Events, Sentiment, and Stocks:
What We Learned from Trying (and Failing) to Predict the Stock Market

Rebecca Baugh, Chris DeMaio, and Aaron Webb
University of California Berkeley
DATASCI 266
Fall 2024

Abstract

Given the success of research that demonstrates the ability of large language models to utilize qualitative data to improve predictions, we sought to discover if an understanding of key events relevant to a stock and the associated company could help to improve these predictions further. We used ARIMA as our quantitative and we use a LSTM with FinBERT sentiment classifications of news articles as our qualitative baselines. We then fine-tuned another FinBERT model on corporate event detection data using LoRA and non-LoRA methods. Contrary to our hypothesis, we found that the addition of event data did not change the sentiment results of the FinBERT baseline. We also had issues generalizing our LSTM across a rolling window. Though we were unable to use business event data to improve stock predictions we believe there are opportunities to use newer models to improve stock predictions by incorporating addition qualitative measures beyond sentiment.

1 Introduction

Attempting to predict the stock market is not a new problem. However, recent advances in large language models create new opportunities to extract and incorporate qualitative information to improve these predictions. We are building off of existing literature that has been able to outperform traditional methods like ARIMA (Box and Jenkins, 1976). These studies use large language models to classify sentiment of news articles for use in stock predictions (Kirtac and Germano, 2024) and (Jiang and Zeng, 2023). Our research aims to go beyond sentiment, focusing on understanding key corporate events within news articles and using them to supplement sentiment analysis and improve short term stock prediction.

Table 1 includes an overview of how each data source (quantitative stock information, news, and event detection) is used within the different models.

2 Background

One of the issues with using sentiment analysis of news articles to predict stocks is the quality of the articles. We use the FNSPID[1] dataset, which works well with transformer-based models. There are other models which attempt to de-noise news data. However, that is not the focus of our research and the filtering pipeline that was used to create FNSPID is publically available (Dong et al., 2024).

We wanted a data set that works well with transformers because multiple experiments show that transformer-based approaches outperform traditional stock prediction methods in the short term (Jiang and Zeng, 2023; Jain and Agrawal, 2024; Kirtac and Germano, 2024).

(Kirtac and Germano, 2024) use a GPT-3-based OPT based model to predict stock market returns with 74.4% accuracy, using returns over a three day period as a success metric. BERT was a close second at 72.5%.

(Jain and Agrawal, 2024) uses the Alpha Vantage API [2] to get news articles, create a trustworthiness score, and use FinBERT to classify the sentiment of the headline and the article summary. This resulting model, FB-GAN, outperforms the best performing stock price prediction that uses only historical price data, WGAN-GP. The performance of these models is evaluated using RMSE for individual stocks.

(Jiang and Zeng, 2023) provides further evidence that FinBERT and BERT models can outperform traditional methods in predicting the next day closing price. This study also surfaces interesting limitations with long term prediction and over-fitting to noise. However, it is consistent with the other studies in the use of BERT or FinBERT (Yang and Huang, 2020) for the highest performing short term stock prediction models using news data.

[1] Available at `https://hugginface.co/datasets/Zihan1004/FNSPID`

[2] Available at https://www.alphavantage.co/

Model	Stock price	FinBERT Sentiment	FinBERT Event
ARIMA	Alpha Vantage	none	none
Sentiment only	Alpha Vantage	FNSPID	none
Sentiment & Events	Alpha Vantage	FNSPID	EDT
Sentiment & Events LoRA	Alpha Vantage	FNSPID	EDT

Table 1: Data source for each model iteration. FNSPID curated financial news dataset (Dong et al., 2024). EDT is a corporate event detection dataset (Zhihan, 2021), and the stock price data is from the Alpha Vantage API (Alpha Vantage Inc., 2024)

3 Methods

Instead of focusing on the news filter side of research with GAN, we are using the existing high-quality FNSPID[3] dataset because we are focused on seeing if we can expand the qualitative measure beyond sentiment to predict short-term stock movement from news data. While de-noising news articles to extract high-quality results is an interesting part of the stock prediction pipeline, it is not our focus.

