Incorporating Sentiment Analysis for Stock Prediction: An Integrated Framework for Embedding Evaluation and Hybrid Forecasting

Stuart Holland and Achyuth Kolluru

Abstract

This paper presents an integrated framework for stock prediction that leverages transformer-based sentiment analysis of financial tweets along with traditional quantitative market data. Our approach involves fine-tuning both a baseline BERT model (`bert-base-uncased`) and a domain-specific FinBERT model to extract rich semantic embeddings using multiple pooling strategies. These representations are then aggregated by day and company ticker and merged with 10 years of historical stock returns. We develop hybrid forecasting models that use dense neural networks and LSTM-based architectures to predict next-day returns. Detailed experimental analysis—including hyperparameter tuning, error stratification, and qualitative evaluation—is provided to understand performance variations across models and embedding types. Our work underscores the importance of embedding quality and robust evaluation metrics, and offers insights into which configurations best capture the nuances of investor sentiment.

1 Introduction

The stock market is influenced by both quantitative data and qualitative signals such as investor sentiment. Advances in Natural Language Processing (NLP) now allow us to extract nuanced signals from social media, especially financial tweets. In this study, we develop a unified system that first fine-tunes transformer-based models on financial tweet data to capture detailed sentiment embeddings and then leverages these embeddings in hybrid forecasting models to predict next-day stock returns. Our dataset comprises financial tweets (sourced from the Ackerman dataset on HuggingFace) focused on five major companies: Apple, Microsoft, AMD, Nvidia, and Amazon. After rigorous cleaning and sentiment label mapping, we apply both a general-purpose BERT model and FinBERT and extract embeddings via different pooling methods (CLS, mean, and max). These daily aggregated sentiment signals are combined with 10 years of Yahoo Finance market data for forecasting.

2 Background and Related Work

Recent research has demonstrated that augmenting stock prediction models with sentiment analysis can significantly improve forecasting accuracy. For instance, Ayyappa et al. (Ayyappa et al., 2023) combined LSTM networks with BERT-based sentiment extraction to capture temporal dependencies and nuanced market sentiment from tweets, achieving improvements over models relying solely on historical prices. Similarly, Chen et al. (Chen, 2019) leveraged contextualized embeddings from BERT to integrate financial news into stock movement prediction, while Zhou et al. (Zhou et al., 2020) provided a comprehensive review of deep learning approaches for financial sentiment analysis.

Other foundational works further support these methodologies. Mikolov et al. (Mikolov et al., 2013) established efficient training methods for word representations, which have influenced later embedding-based techniques. Loughran and McDonald (Loughran and McDonald, 2011) highlighted the importance of domain-specific textual analysis in financial documents. Additionally, Akkerman's work on FinTwitBERT-sentiment (Akkerman, 2023) provides a domain-tailored sentiment model that has informed subsequent studies, including ours.

In our work, we extend these approaches by focusing on a unique dataset that not only provides real-time sentiment labels (Bullish, Neutral, Bearish) but also includes technical analysis details. By concentrating our study on five prominent companies from the NYSE Fortune 100, we enable a detailed analysis of the interplay between investor sentiment and stock price movements.

3 Experimental Analysis and Model Comparison

Our experimental process is tightly interwoven throughout the project. We start by fine-tuning our transformer models on the preprocessed tweet dataset. Our fine-tuning configurations include:

- **Baseline:** 3 epochs, learning rate = 2×10^{-5}, batch size = 16.
- **Iteration Adjustments:** 5 epochs with increased learning rate (3×10^{-5}) or reduced batch size (8) to induce more frequent updates.
- **Regularization Enhancements:** Employing lower learning rates with warmup (e.g., 500 steps), increased dropout (0.3), and gradient clipping (max norm = 1.0).
- **Domain-Specific Training:** Fine-tuning FinBERT for 3 epochs.
- **Extended Fine-Tuning:** Further training the best configuration (with increased dropout) for 10 epochs.

Evaluation is conducted using standard metrics (accuracy, Mean Absolute Error) and is detailed in Table 1. We analyze model performance by stratifying results based on tweet characteristics (e.g., length, jargon density) and by company. This level of analysis ensures that our results speak not only to aggregate performance but also to the underlying strengths and weaknesses of each model.

