

Sentiment Analysis on Financial News Project Report

Zhenhan Chen

1. Introduction

Financial markets react to the continuous flow of information, and news articles are one of the main channels through which investors learn about companies. The idea of using natural language processing to systematically extract sentiment from news text and leverage it for predicting stock returns has gained significant attention in recent years, especially with the development of transformer-based models like BERT.

However, the Efficient Market Hypothesis suggests that publicly available information, including news, should already be reflected in stock prices. This makes predicting future returns from news text a fundamentally difficult task. Despite this challenge, there is some empirical evidence that sentiment extracted from financial text can carry predictive value, particularly when aggregated over time or across many articles.

This project applies NLP techniques to analyze sentiment in Wall Street Journal articles covering S&P 500 companies. Two approaches are compared: (1) FinBERT (Huang et al., 2022), a pretrained model specifically designed for financial sentiment classification, and (2) a fine-tuned neural network trained on BERT embeddings to predict daily stock returns. Both models are evaluated on individual return prediction, binary price direction classification, and whether aggregated sentiment can forecast broader market trends.

2. Data Description

2.1 Datasets

The news dataset (news.csv) contains 20,550 articles from the Wall Street Journal about S&P 500 firms, spanning from January 3, 2017 to December 30, 2020. Each record includes the publication date, headline, full article body, and the stock ticker being discussed. The dataset covers 381 unique tickers, though coverage is not evenly distributed. Some heavily-covered companies like AAPL and AMZN have significantly more articles than others.

The price dataset (price.csv) has 603,840 rows of daily closing prices for 480 S&P 500 constituent stocks and the S&P 500 index (ticker SPX). It covers January 2017 to December 2021, providing an extra year of price data beyond the news coverage which is helpful for computing forward returns for articles published near the end of 2020.

Figure 1 shows the monthly distribution of articles in the dataset. There is noticeable variation in article volume across months, with some periods having significantly more coverage than others.

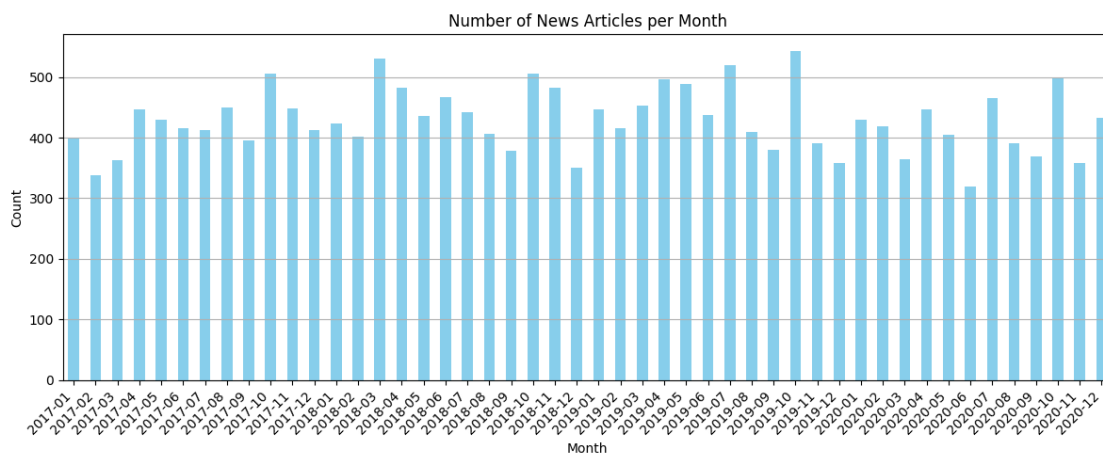


Figure 1: Number of news articles per month

2.2 Preprocessing

Daily stock returns are computed from closing prices using percentage change, grouped by ticker so that returns are calculated within each stock's series independently. Both date columns are converted to datetime format for proper alignment between the two datasets.

For articles published on non-trading days (weekends and holidays), each article is matched to the return on the next available trading day for that ticker. This reflects the assumption that the market's reaction to the news would show up on the next open session. The article body text is tokenized with a maximum length of 512 tokens since that is BERT's limit. Longer articles get truncated, which means some information from lengthy pieces may be lost, though the most important content usually appears early in financial articles.

