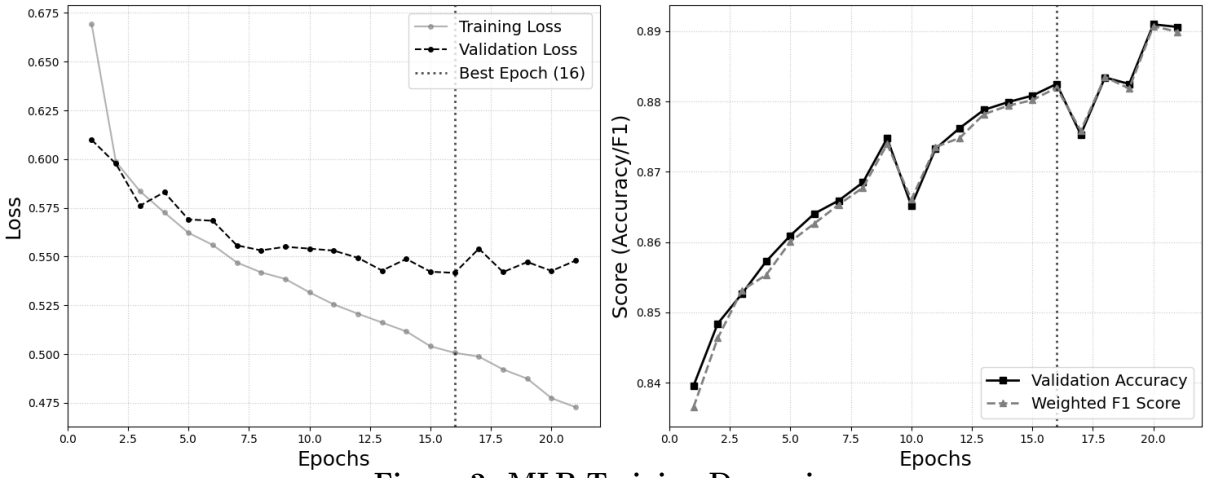

Figure 2: MLP Training Dynamics

Training and validation loss/accuracy curves for the MLP model over the training process. The model converged at epoch 16 with stable validation performance and minimal overfitting, achieving 88.2% accuracy.

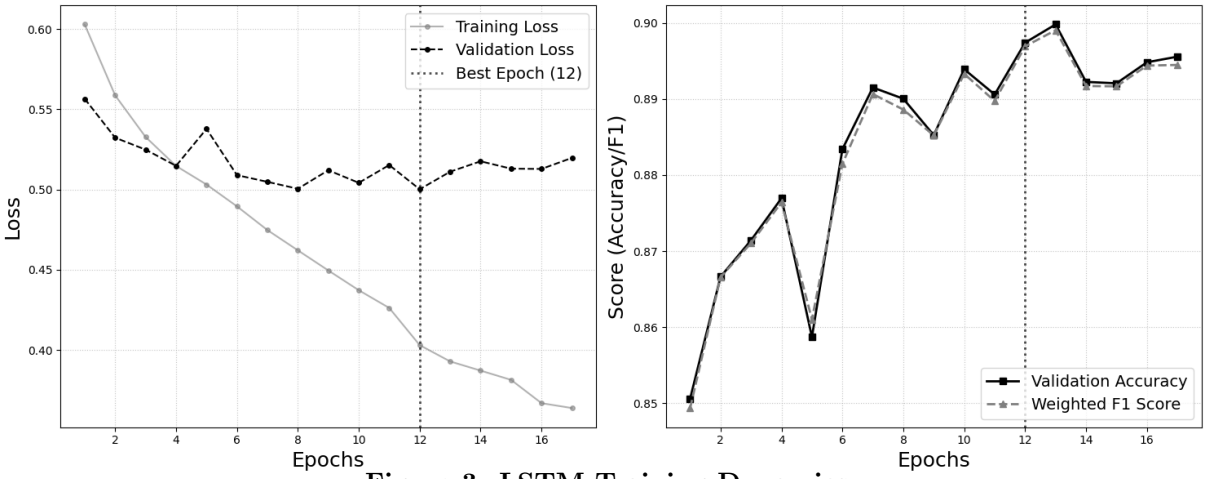

Figure 3: LSTM Training Dynamics

Training and validation loss/accuracy curves for the LSTM model over the training process. The model reached best performance at epoch 12 with 89.7% accuracy, showing a smooth convergence with slight validation changes.

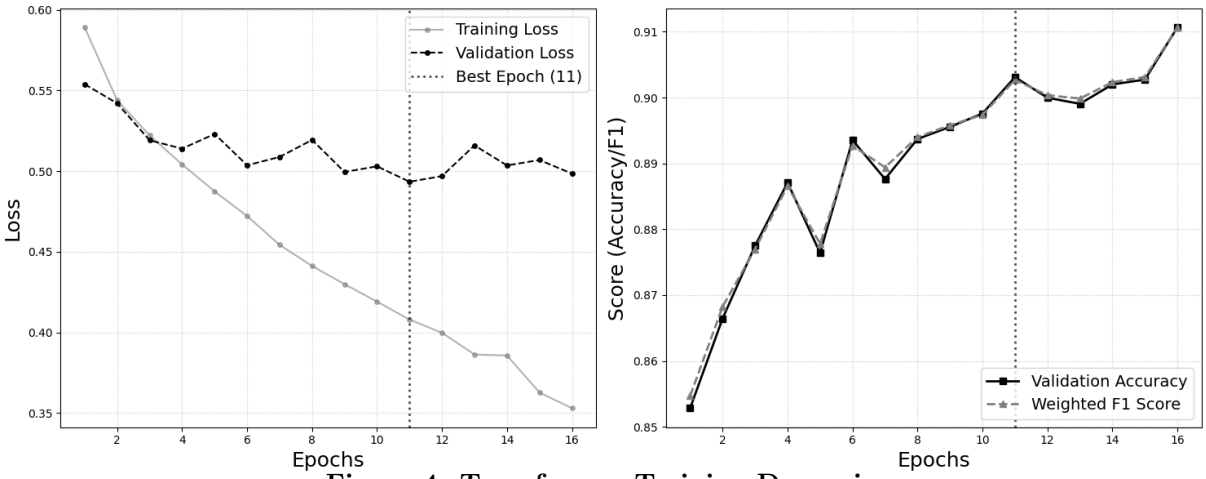

Figure 4: Transformer Training Dynamics

Training and validation loss/accuracy curves for the Transformer model over the training process. Best performance achieved at epoch 11 with 90.3% validation accuracy, demonstrating stable learning progression with minimal train-validation gap.