Summer 8-28-2023

Predictive AI for the S&P 500 Index

Jacqueline Rose Perry
Dartmouth College, jacqueline.r.perry.23@dartmouth.edu

PREDICTIVE AI FOR THE S&P 500 INDEX

A Thesis

Submitted to the Faculty

in partial fulfillment of the requirements for the

degree of

Bachelor of Arts

in

Computer Science

by

Jacqueline Perry

DARTMOUTH COLLEGE

Hanover, New Hampshire

August 2023

Abstract

Artificial intelligence has powerful applications in virtually every field, and the financial world is no exception. Utilizing various elements of artificial intelligence, this research aims to predict the future value of the S&P 500 index using numerous models, and in doing so, identify relevant features. More specifically, models that include combinations of historical data, public sentiment, and technical indicators were employed to predict the stock price one day and three days forward. To account for public opinion, the sentiment of tweets and news headlines from the beginning of 2015 through the end of 2019 was calculated using FinBERT, a pre-trained version of BERT retrieved from the HuggingFace Model Hub and designed specifically for financial-related text. For each textual input, FinBERT provides three outputs: the probability that the text is positive, negative or neutral. These probability values were applied to approximate the number of positive, negative, and neutral tweets and news headlines each day. The following features were used in complex LSTM models: open, close, low and high prices; volume; the number of positive, negative, and neutral tweets and news headlines; relative strength index; and earnings per share. The highest performing predictive model for one day forward and three days forward utilized historical data, tweet sentiment, and the relative strength index. Coupled with other tools wielded by investors, this model can help anticipate market movements and inform decisions.

Keywords: LSTM, S&P 500 Index, natural language processing, sentiment analysis, RSI, EPS

Acknowledgments

This work would not have been possible without the help, thoughtful guidance, and incredible support of my advisor Dr. Soroush Vosoughi. Thank you for always making your classes engaging, fun and challenging. Your teaching provided me with the tools and inspiration I needed to carry out this research. Thank you for advising me while on sabbatical and for always finding the time to help, even from different time zones.

I would also like to recognize the Dartmouth Computer Science department for enabling me to pursue this thesis research during the spring and summer of 2023.

Contents

Chapter 1

Introduction

The S&P 500 index (ticker "GSPC") is an extremely powerful and useful tool which measures the value of the aggregate S&P 500 with important implications for the overall performance of the United States economy. The S&P 500 and the general state of the market are driven by many factors such as current events (wars, natural disasters, pandemics, etc.), interest rates, earnings, the economy, actions by governments, etc.; however, it is also hugely influenced by a factor that is challenging to measure–how people feel about the market. The behavior of financial markets is often erratic and very difficult to predict, especially given the influence of investors' sentiment. Artificial intelligence has made enormous strides in the past twenty years, becoming more efficient, comprehensive and accurate, and yet more accessible. This presents an exciting opportunity to utilize state-of-the-art technology for the prediction of financial markets and stock pricing. Long Short-Term Memory (LSTM) models [Gee, 2023] in particular exhibit promise as studies show they outperform existing models for stock price prediction such as ARIMA, artificial neural networks [Ma, 2020], random forests, standard deep networks, and linear regression [Bhandari et al., 2022]. LSTM networks are a type of recurrent neural network which are specifically useful for time series predictions. LSTM models are preferred to these other models because they al-low for more complex features than stock-related data and they do not suffer from the vanishing gradient issue. LSTM models resolve the vanishing gradient by incorporating additional gates to store longer series of input data, and therefore larger amounts of past information [Siami-Namini et al., 2019]. There are two main challenges when developing an LSTM model: determining the structure of the neural network and selecting relevant, impactful features. This study seeks to compare the performance of models utilizing various sets of features and thus determine the importance of these features.

