Linear Forecasting: Predictive Stock Market Trends

Rohan Shah and Joshua Kannagala

Abstract This research explores the application of machine learning techniques to predict stock market prices, integrating historical stock data with sentiment analysis. We examine various machine learning models, including Long Short-Term Memory (LSTM) networks and Principal Component Analysis (PCA), to assess their effectiveness in predicting stock trends. After careful evaluation, sentiment analysis using social media and news data was incorporated to capture market sentiment, providing a more comprehensive approach to stock price prediction. The model developed in this study demonstrates a 3.2% improvement in accuracy when sentiment analysis is included, as measured by the reduction in Mean Absolute Error (MAE). Despite the model's enhanced accuracy, challenges such as the limited availability of historical sentiment data and the potential for overfitting are acknowledged. The findings suggest that combining quantitative financial data with qualitative sentiment insights can lead to more accurate stock price predictions. This research contributes to the growing body of work focused on improving predictive models in finance by incorporating both traditional and sentiment-based analysis techniques.

Rohan Shah
University of Illinois, Champaign, IL, e-mail: rohan.m.s1211@gmail.com

Joshua Kannagala
University of Illinois, Champaign, IL e-mail: joshuakann23@gmail.com

1 Introduction

Predicting stock prices is a complex and challenging task, yet it remains such a fundamental objective for investors and financial analysts aiming to make informed decisions. The motivation for this project arises from the constantly changing nature of the stock market, influenced by economic indicators, market sentiment, and global events. In a world where financial markets can turn on a dime, harnessing the power of machine learning presents an exciting opportunity to enhance the accuracy of stock price predictions [1].

This research explores the potential of machine learning techniques to predict stock prices, providing a thorough analysis of various models and their performance. By combining historical stock data with sentiment analysis, the study aims to uncover patterns and insights that can improve prediction models.

2 Related Work and Background

Previous research on stock price prediction has implemented many different different machine learning methods, including LSTM models, PCA for reducing data complexity, and sentiment analysis. We combined the best ideas and positive outcomes of these approaches to build a model that uses both past data and sentiment scores to improve prediction accuracy.

2.1 Stock Prediction with High Accuracy using Machine Learning Techniques

In this article, authors Bansal, Goyal, and Choudhary examine and compare the use of five different machine learning algorithms to accurately predict stock market trends in the Indian Stock Market. These five models include Linear Regression, Support Vector Machines (SVM), and Long Short-Term Memory (LSTM) networks, K-Nearest Neighbors, and Decision Tree Regression networks. Through a comparative analysis of the performances of each method, the study concludes that deep learning algorithms proved to be the most accurate [2].

2.2 Stock Price Prediction using LSTM and its Implementation

In this article, Siddharth M explores how accurate Long Short-Term Memory (LSTM) networks are in analyzing complex patterns in historical stock price data. The article provides a detailed tutorial on using LSTM networks for stock price prediction, covering the theoretical background of LSTM, the challenges of traditional methods, the architecture of LSTM, and a practical implementation using Python. By leveraging LSTM's ability to capture long-term dependencies, the study demonstrates how LSTM networks can effectively analyze time-series data and predict stock prices [3].

2.3 Stock Price Prediction using Principal Components (PCA)

In this article, authors Ghorbani and Chong investigate the application of Principal Component Analysis (PCA) in predicting stock market prices. The study uses PCA to reduce dataset dimensionality, focusing on principal components that capture the most variance. The authors apply PCA to historical stock data and use these components in machine learning models to predict future prices. Their findings demonstrated that PCA enhances prediction accuracy by isolating key patterns and trends, making the models more robust compared to those without PCA integration [4].

2.4 Stock trend prediction using sentiment analysis

In this article, Xiao and Ihnaini introduce a novel weighted sentiment index for predicting stock trends using Twitter and news data, analyzed with VADER and FinBERT, respectively. They compare the predictive performance of tweets and news across different time divisions, employing algorithms like K-Nearest Neighbors, Decision Trees, SVM, Random Forests, Naïve Bayes, and Logistic Regression. The study finds that Naïve Bayes consistently delivers the best accuracy, particularly with skewed datasets, demonstrating that combining news and tweet sentiment scores improves prediction accuracy [5].