2.3 Train-Validation-Test Split

A time-series split is used to avoid look-ahead bias, which is critical in financial applications. The data is

split chronologically: the training set includes articles before January 2019 (10,320 articles), the validation set covers 2019 (5,338 articles), and the test set includes 2020 onwards (4,892 articles). This ensures the model never sees future data during training. It is worth noting that the test period coincides with COVID-19, which introduced unusual volatility. Figure 2 visualizes the split.

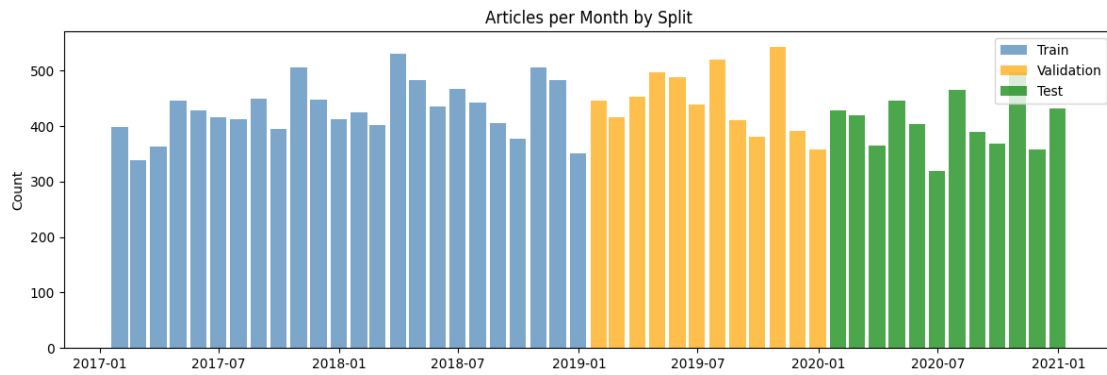


Figure 2: Articles per month by train/validation/test split

3. Methodology

3.1 Pretrained Model: FinBERT

FinBERT is a BERT model that was further fine-tuned on a large corpus of financial text for 3-class sentiment classification: positive, negative, and neutral (Huang et al., 2022). The yiyanghkust/finbert-tone checkpoint from HuggingFace is used. Each article body is tokenized with `max_length=512` and passed through the model, which outputs probability scores for the three sentiment classes that sum to 1.

To derive a single sentiment score, the following formula is applied: $\text{sentiment} = P(\text{positive}) - P(\text{negative})$. This gives a value from -1 (strongly negative) to +1 (strongly positive), with values near zero indicating neutral or mixed sentiment. Inference is performed in batches of 16 on GPU for efficiency. Figure 3 shows the distribution of the three probability outputs across all articles.

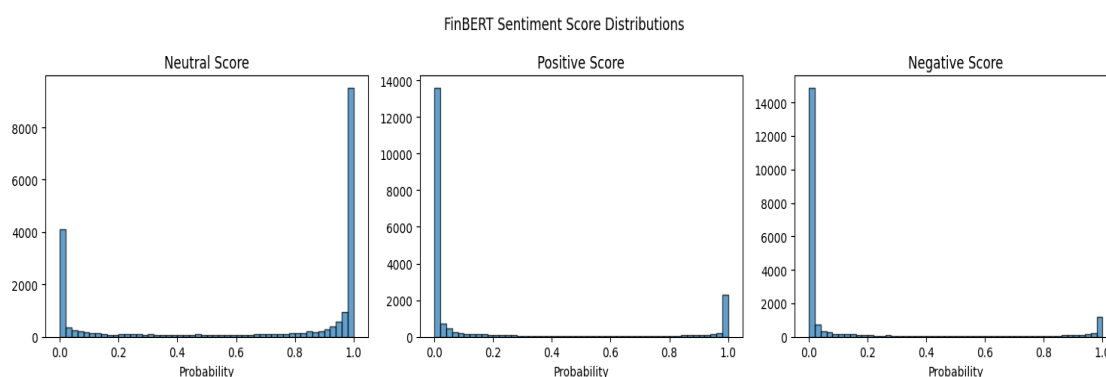


Figure 3: FinBERT sentiment probability distributions

The distributions show that most articles receive high neutral probabilities, which makes sense for Wall Street Journal reporting. The positive and negative distributions are right-skewed with most values near zero, meaning strong sentiment signals are relatively rare.