Related Works. In recent years, there has been significant research into machine learning and stock market forecasting. As machine learning has evolved, a variety of algorithms have been employed to predict stock prices. Some studies compare the efficacy of various predictive models while others identify and compare relevant features. Similar to the works of Ma and Bandari et al. referenced in the introduction, Karmiani et al. compare the following predictive algorithms for nine selected companies: backpropagation, support vector machines (SVM), Kalman Filter and LSTM [Karmiani et al., 2019]. Karmiani et al. concluded that in terms of high accuracy and low variance, LSTM produces the best results. T-test results indicated that LSTM is more reliable than backpropagation and SVM [Karmiani et al., 2019]. Htun et al. survey a variety of feature selection and extraction techniques and conclude that "accurate stock market predictions strongly depend on the selection of appropriate features" [Htun et al., 2023]. Bhandari et al. similarly construct a LSTM model with feature types such as market data, economic data, and technical indicators to predict stock price; however, they do not incorporate sentiment data [Bhandari et al., 2022]. Bhandari et al. also experimented with various network structures by modifying hyperparameters such as the number of layers, the number of neurons on each layer, and the batch size. Of significance, finding the best LSTM structure is a challenging task solved by trial and error. The process can be accelerated by investigating the structures of models used for similar purposes; although, ultimately the discovery effort is still subject to trial and error.

This thesis takes the LSTM model construct a step further by integrating sentiment analysis with market, economic and technical data. Other studies use similar sets of data to predict the future value of the CBOE Volatility Index [Hosker et al., 2018, Deveikyte et al., 2022], the expected value of volatility based on S&P 500 index options. Some studies employ a classification approach [Nelson et al., 2017], aiming to predict whether a stock's price will go up or down rather than predicting the actual price. There has been extensive research in stock price prediction; however, due to the complex, unpredictable nature of the stock market and the fact that machine learning techniques are constantly evolving, there remain numerous untried methods incorporating hundreds of factors across thousands of combinations which would likely yield improved results.

Chapter 2

Methodology

Section 2.1 Features: Market Data, Indicators and Sentiment

The most basic set of features, which is included in every financial market forecasting model, is the historical data of the S&P 500 index–open, low, high and close prices and volume. Investors often use a variety of indicators to inform their trading decisions. For example, earnings per share (EPS) indicates a company's profitability; in practice, investors might pay more for shares in a company if that company has greater earnings or profits relative to the share price [Fernando, 2023a]. For this reason, EPS was a considered feature. The relative strength index (RSI) is a momentum indicator which determines if a security is overbought or oversold [Fernando, 2023b]. By extension, it may anticipate trend reversals or a pullback in price. RSI is commonly used in technical analysis, so it is also utilized as a feature in some of the models. All of these features pertain to the actual stock price; however, as stated in the introduction, investor sentiment also plays a role in the state of the market. Investor sentiment was extracted from two sources: Twitter and news headlines. Twitter is a platform on which anyone can openly express their opinions about the market, mostly unfiltered. Thus, Twitter is especially useful in capturing the general public's feelings regarding the stock market. On the other hand, people often have inaccurate perceptions of the market. News headlines provide another source of sentiment and arguably, a more professional and reliable perspective of the stock market.

Section 2.2

Data Collection and Processing

Three main sources of data were used in this study: historical stock data retrieved using the Yahoo Finance python package yfinance, a csv file consisting of news headlines about more than 6,000 stocks [Kag, 2020], and another csv file containing tweets referring to top S&P 500 companies such as Amazon, Google, Microsoft, Apple, and Tesla [Metin, 2020]. At a first glance, this seems like it might be a source of bias in the model. Although the Twitter dataset did not consist of tweets referencing all S&P 500 companies, tweets often contained references to multiple companies. As a result, many other companies in the S&P 500 were referenced. The dataset probably does not provide a perfect representation of all S&P 500 companies, but this is not a major issue as a significant majority of the companies make up a small portion of the S&P 500's total value. Both csv files were downloaded from Kaggle and read into Pandas DataFrames. The news dataset was filtered to include only companies in the S&P 500 and, in order to correspond to the same dates as in the Twitter dataset, headlines from the beginning of 2015 through the end of 2019. The final tweets and headlines datasets consisted of roughly 3.5 million tweets and 165,000 news headlines. There are approximately 140,000 unique users in the Twitter dataset. Given the information in the dataset, it is not possible to determine how diverse, and therefore representative, the users are. However, the results indicate that tweets are a valuable feature. The headlines dataset consists of headlines for 379 S&P 500 stocks scraped from Benzinga, a financial news website that aggregates content from various other sources such as TheStreet and the Wall Street Journal.