3 Data and Methodologies

To build a comprehensive stock price prediction model, various tools and methods were utilized to ensure accuracy and relevance. Google Cloud APIs played an integral role in efficiently retrieving data using Python. This allowed large volumes of information from historical stock data sources to be automated onto spreadsheets. Yahoo Finance served as the primary source for historical and real-time stock data, providing extensive and reliable market information. Additionally, sentiment scores and data were gathered using VADER and BeautifulSoup to incorporate market sentiment into the model, enhancing its predictive power by combining both quantitative financial data and qualitative insights [6].

3.1 Data Collection

After gathering historical stock data from many companies over the past ten years, the data was organized in Google Sheets. The spreadsheets focused on technology companies, including major companies like Apple, Microsoft, and Google. Following the collection of basic information from Yahoo Finance, additional metrics were calculated using Python and subsequently stored in the spreadsheets for further analysis with machine learning.

3.1.1 Company Historical Stock Price Data

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits	Symbol	SMA_20	EMA_20	Volatility_20	BB_upper	BB_lower
2024-08-13 00:0	219.0099945068	221.8899993896	219.0099945068	221.2700042724	44155300	0.0	0.0	AAPL	nan	221.2700042724	nan	nan	nan
2024-08-12 00:0	216.0700073242	219.5099945068	215.6000061035	217.5299987792	38028100	0.25	0.0	AAPL	nan	220.9138132731	nan	nan	nan
2024-08-09 00:0	211.8547924511	216.5293744807	211.7249378699	215.9900054931	42201600	0.0	0.0	AAPL	nan	220.4448791988	nan	nan	nan
2024-08-08 00:0	212.8636192825	213.9523554508	208.5885687073	213.0633850097	47161100	0.0	0.0	AAPL	nan	219.7418797522	nan	nan	nan
2024-08-07 00:0	206.6607856530	213.3929986622	206.1513807783	209.5774230957	63516400	0.0	0.0	AAPL	nan	218.7738362611	nan	nan	nan
2024-08-06 00:0	205.0626409492	209.7472209414	200.8375358240	206.9904022216	69660500	0.0	0.0	AAPL	nan	217.6516044478	nan	nan	nan
2024-08-05 00:0	198.8598223925	213.2531662150	195.7733984925	209.0280609130	119548600	0.0	0.0	AAPL	nan	216.8303145874	nan	nan	nan
2024-08-02 00:0	218.8966186199	225.3391734979	217.4582963114	219.6058044433	105568600	0.0	0.0	AAPL	nan	217.0946469546	nan	nan	nan
2024-08-01 00:0	224.1105900038	224.2204634371	216.7690968426	218.1075439453	62501000	0.0	0.0	AAPL	nan	217.1911133347	nan	nan	nan
2024-07-31 00:0	221.1839827384	223.5612359588	220.3749216653	221.8232421875	50036300	0.0	0.0	AAPL	nan	217.6322684635	nan	nan	nan
2024-07-30 00:0	218.9365913481	220.0752727557	215.8701333369	218.5470428466	41643800	0.0	0.0	AAPL	nan	217.7193898333	nan	nan	nan
2024-07-29 00:0	216.7091672552	219.0464581940	215.5005594970	217.9876861572	36311800	0.0	0.0	AAPL	nan	217.7449418642	nan	nan	nan
2024-07-26 00:0	218.4471424920	219.2362376525	215.7602501518	217.7080078125	41601300	0.0	0.0	AAPL	nan	217.7414243354	nan	nan	nan
2024-07-25 00:0	218.6768782594	220.5946718767	214.3718636751	217.2385559082	51391200	0.0	0.0	AAPL	nan	217.6935321043	nan	nan	nan
2024-07-24 00:0	223.7410169933	224.5400951022	216.8789647868	218.2873229980	61777600	0.0	0.0	AAPL	nan	217.7500836179	nan	nan	nan
2024-07-23 00:0	224.1105879628	226.6776239526	222.4225394312	224.7498474121	39960300	0.0	0.0	AAPL	nan	218.4167277888	nan	nan	nan
2024-07-22 00:0	226.7475419395	227.5166559886	222.8320757894	223.7010803222	48201800	0.0	0.0	AAPL	nan	218.9199994587	nan	nan	nan
2024-07-19 00:0	224.5600792921	226.5377858266	223.0218512443	224.0506591796	49151500	0.0	0.0	AAPL	nan	219.4086337178	nan	nan	nan
2024-07-18 00:0	230.0137604637	230.1735791375	222.0130267153	223.9208068847	66034600	0.0	0.0	AAPL	nan	219.8383644956	nan	nan	nan
2024-07-17 00:0	229.1847200362	231.1924059439	226.3779712354	228.6153869628	57345900	0.0	0.0	AAPL	218.3896133422	220.6742713972	5.570622380100	229.5308581024	207.2483685820
2024-07-16 00:0	234.7283075285	235.9968435006	232.0613962464	234.5485229492	43234300	0.0	0.0	AAPL	219.0535392761	221.9956286879	6.623731637050	232.3010025502	205.8060760020
2024-07-15 00:0	236.2065948881	236.9557277930	232.8205147673	234.1289978027	62631300	0.0	0.0	AAPL	219.8834892272	223.1511876512	7.415398384822	234.7142859969	205.0526924576
2024-07-12 00:0	228.6553306110	232.3710309246	228.4156026019	230.2734527587	53046500	0.0	0.0	AAPL	220.5976615905	223.8294986138	7.702924305316	236.0035102012	205.1918129799
2024-07-11 00:0	231.1224755684	232.1213194086	225.5089780638	227.3069000244	64710600	0.0	0.0	AAPL	221.3098373413	224.1606797006	7.627755705601	236.5653487525	206.0543259301