Baselines and Experiments

1. ARIMA model as our quantitative baseline
2. FinBERT sentiment classifier as our qualitative baseline
3. FinBERT fine tuned on corporate event detection (LoRA and non LoRA approaches) as our experiment component

Using corporate event detection (EDT[4]) data, we want to combine the sentiment classifier and build an event classification layer to see if we can improve the model. We believe key business events can be as indicative towards stock price movement in the short term as the overall sentiment of the news article.

3.1 Design

3.1.1 Data Sources

Table 1 provides an overview of the data sources for stock price and news articles for sentiment analysis. It also outlines which of our models depend on those datasets to create features for prediction. We pulled the stock data from the AlphaVantage API. We choose SP100 stocks as of December 2nd, 2024, assuming that larger market cap stocks are more closely associated with the news cycle. The FNSPID dataset provided us with news articles for these stocks. Our training sets consisted of news articles about SP100 companies from 2018 until 2022 and the OHLCV (Daily Open, Daily High, Daily Low, Daily Close, Volume) data for the same companies and time range. Our testing set includes SP100 OHLCV and news articles throughout the whole 2023 calendar year. The news article sentiment and business events were classified by FinBERT and further technical indicators were engineered to be included in the LSTM training and testing sets.

3.1.2 Success Measures

We are using accuracy to measure the success of our models. The focus of our experiment is to see if sentiment supplemented by business events can improve a model in short-term predictions. While there are more complex means of tracking overall return on investment as described in the background, we want our first pass to be simple and predict if the stock price goes up or down at next day open.

3.2 Implementation

3.2.1 ARIMA

Though sentiment classifiers have outperformed ARIMA models for years, we decided to include a quantitative only ARIMA model as our baseline. Because this model is univariate, the features and target are both the Daily Open price. We leveraged the auto arima function from pmdarima to fit 101 separate ARIMA models for each stock in the SP 100. The parameters of the ARIMA grid search are outlined in Table 2. Because fitting 101 models is compute intensive, we've set the max p and max q to 7 days, and fit the models in batches of 10 stocks.

[3] Available at `https://hugginface.co/datasets/Zihan1004/FNSPID`

[4] Available at `https://github.com/Zhihan1996/TradeTheEvent/tree/main/data`

seasonal = False
max p = 7
max d = 1
max q = 7
stepwise = True

Table 2: ARIMA Grid Search Parameter

The models were initially trained on 2018 to 2022, then a 60-day rolling window updated the model parameters and made a forward prediction starting in 2023 and slide forward until the end of 2023. The model predicted the next-day stock price, then we converted that into a 1 for upward movement and a 0 for no movement/downward movement to evaluate accuracy.

3.2.2 LSTM

Since most of the modern research on stock predictions use an LSTM architecture to make the forward prediction, we have followed that same approach. Hyper parameter tuning was applied to the sentiment-only dataset and the same network architecture was used to evaluate if the impact of Business Events identified from the fine-tuned FinBERT model and LoRA-adapted FinBERT model would improve accuracy on top of sentiment alone.

The process of training the LSTM follows logic similar to the ARIMA baseline. We initially trained the LSTM on 2018 to 2022, then applied a 60-day rolling window process to make the predictions into 2023. The LSTM model weights were updated during the 60-day rolling windows, similar to the ARIMA model.

Convergence of this network required a lot of adjustments. After numerous initial training rounds, the training and validation loss seemed to be caught in a local minima with no loss reduction until further feature engineering was applied to the stock data and the hyper parameters were adjusted significantly.

The stock data itself started off with the OHLCV features, average news sentiment, and and indicator if there was news on that day for that stock or not. We further added technical indicators such as the RSI (Relative Strength Index), MACD (Moving Average Convergence/Divergence) Bollinger-Width to provide additional information around a stock's momentum and variance. Lagged features such as Lag 1 and Lag 2 of Daily Open prices and Lag 1 of Volume were also included. Finally, two temporal features were added as well: day of the week and month of the year.