Using FinBERT [Araci, 2019], sentiment analysis was performed on both Kaggle datasets. Each tweet or headline was tokenized using FinBERT's corresponding pre-trained tokenizer, and arrays of tokenized text were passed to FinBERT as input in batches of ten. The output is passed through a softmax function and consists of three values for each tweet or headline: the probability that the text is positive, negative, or neutral. The historical data for the S&P 500 index over the same time frame as the other two datasets was retrieved using yfinance. The number of positive, negative, and neutral headlines or tweets per day were summed up using the probabilities. Prior to being passed into the LSTM model and with respect to individual features, all input data was scaled to values between 0 and 1 using sklearn's MinMaxScaler. Predicted values must then be transformed back to true values using the specific scaler instance applied to the open prices initially passed into the model. Other sklearn scalers, such as StandardScaler and RobustScaler, were also considered; however, MinMaxScaler produced the best results. Historical data, namely volume and open, close, high and low prices, and sentiment values were used to establish four initial sets of features: 1) only historical data; 2) historical data and news headlines; 3) historical data and tweets; and 4) historical data, tweets and news headlines.

Section 2.3

Model Structure

Two distinct LSTM structures were created to make two kinds of predictions: one to predict stock price one day forward and the other to predict 3 days forward. The simplest models, one day forward without the addition of RSI or EPS, require a slightly different LSTM structure than the rest of the models. They consist of two LSTM layers followed by a dropout layer with a rate of 0.2, another LSTM, a second dropout layer with a rate of 0.2 and then two dense layers with one unit each. The first LSTM layer has 150 units and a ReLU activation function, the second LSTM layer has 80 units, and the third has 60 units. All other models consist of two LSTM layers followed by a dropout layer with a rate of 0.2 and then two dense layers, both with three units. The first LSTM layer has 150 units and a ReLU activation function, and the second LSTM layer has either 50 or 60 units. The models are compiled using Adam optimizers and mean squared error loss functions and run for 150 epochs with batch sizes of 30. Models were created using the four initial sets of features with and without the addition of some technical market indicators. These technical indicators include earnings per share (EPS) of the top 30 S&P 500 companies based on market weight and the relative strength index (RSI) of the S&P 500 index, an indication of whether a stock is overbought or oversold. This study considers eight base models including and excluding technical indicators. Let model types I through IV refer to models using only historical data; historical data and news headlines; historical data and tweets; and historical data, tweets, and news headlines, respectively. Then let A indicate that the model predicts the open price for the next day, and let B indicate that the model predicts the price for the next three days. Thus models are labeled Type IA, Type IB, Type IIA, etc. The table below summarizes the model types. The random state is set so that results are reproducible and the effect of various features is more easily detected. The data is a time series, meaning that the data must be passed into the neural network in chronological order. The training set consists of the first 80% of the dataset, or approximately from 2015 to the end of 2018, and the last 20%, roughly all of 2019, is the test set.

Model Name	Historical Data	News headlines	Tweets	One day forward	Three days forward
Type IA	✔			✔	
Type IIA	✔	✔		✔	
Type IIIA	✔		✔	✔	
Type IVA	✔	✔	✔	✔	
Type IB	✔				✔
Type IIB	✔	✔			✔
Type IIIB	✔		✔		✔
Type IVB	✔	✔	✔		✔

Table 2.1: Summary of the LSTM models according to the number of days being predicted and the features used.

Chapter 3

Results

Numerous models were trained and tested in order to determine the efficacy of various features. The following metrics were used to evaluate each model: root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and R2 score.