4 Embedding Extraction and Daily Aggregation

After fine-tuning our transformer-based sentiment classifier, we extract tweet-level embeddings using both the standard `bert-base-uncased` model and FinBERT. For each tweet, we generate three types of embeddings by applying distinct pooling strategies on the transformer output:

- **CLS Embeddings:** The representation of the [CLS] token, which is conventionally used as a summary of the input sequence.
- **Mean-Pooled Embeddings:** The average of the token embeddings from the last hidden layer, capturing the overall semantic context across the entire tweet.
- **Max-Pooled Embeddings:** The maximum value taken across token embeddings, emphasizing the most prominent features.

Each tweet's embedding is tagged with its associated company ticker and the date (extracted from its timestamp). We then aggregate the embeddings on a daily basis per ticker. Specifically, for each day and each company, we compute the mean of all tweet embeddings to form a unified daily sentiment representation. This aggregation step is critical, as it reduces the variability of individual tweet predictions and provides a stable sentiment signal that aligns with the daily frequency of market data.

In parallel, we preprocess 10 years of Yahoo Finance market data by converting it into a long format that records the daily return for each company. The daily aggregated sentiment embeddings are then merged with the market data on the basis of date and ticker, forming a comprehensive dataset for our forecasting module.

Model	Train Loss	Validation Loss	Accuracy
Baseline (3 epochs)	0.87	0.84	63.1%
Iteration 2 (5 epochs, lr=3e-5)	0.89 / 0.21	0.85 / 0.99	63.3% / 71.5%
Iteration 3 (5 epochs, batch=8)	0.88 / 0.12	0.81 / 1.26	65.6% / 69.2%
Lower LR + Warmup	1.0/0.74	0.93/0.81	66.5%
Increased Dropout	0.94/0.34	0.88/0.86	67.7%
Gradient Clipping	0.94/0.34	0.88/0.86	67.7%
FinBERT	0.87	0.82	64.5%
Bert Extended Fine-Tuning	0.89/0.31	0.82/0.78	67.7%

Table 1: Evaluation Metrics for Models Tested

Visualization of Embeddings with t-SNE

To evaluate the quality and coherence of the extracted embeddings, we applied t-distributed

Stochastic Neighbor Embedding (t-SNE) to reduce the high-dimensional embedding space to two dimensions. The resulting t-SNE plots (see Figures 1 and 2) provide valuable insights into the semantic structure captured by our models.

Figure 1 displays the t-SNE visualization of the CLS embeddings. The clusters in this plot correspond to different sentiment classes (Bearish, Neutral, Bullish). There is a separation among clusters indicating that the model has learned to encode distinct sentiment features, but not well enough to accurately distinguish the data points in the center. However it supports the reliability of the aggregated daily sentiment.

Similarly, Figure 2 shows the t-SNE visualization for an alternative embedding extraction method (aggregated [CLS] embeddings from the last four hidden layers). This plot also validates that our alternative pooling strategy also produces semantically meaningful representations, with clusters that are consistent with the sentiment labels, however still struggles like the previous rendition.

Together, these t-SNE visualizations confirm that the extracted embeddings are robust and that the daily aggregation of these embeddings provides a stable, interpretable sentiment signal. This, in turn, justifies their integration with historical market data for enhanced stock prediction.