3.2 Fine-tuned Neural Network

The fine-tuned approach uses a two-stage pipeline. In the first stage, `bert-base-uncased` is used as a frozen feature extractor. Each article is tokenized and fed through BERT, and the 768-dimensional [CLS] token embedding from the last hidden layer is taken as the document representation. The [CLS] token is designed to capture the overall meaning of the input sequence. All embeddings are pre-computed and cached as `.numpy` files to avoid redundant computation.

In the second stage, these embeddings are fed into a fully connected neural network (ReturnPredictor) that predicts the next trading day's return. The architecture is: `Linear(768 -> hidden_size) -> ReLU -> Dropout -> Linear(hidden_size -> hidden_size/2) -> ReLU -> Dropout -> Linear(-> 1)`. The model is trained with MSE loss using Adam optimizer. The target variable is the daily return on the next trading day after the article's publication.

This approach differs from FinBERT in an important way: instead of classifying sentiment directly, it tries to learn a mapping from text representations to actual stock returns. The sentiment is implicitly captured in the predicted return value. A positive predicted return implies positive sentiment, and vice versa.

3.3 Hyperparameter Experimentation

To understand how sensitive the model is to architectural choices, four configurations are tested varying hidden layer size, learning rate, dropout rate, and number of training epochs. Table 1 summarizes the configurations and results.

Configuration	Hidden	LR	Dropout	Epochs	Best Val Loss
Base	256	0.001	0.2	10	0.000604
Wide	512	0.0005	0.3	15	0.000604
Narrow	128	0.002	0.1	10	0.000604
Slow	256	0.0001	0.2	20	0.000606

Table 1: Hyperparameter experiment results.

All four configurations converged to nearly identical validation losses around 0.000604, with only the Slow configuration slightly worse at 0.000606. This insensitivity to hyperparameters suggests that the bottleneck is not the network architecture but rather the inherent difficulty of the prediction task. The frozen BERT embeddings may simply not contain enough information to meaningfully predict daily returns, regardless of how the downstream network is configured.

The Narrow configuration (hidden_size=128, lr=0.002, dropout=0.1) is selected as the best model. Figure 4 shows the training and validation loss curves for all four configurations.