Section 3.1

Predicting One Day Forward

Using the previous 10 days, the models make a prediction for the open price on the 11th day. The table below summarizes the performance of the base models with and without the addition of RSI and EPS. The use of EPS improves the Type IA model significantly and all others by a few percentage points. The addition of RSI improves all of the models significantly. Interestingly, the use of both RSI and EPS does not produce better results overall than either of them individually. The graph below illustrates the predictions of the models using RSI compared to the actual value of the S&P 500 index.

Model	RMSE	MAE	MAPE	R2 Score
Type IA	79.68	69.04	0.023	0.717
Type IIA	40.96	30.85	0.010	0.925
Type IIIA	38.16	29.16	0.0098	0.935
Type IVA	38.65	29.75	0.010	0.933
Type IA with RSI	25.30	20.34	0.007	0.971
Type IIA with RSI	28.76	23.85	0.008	0.963
Type IIIA with RSI	23.86	19.19	0.007	0.975
Type IVA with RSI	28.82	22.40	0.008	0.963
Type IA with EPS	42.82	37.48	0.013	0.918
Type IIA with EPS	27.80	21.76	0.008	0.966
Type IIIA with EPS	27.79	21.61	0.008	0.966
Type IVA with EPS	34.66	28.84	0.010	0.946
Type IA with RSI and EPS	25.11	20.17	0.007	0.972
Type IIA with RSI and EPS	32.89	24.79	0.008	0.952
Type IIIA with RSI and EPS	51.04	46.10	0.016	0.884
Type IVA with RSI and EPS	39.34	32.19	0.011	0.931

Table 3.1: Performance metrics for the LSTM models predicting one day forward.