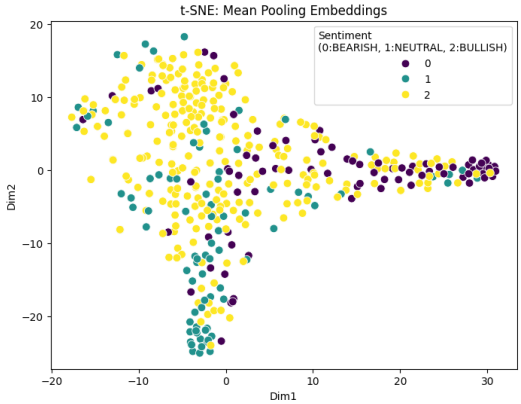

Figure 1: t-SNE mean pooling Embeddings

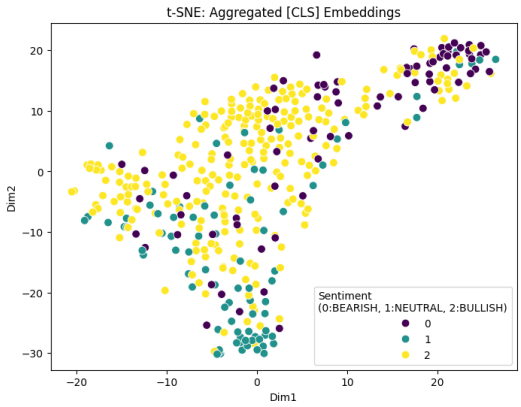

Figure 2: t-SNE Aggregated CLS Embeddings

5 Stock Return Prediction Using Integrated Sentiment and Price Data

Leveraging our unified dataset—where daily aggregated sentiment embeddings are merged with 10 years of historical market data—we develop hybrid forecasting models to predict next-day stock returns. Our approach integrates qualitative sentiment signals extracted from financial tweets with traditional quantitative features, enabling a greater representation of market dynamics.

We explored two neural network architectures:

- **Dense Neural Networks:** A fully-connected network applied on flattened embedding features. This model is designed to capture nonlinear relationships between sentiment embeddings and stock returns.
- **LSTM-based Recurrent Models:** A model that captures temporal dependencies by processing a sliding window of daily sentiment embeddings, which are combined with price features. This architecture is well-suited for time-series forecasting.

We experimented with different feature configurations, including:

- **Price-only Features:** Using only historical market data.
- **Sentiment-only Features:** Using daily aggregated sentiment embeddings.
- **Combined Features:** Integrating both price data and sentiment embeddings.

Our results indicate that integrating sentiment features with market returns consistently improves

predictive performance. For example, our experiments show that the LSTM model using mean-pooled FinBERT embeddings achieves the highest directional accuracy (approximately 56.8%), while dense networks based on CLS embeddings offer competitive Mean Absolute Error (MAE) with overall accuracy near 55.5%. Figures 3, 4, 5 and 6 illustrate these findings.

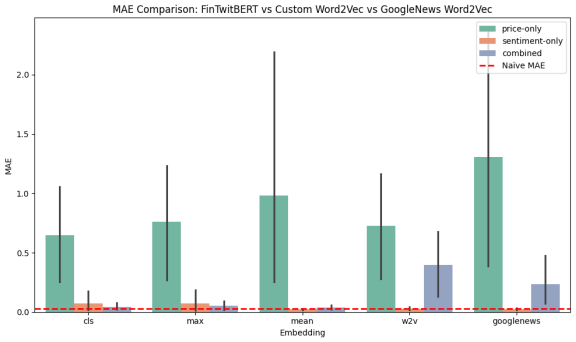

Figure 3: MAE Comparison Across Embedding Types and Models.

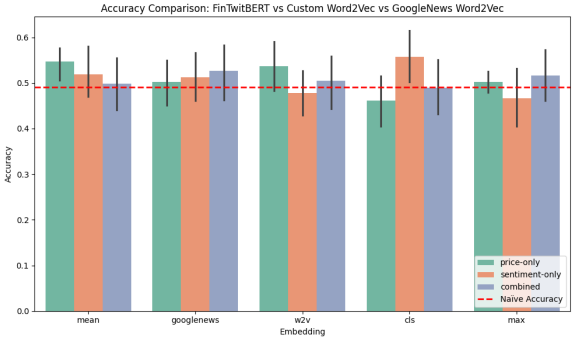

Figure 4: Accuracy Comparison Across Embedding Types and Models.