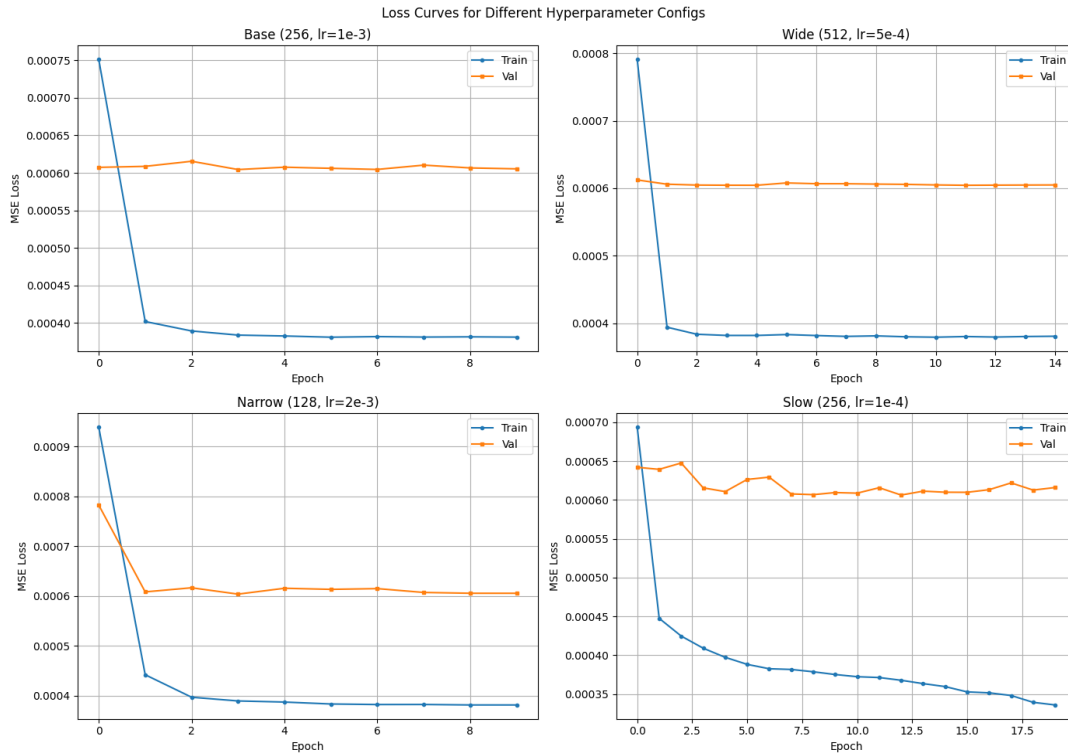


Figure 4: Loss curves for all hyperparameter configurations

4. Results

4.1 Fine-tuned Model: Regression Performance

Figure 5 shows the baseline model's training and validation loss over 10 epochs. Both curves drop sharply in the first 2-3 epochs and then flatten out, indicating that the model quickly converges to its best solution. The gap between training and validation loss is small, suggesting the model is not overfitting. Rather, it has simply learned to predict values close to zero, which is the mean of daily returns.



Figure 5: Baseline model training and validation loss

The test set evaluation metrics for the best model (Narrow configuration) are as follows:

MSE: 0.00137 | R^2 (in-sample): -0.0003 | R^2 (OOS, benchmark=0): 0.0005

The MSE of 0.00137 seems low in absolute terms, but this is simply because daily returns are small numbers. The in-sample R^2 is slightly negative, meaning the model performs worse than just predicting the mean of the test set. The out-of-sample R^2 using zero as the benchmark is 0.0005, which is essentially zero. These results are consistent with what most academic studies find when trying to predict daily returns from text data. Stock returns at the daily level are dominated by noise, and the signal from any single news article is extremely weak.

The actual vs predicted scatter plot (Figure 6) makes this very clear. The model's predictions are tightly clustered in a narrow band around zero, while the actual returns span a much wider range from roughly -15% to +15%. The model has essentially learned to output a near-constant prediction.

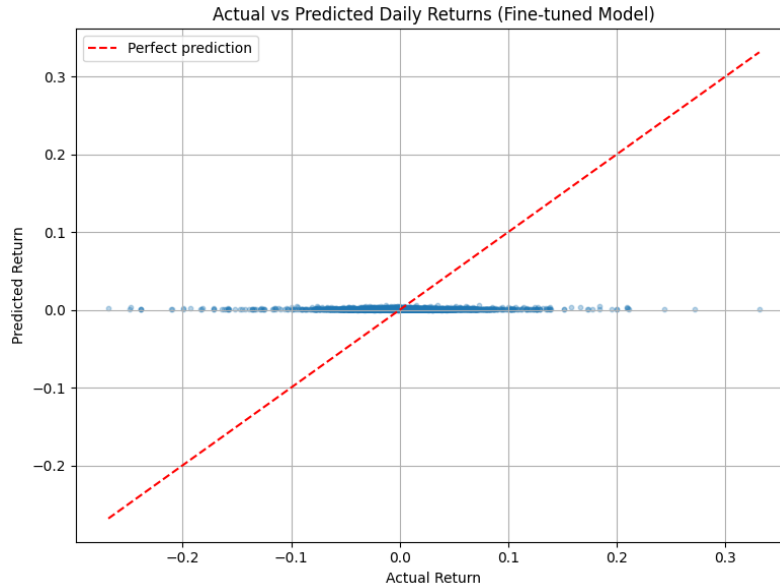


Figure 6: Actual vs predicted daily returns (fine-tuned model)

4.2 Binary Classification: Price Movement Prediction

Since continuous return prediction proved difficult, both models are also evaluated on a simpler task: predicting the direction of price movement (up or down). Predictions and actual returns are converted to binary labels where a positive return counts as 1 (increase) and negative as 0 (decrease). Table 2 shows the results.

Model	Accuracy	F1	Precision	Recall
FinBERT	49.3%	0.49	0.49	0.49
Fine-tuned NN	50.0%	0.67	0.50	1.00

Table 2: Binary classification performance comparison.

FinBERT achieves 49.3% accuracy with balanced precision and recall around 0.49. This is essentially random guessing, meaning the FinBERT sentiment direction has no predictive power for next-day price movements at the individual article level.

The fine-tuned model shows 50.0% accuracy with a seemingly better F1 of 0.67. However, this is misleading. The recall of 1.00 and precision of 0.50 reveal that the model predicts virtually every sample as "increase." Since roughly half of daily returns are positive, this trivial strategy achieves about 50% accuracy. The high F1 is just an artifact of the class imbalance in the metric calculation, not actual predictive skill.