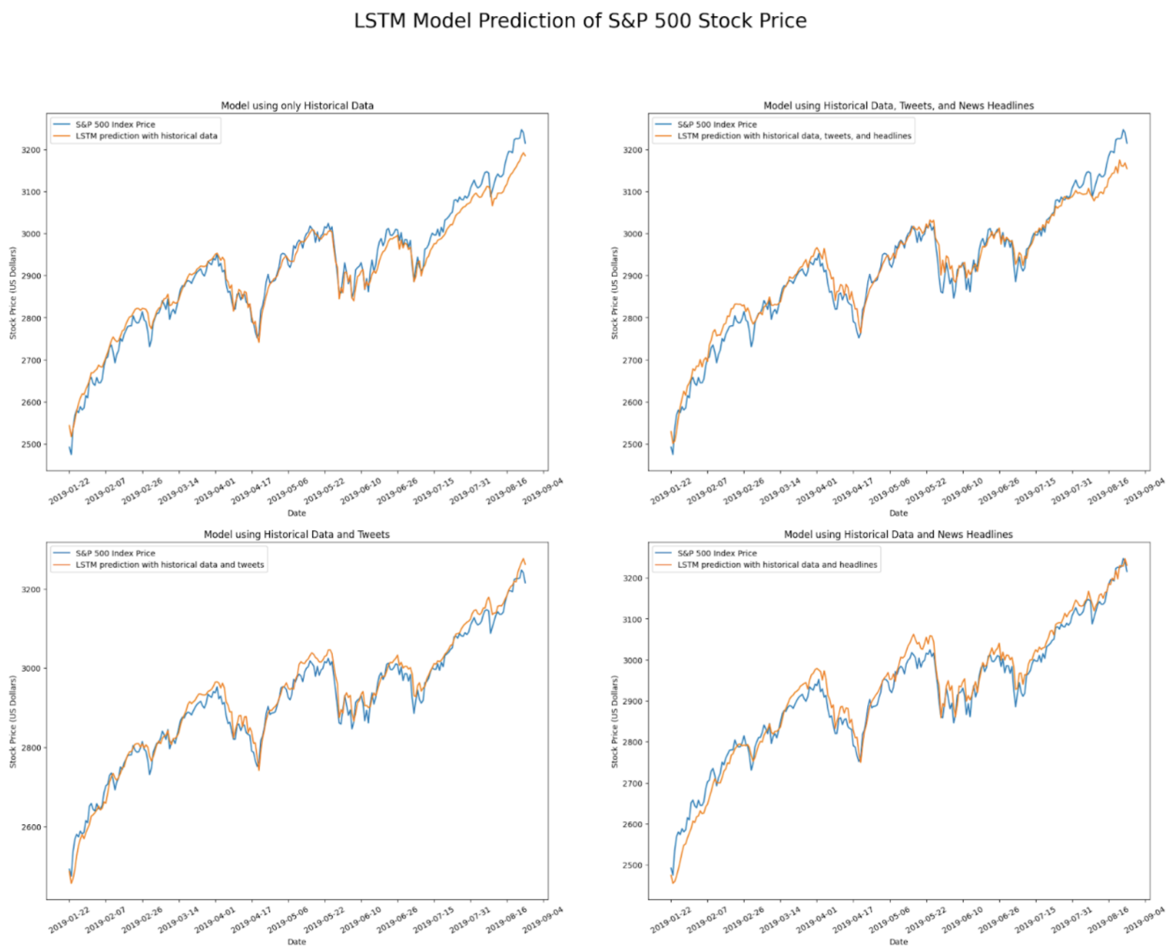

Figure 3.1: Graph of the LSTM models with RSI predicting one day forward vs. the actual value of the S&P 500 index.

Section 3.2

Predicting Three Days Forward

Using the previous 20 days, the models make predictions for the open prices over the next three days. The table below summarizes the performance of the models with and without RSI and EPS. As was the case when predicting one day forward, the use of RSI greatly improves the predictions by all the models. The use of EPS results in an overall decline in performance, and the models using both RSI and EPS perform worse than the models only using RSI. These predictions were not graphed because each data point consists of three prediction values, and this would not be well represented with a graph.

Model	RMSE	MAE	MAPE	R2 Score
Type I	41.11	31.93	0.011	0.927
Type IIB	55.13	41.89	0.014	0.868
Type IIIB	38.37	30.33	0.011	0.936
Type IVB	34.08	26.35	0.009	0.950
Type IB with RSI	35.06	27.02	0.009	0.947
Type IIB with RSI	34.22	27.61	0.010	0.949
Type IIIB with RSI	28.00	21.86	0.008	0.966
Type IVB with RSI	30.55	23.91	0.008	0.959
Type IB with EPS	49.06	41.54	0.014	0.896
Type IIB with EPS	48.70	39.16	0.014	0.897
Type IIIB with EPS	47.00	38.04	0.013	0.904
Type IVB with EPS	58.69	46.36	0.016	0.850
Type IB with RSI and EPS	63.95	53.84	0.019	0.822
Type IIB with RSI and EPS	42.43	32.43	0.011	0.923
Type IIIB with RSI and EPS	34.55	27.21	0.009	0.948
Type IVB with RSI and EPS	35.79	28.29	0.010	0.944

Table 3.2: Performance metrics for the LSTM models predicting three days forward.

Chapter 4

Discussion

Nearly all the models have low error values relative to the scale of the stock prices which range from approximately \$2,500 to \$3,200 in the test set. In fact, most of the MAE and RMSE values are similar to the difference between the high and low price on any given day. From the results, it is clear that features have varying impacts on model performance. RSI seems to have the greatest impact on the performance of all the models, decreasing errors and increasing R2 scores. It is not surprising that RSI is helpful in next-day predictions given that in practice it may "provide short-term traders with buy and sell signals" [Fernando, 2023b]. Interestingly, when predicting three days forward, the use of RSI improved when a window of eight days was used rather than the standard 14. However, the standard 14-day window produced better results when predicting one day forward. Traders typically use windows smaller than 14 days when they want to discern short-term trends [Rob, 2023]. This seems counter-intuitive given that one day is more short-term than three days. On the other hand, EPS did not prove particularly helpful. For type A models, EPS mildly improved the results; for type B models, it actually worsened the results. For both types A and B, the highest performing model (Type IIIA with RSI and Type IIIB with RSI) had the following features: open, close, high, and low prices; volume; sentiment of tweets; and relative strength index. The Type III models with RSI had the lowest RMSE, MAE, and MAPE values and the highest R2 scores, 0.975 for Type IIIA and 0.966 for Type IIIB. The high R2 score for the Type IIIA model is visualized in figure 3.1; the bottom left graph illustrates how the Type IIIA model's predictions closely fit the true S&P 500 index price. When considering these results, it is important to note that they may vary across operating systems and with different package versions. These models were run with macOS Ventura version 13.3. The following packages and versions were utilized: Python 3.9.13, Pandas 2.0.3, Numpy 1.25.0, TensorFlow 2.13.0, Scikit-Learn 1.3.0, Keras 2.13.1, and yfinance 0.2.26.