3.2.3 Sentiment Only

Based on the success of the FinBERT (Yang and Huang, 2020) model and our high-quality news articles, we used a standard FinBERT sentiment classifier to supplement our stock data with a sentiment score of "positive", "negative", or "neutral". The label with the highest probability was chosen and encoded as a -1 (negative) to 1 (positive). We then grouped all the articles for a given stock on a given day and averaged them to be the average sentiment feature used in the LSTM. For those days where no news was recorded in our data, we included a 'news day' indicator variable to denote that. Because we had multi-day news gaps, we used an exponential moving average to dilute the news sentiment from one given day over the next 5 days when news did not exist. The hope is that the network will learn, in a generalizable fashion, the association between a company's daily stock indicators and it's news sentiment to better predict whether the Open price will go up or down the next day.

3.2.4 Sentiment and Events (LoRA and Non-LoRA)

The sentiment and event models required three steps. First, we used FinBERT to classify the FNSPID news articles and create a "positive", "negative", or "neutral" sentiment. Then, we used the EDT data to fine-tune a FinBERT model to be label to generate one of twelve classes of business events shown in Table 4. This combined dataset was then fed into the same LSTM as the sentiment only model.

Key	Event
I-A	Aquisition
I-CT	Clinical Trial
I-RD	Regular Dividend
I-DC	Dividend Cut
I-DI	Dividend Increase
I-GI	Guidance Increase
I-NC	New Contract
I-RSS	Reverse Stock Split
I-SD	Special Dividend
I-SR	Stock Repurchase
I-SS	Stock Split(SS)
0	No Event

Table 4: Business event labels (Zhihan, 2021)

Hyper Parameter	Values Tested	Final Selection
LSTM layers	[1, 2, 3]	2
Hidden Size	[64, 96, 128, 256]	96
Dropout	[0.1, 0.2, 0.3]	0.2
Learning Rate	[0.001, 0.0005, 0.0001]	0.0005
Learning Weight Decay	[0, 0.0001]	0.0001
Learning Rate Scheduler	[0, 1]	1
Number of Epochs	[20, 50, 100]	100

Table 3: Hyper Parameters Evaluated

This set of labeled data focuses more on events that happen to a company's stock rather than events that happen to a company itself. The data was originally labeled at the word level rather than the sentence level. But, on inspection of the metadata we understood that we needed to combine the data into sentences using a sentence-level rule. We chose to take the most frequently recurring label that was not "No Event" in order to stay most in line with the metadata instructions. We explored the possibility of a multi-label classification, but decided to use a max-recurring approach for simplicity.

When broken down at the sentence level, the training event detection data has 9,718 rows. 7,557 of these rows were labeled as "No Event." This means a majority class represented 78% of the data. To offset the imbalance, we implemented stratification in the train and evaluation splits, as well as a weighted sampler applied to each batch.

Even with balancing techniques, the size of the data set did not feel sufficient for fine-tuning a FinBERT (Yang and Huang, 2020) model. Additionally, the data imbalance was so large that we opted to try other approaches, such as LoRA (Hu et al., 2021).

However, the LoRA model was highly sensitive to the imbalance of the dataset. We decided to undersample because LoRA does not require large amounts of data to train. We ended up with 50 examples per label. We trained our data on a perfectly balanced 440 rows evaluated on 110 rows. The cross-entropy loss for the validation set did not improve beyond a near-random guess, which is reflective in our results.

4 Results

Looking at Table 5, the ARIMA baseline performs as expected with other published models. The Sentiment only model, achieves comparable results in pre-training, but suffers a steep drop in the rolling window, likely due to the temporal factors of our dataset.

Comparing our two experimental models to our baselines, we are unable to demonstrate any meaningful increase in predictive power.

4.1 Economic Shifts at the Train / Test Split

Looking first at our inability to generalize from pretraining to the rolling windows in the LSTMs, we see a downward economic trend from 2018 to the end of 2022. This cycle begins to reverse right at the start of our 2023 test data. Given the train data time range, it would be difficult to predict upward movement.

4.2 Low Quality Fine Tuning For Event Detection

Ultimately, our fine-tuned models for event classification failed to learn much. The cross entropy loss for both barely improved beyond a random guess, and therefore it is unsurprising that we did not see benefits in stock prediction with event detection. Going forward, we would likely try to find more balanced datasets with more relevant and diverse labels to see if the issue was in the type of data we used or in our approach.