Neither model can reliably predict daily price direction from news, which is consistent with the efficient market hypothesis. The information in news articles is likely already priced in by the time the next trading day opens.

4.3 Sentiment and Market Trend

While individual article-level prediction failed, sentiment might still be useful when aggregated. Average

monthly sentiment scores are computed for each model and compared with the S&P 500 return in the following month (month t sentiment vs month $t+1$ return). This lead-lag setup tests whether current sentiment can forecast next month's market direction. OLS regressions are fitted and correlations computed. Table 3 shows the results.

Model	OLS Slope	p-value	R ²	Correlation
FinBERT	-0.308	0.314	0.101	-0.318
Fine-tuned NN	42.765	0.947	0.000	0.021

Table 3: Month t sentiment vs month $t+1$ S&P 500 return.

For FinBERT, the OLS regression reveals a negative slope of -0.308 with an R^2 of 0.101 and correlation of -0.318. The p-value of 0.314 means this is not statistically significant, but the direction is noteworthy. The negative relationship suggests that months with more positive sentiment tend to be followed by lower market returns. This aligns with contrarian investment theory, where periods of excessive optimism can precede corrections. Figure 7 shows the scatter plot with OLS regression line and the time series comparison for FinBERT.

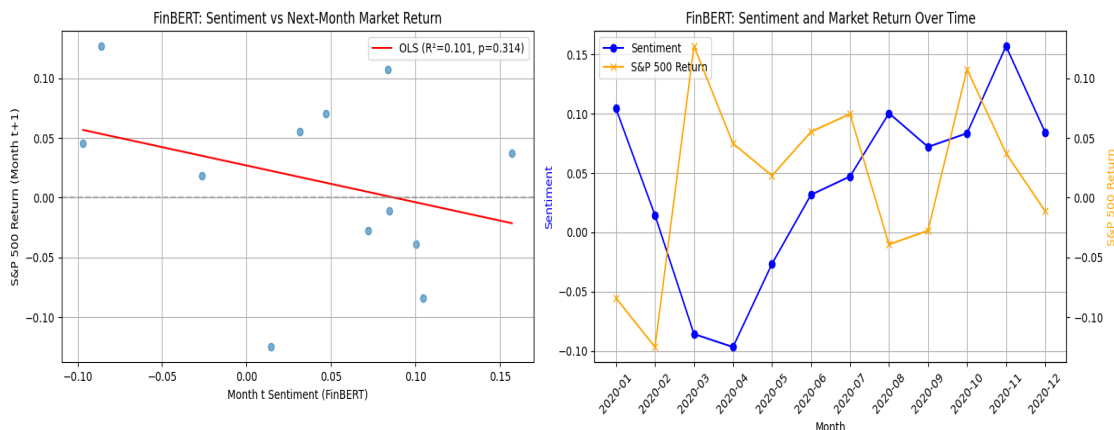


Figure 7: FinBERT - month t sentiment vs month $t+1$ S&P 500 return

The fine-tuned model shows essentially zero relationship with future returns ($p=0.947$, $R^2 \approx 0$, correlation=0.021). This is because the model's predictions are so tightly clustered around zero that monthly aggregation produces very little variation. Without meaningful variation in the sentiment signal, there is nothing to correlate with market returns. Figure 8 shows this clearly, with the sentiment values concentrated in an extremely narrow range.