Section 4.1

Limitations

There are numerous sources of limitations in this particular study, including time and computational restrictions, the lack of interpretability of artificial intelligence, and data limitations. Time complexity was particularly limiting with respect to sentiment data. BERT and FinBERT are extremely slow, especially when calculating the sentiment of large sets of text. A larger amount of sentiment data would have been preferable for this study, but without more computational power, it was not feasible. Furthermore, it proved difficult to obtain additional data on which to perform the sentiment analysis. Twitter's (now called "X") recent changes to its API restrict the number of tweets that can be freely retrieved in a month (above the threshold, users are required to pay a fee in order to access additional tweets). A very large amount of data was required to capture representative, broad-based public sentiment for each day; as such, sources were limited to large open-access datasets. Reddit was also initially considered as a possible source of sentiment; however, given the typically long length of reddit posts, sentiment analysis of a sufficiently large dataset would have taken a prohibitively long time to process with the available computational power. Furthermore, the lack of interpretability of neural networks poses a limitation in how the results are utilized. Because neural networks are essentially "black boxes," how the models make predictions cannot be well explained. Although the models perform fairly well, there is considerable opportunity for improvement. Thus, investors should not solely rely on the predictions of LSTM models; LSTM models are but one of many tools traders should consider when making investment decisions.

Section 4.2

Future Directions and Applications

4.2.1. Future Directions

There are a multitude of directions this study could expand upon, and all have the potential to add value. The most immediate extension of this work is to improve the models by considering more features and evaluating their impact on performance. Some potential feature types that may add value are technical and macroeconomic indicators. More sources of public sentiment could also provide a robust, representative perspective of the market. Similarly, more financial news data from sources such as Robinhood, the Wall Street Journal and Barons could improve the predictability of the models. It would also be interesting to investigate whether some features perform best under certain conditions. For example, it is plausible that some features produce better results during an economic bubble, a rapid increase in stock price beyond its intrinsic value. More accurate predictions could also be produced with the use of a similar type of network: a Bidirectional LSTM (BiLSTM). BiLSTMs traverse input data twice, once from left to right and again from right to left, and therefore undergo additional training. Research shows that BiLSTM models produce better predictions than LSTM models; however, BiLSTM models are much slower than LSTM models [Siami-Namini et al., 2019]. Another possible extension of this work is to use the predictions in conjunction with deep reinforcement learning to construct effective trading strategies. Kabbani and Duman found that including technical indicators and sentiment scores in state representation improved the trading agent's performance [Kabbani and Duman, 2022]. The LSTM predictions could be included in the state representation and used by the agent to make investment decisions.

4.2.2. Applications

With the recent growing popularity of 0DTE (0 days to expiration) options contracts, LSTM models have the potential to inform traders of potential short-term market changes. The Financial Industry Regulatory Authority (FINRA) has noted that between January 2022 and January 2023 "the number of opening 0DTE option contracts positions increased approximately 60 percent" [FIN, 2023]. More specifically, as of this past February, 0DTE options contracts accounted for around 44% of the 10-day average daily S&P 500 options volume (Ahmed, 2023). These models can also be used for other short-term decisions. This overall methodology can be adapted for any security, combination of securities, or market indices given the proper information.