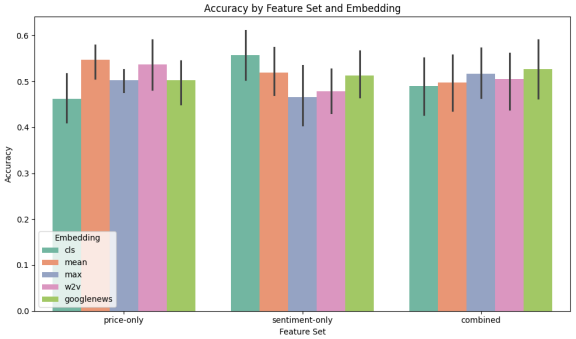

Figure 5: Grouped accuracy by feature configuration

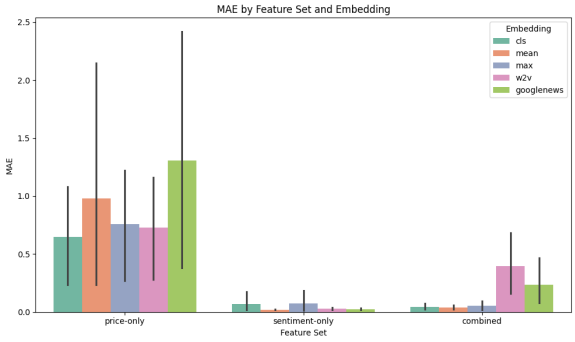

Figure 6: Grouped Mean Absolute Error (MAE) by feature configuration

We evaluate model performance across various embedding types (FinTwitBERT: CLS, Mean, Max; Word2Vec: custom and GoogleNews) and feature configurations (price-only, sentiment-only, and combined). Results are reported using both classification accuracy and Mean Absolute Error (MAE). Across figures, grouped comparisons reveal that price-only features consistently outperform individual inputs, while FinTwitBERT embeddings especially CLS and Mean tend to yield superior performance under the right circumstances. Red dashed lines in relevant plots denote naive baselines based on previous-day returns, providing a reference point. The visualizations collectively highlight the predictive strength of integrating sentiment and price signals, especially when encoded through domain specific embeddings.

6 Error Analysis and Qualitative Evaluation

A critical component of our study is the detailed error analysis and interpretation of results. We analyzed our prediction errors by stratifying the test set by tweet length, hypothesizing that longer tweets (which often contain more financial jargon and may suffer from truncation) present greater challenges.

Our analysis revealed:

- Longer, jargon-dense tweets tend to have higher misclassification rates.
- Certain companies exhibit unique error patterns that may be associated with industry-specific language or market volatility.
- Qualitative evaluations, supported by example tweets and corresponding predictions, provide valuable insights into model strengths and weaknesses.

7 Discussion and Future Work

Our integrated framework demonstrates that transformer-based sentiment analysis, when harmonized with quantitative market data, can enhance stock return predictions. In the process of writing this paper, it is important to critically analyze every choice made—from the evaluation metrics used (accuracy, MAE) to the method of aggregation and the selection of hyperparameters. For example, why do we use mean-pooled embeddings for certain models and CLS embeddings for others? Our analysis suggests that the choice of pooling affects the representation quality and thereby the forecasting accuracy; further breakdowns by company and tweet length support these hypotheses.

Additional aspects for future exploration include:

- Refining text preprocessing to minimize truncation effects and better preserve domain-specific terminology.
- Exploring additional segmentation strategies for longer tweets.
- Further dis-aggregating the results by categories (e.g., high-jargon vs. low-jargon) to provide deeper insights.
- Investigating the impact of combining transformer-derived embeddings with traditional technical indicators in a unified forecasting model.

8 Conclusion

In this study, we have presented a comprehensive, integrated framework that combines transformer-based sentiment analysis and quantitative market data for stock return prediction. Our work demonstrates that fine-tuning both a baseline BERT and a domain-specific FinBERT model significantly enhances the extraction of semantic embeddings from financial tweets. The integrated forecasting module shows that combining these qualitative features with historical stock data yields promising improvements in prediction accuracy. As we move forward, further refinement and expanded error analysis will help in forming robust, deployable hybrid forecasting systems that leverage both textual and numerical signals.

References

Stephan Akkerman. 2023. Fintwitbert-sentiment. https://huggingface.co/StephanAkkerman/FinTwitBERT-sentiment. Accessed: [insert date].

Y. Ayyappa, B. V. Kumar, S. P. Priya, S. Akhila, T. P. V. Reddy, and S. M. Goush. 2023. Forecasting equity prices using lstm and bert with sentiment analysis. In *2023 International Conference on Inventive Computation Technologies (ICICT)*, pages 643–648.

Q. Chen. 2019. Stock movement prediction with financial news using contextualized embedding from bert. *arXiv preprint arXiv:2107.08721*.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66(1):35–65.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

X. Zhou, X. Song, and Q. Li. 2020. Deep learning for financial sentiment analysis: A survey. *Expert Systems with Applications*, 154:113468.