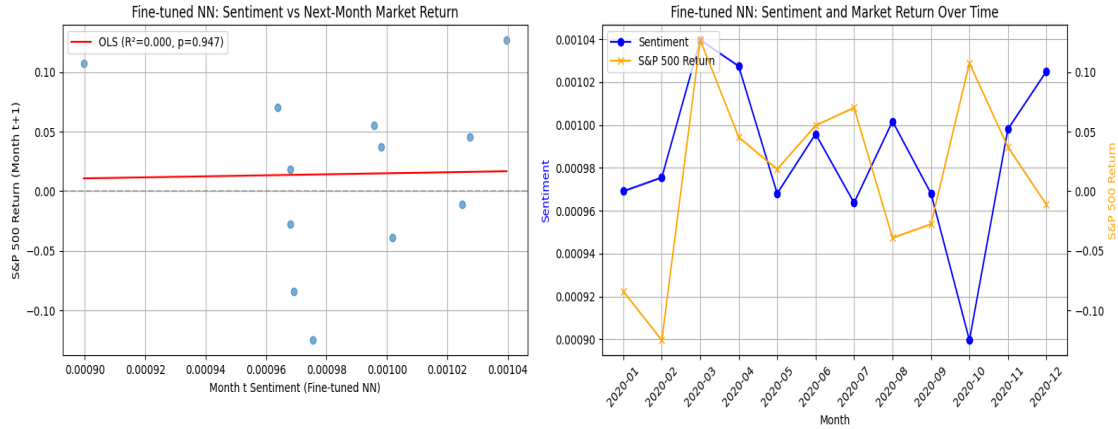


Figure 8: Fine-tuned NN - month t sentiment vs month $t+1$ return

Figure 9 directly compares the monthly sentiment trajectories of both models. FinBERT sentiment shows clear month-to-month variation, capturing shifts in the overall tone of financial news coverage. The fine-tuned model's sentiment is essentially flat, confirming that it lacks the discriminative power to differentiate between positive and negative news environments at any scale.

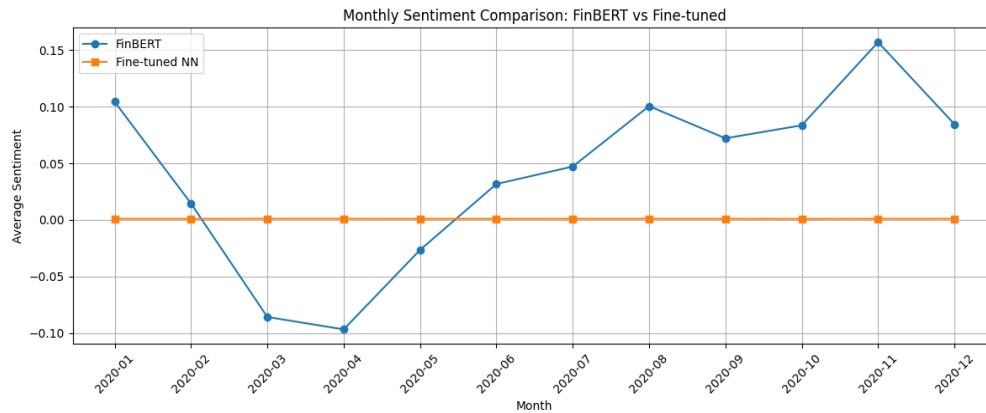


Figure 9: Monthly sentiment comparison - FinBERT vs Fine-tuned NN

Overall, FinBERT is clearly the better model for capturing meaningful sentiment signals. Its advantage comes from being explicitly trained on financial text for sentiment classification, giving it the ability to distinguish positive from negative news. The fine-tuned model, trained on the noisy task of predicting exact daily returns, collapses to predicting the mean and loses all sentiment information in the process. However, the small sample size of roughly 12 months in the test period makes it difficult to draw strong conclusions about the statistical significance of the sentiment-market relationship.

5. Conclusion

This project compared two NLP approaches for financial news sentiment analysis. FinBERT, a pretrained financial sentiment model, was used to directly classify article sentiment. A fine-tuned neural network was trained on BERT embeddings to predict daily stock returns as an implicit sentiment measure. Both are evaluated on regression accuracy, binary price prediction, and sentiment-market trend analysis.

The main findings are: (1) neither model can meaningfully predict daily stock returns or price direction at the individual article level, which is consistent with market efficiency; (2) FinBERT produces more useful sentiment signals than the fine-tuned model, which collapses to near-zero predictions; (3) there is a weak negative correlation between FinBERT sentiment and next-month S&P 500 returns, though it is not statistically significant given the limited test period.

The project has several limitations. BERT's 512-token limit truncates longer articles. Some articles mention multiple companies but only one ticker is used. The test period is limited to 2020, which was dominated by COVID-19 market dynamics. And with only about 12 monthly observations, the trend analysis has low statistical power.

Future work could explore longer-context models like Longformer to handle full articles. Predicting weekly or monthly returns instead of daily ones might reduce noise. End-to-end fine-tuning of BERT rather than using frozen embeddings could help the model learn better representations for the prediction task. Finally, a longer evaluation period spanning multiple years would provide more reliable evidence about the sentiment-market relationship.

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 4171-4186.

Huang, A. H., Wang, H., & Yang, Y. (2022). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*, 39(4), 2314-2340.