Chapter 5

Conclusion

This study evaluates the effectiveness of RSI, EPS, and sentiment of tweets and news headlines as features for predicting the movement of the S&P 500 index one day forward and three days forward using a complex, multilayer LSTM model. Multiple models utilizing the various features show encouraging results in predicting the price of the S&P 500 index in the short-term. The use of sentiment data and RSI as features, in addition to historical data, produced the best results overall. More specifically, the incorporation of RSI greatly improves nearly all predictions. It is possible that a larger set of sentiment data from a greater number of sources could further improve results. Although more research is necessary, the initial results show promise, suggesting that the technical indicators and public sentiment–through both official and unofficial channels–can effectively improve predictions for the future value of the S&P 500 index. Next steps include identifying other meaningful features, exploring data types and datasets which strengthen the model, and experimenting with various model structures to identify the most effective architecture.

Bibliography

[Kag, 2020] (2020). *Daily Financial News for 6000+ stocks.* Kaggle. https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests.

[Gee, 2023] (2023). *Deep learning: Introduction to long short term memory.* GeeksforGeeks. https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/.

[Rob, 2023] (2023). *Which timeframe is best for RSI?* The Robust Trader. https://therobusttrader.com/which-timeframe-is-best-for-rsi/.

[FIN, 2023] (2023). *Zeroing in on an options trading strategy: 0DTE.* FINRA. https://www.finra.org/investors/insights/zeroing-in-options-trading-strategy.

[Araci, 2019] Araci, D. T. (2019). *Prosusai/finbert.* Hugging Face. https://huggingface.co/ProsusAI/finbert.

[Bhandari et al., 2022] Bhandari, H. N., Rimal, B., Pokhrel, N. R., Rimal, R., Dahal, K. R., and Khatri, R. K. (2022). Predicting stock market index using lstm. *Machine Learning with Applications*, 9:100320.

[Deveikyte et al., 2022] Deveikyte, J., Geman, H., Piccari, C., and Provetti, A. (2022). A sentiment analysis approach to the prediction of market volatility. *Frontiers in Artificial Intelligence*, 5.

[Fernando, 2023a] Fernando, J. (2023a). *Earnings per share (EPS): What it means and how to calculate it.* Investopedia. https://www.investopedia.com/terms/e/eps.asp.

[Fernando, 2023b] Fernando, J. (2023b). *Relative strength index (RSI) indicator explained with formula.* Investopedia. https://www.investopedia.com/terms/r/rsi.asp.

[Hosker et al., 2018] Hosker, J., Djurdjevic, S., Nguyen, H., and Slater, R. (2018). Improving vix futures forecasts using machine learning methods. *SMU Data Science Review*, 1(4).

[Htun et al., 2023] Htun, H. H., Biehl, M., and Petkov, N. (2023). Survey of feature selection and extraction techniques for stock market prediction. *Financial Innovation*, 9. Article 26.

[Kabbani and Duman, 2022] Kabbani, T. and Duman, E. (2022). Deep reinforcement learning approach for trading automation in the stock market. *IEEE Access*, 10:93564–93574.

[Karmiani et al., 2019] Karmiani, D., Kazi, R., Nambisan, A., Shah, A., and Kamble, V. (2019). Comparison of predictive algorithms: Backpropagation, svm, lstm and kalman filter for stock market. In *2019 Amity International Conference on Artificial Intelligence (AICAI)*, pages 228–234.

[Ma, 2020] Ma, Q. (2020). Comparison of arima, ann and lstm for stock price prediction. *E3S Web Conf.*, 218:01026.

[Metin, 2020] Metin, (2020). *Tweets about the top companies from 2015 to 2020.* Kaggle. https://www.kaggle.com/datasets/omermetinn/tweets-about-the-top-companies-from-2015-to-2020.

[Nelson et al., 2017] Nelson, D. M. Q., Pereira, A. C. M., and de Oliveira, R. A. (2017). Stock market's price movement prediction with lstm neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1419–1426.

[Siami-Namini et al., 2019] Siami-Namini, S., Tavakoli, N., and Namin, A. S. (2019). The performance of lstm and bilstm in forecasting time series. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 3285–3292.