Also, we need to further examine how the labels available in the EDT dataset were over-focused on things that can happen to stocks. Since we have a lot of quantitative information about the stocks themselves, it may be better to find event labels that supplement stock events, rather than repeat them. The value of the business event feature was likely limited because it was redundant.

4.3 Additional Quantitative Details in LSTM

In order to get out of the local minimum, features were calculated from quantitative stock data. This demonstrated the strength of quantitative features in stock prediction, including temporal features. (Alpha Vantage Inc., 2024). The improvement

	Model	Features	Accuracy
1	ARIMA	Quantitative only	49% (100 stocks)
2	LSTM	Quantitative and FinBERT Sentiment	70%: pre-training 52%: rolling-window
3	LSTM	Quantitative and FinBERT Sentiment & Events	70%: pre-training 50%: rolling-window
4	LSTM	Quantitative and FinBERT Sentiment & Events (LoRA)	70%: pre-training 53%: rolling-window

Table 5: Results: The ARIMA model is stand alone, the remaining models were fed through the LSTM, while sentiment and event features were added using the referenced models as classifiers

over the ARIMA baseline demonstrates the power LSTM models supplemented by sentiment data, but the lack of improvement with event based detection indicates that the specific events might not add any additional information that is not already available.

5 Conclusion

We were unable to show that supplementing sentiment analysis with event-based information could improve stock prediction models. On the topic of the low performance of the rolling window predictions. We believe this could be caused by the timing of the economic shift in the middle of our train test split, and suggest more advanced techniques to address temporal dependence. The lack of improvement when adding event-based data, raises two key considerations. The first is the fit of the data we used to fine-tine the model. The issues with balance, and the types of labels did not add additional information to our baseline models. The second is the value of events in addition to sentiment. The sentiment of the article may give you all the information you need to know and the event itself increase information gain. We still believe there is more information to be extracted from news articles about stock prices; but with regard to further research, we recommend attempting to understand what information can be extracted, which is not correlated with sentiment.

6 References

References

Alpha Vantage Inc. 2024. Alpha Vantage API. Accessed: 2024-11-25.

George E. P. Box and Gwilym M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control*. Holden-Day.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. 2024. Fnspid: A comprehensive financial news dataset in time series.

Ramesh Gupta, Lihua Chen, and Wei Zhang. 2024. Fnspid: A comprehensive financial news dataset in time series. *arXiv preprint arXiv:2402.06698*.

Ramsundar K. Hu, Jie Liu, Lutz Horn, and Nicholas S. Watson. 2021. Lora: Low-rank adaptation of large language models. *arXiv*.

Jainendra Kumar Jain and Ruchit Agrawal. 2024. FB-GAN: A novel neural sentiment-enhanced model for stock price prediction. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 85–93, Torino, Italia. Association for Computational Linguistics.

Tingsong Jiang and Andy Zeng. 2023. Financial sentiment analysis using finbert with application in predicting stock movement. *arXiv preprint arXiv:2306.02136*.

Kemal Kirtac and Guido Germano. 2024. Enhanced financial sentiment analysis and trading strategy development using large language models. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 1–10, Bangkok, Thailand. Association for Computational Linguistics.

Xiaodong Li, Pangjing Wu, and Wenpeng Wang. 2020. Incorporating stock prices and news sentiments for stock market prediction: A case of hong kong. *Information Processing & Management*, 57(5):102212.

Wen-Jie Liu, Ye-Bo Ge, and Yu-Chen Gu. 2024. News-driven stock market index prediction based on trellis network and sentiment attention mechanism. *Expert Systems with Applications*, 250:123966.

Yi Yang and Allen Huang. 2020. Finbert: A pretrained language model for financial text mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Qiuyue Zhang, Chao Qin, Yunfeng Zhang, Fangxun Bao, Caiming Zhang, and Peide Liu. 2022. Transformer-based attention network for stock movement prediction. *Expert Systems with Applications*, 202:117239.

L. Zhihan. 2021. Tradetheevent: A financial event classification dataset. https://github.com/Zhihan1996/TradeTheEvent/tree/main/data.