Fig. 1 AAPL Stock Data Metrics

3.1.2 Sentimental Scores

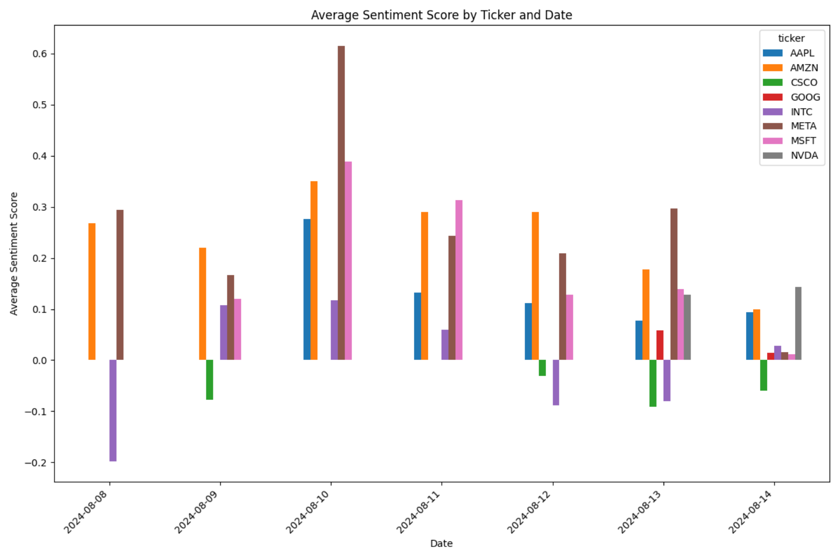

Fig. 2 Average Sentiment Scores of Several Technology Companies in a week

3.2 Methods

3.2.1 Visual Studio Code

Visual Studio Code served as the main coding platform due to its flexibility and extensive support for Python libraries. As a result, Python was also chosen as the main programming language, as many of these libraries were needed for machine learning. Key libraries included:

- **Pandas**: Used for data manipulation and analysis, particularly for organizing stock price data into data frames for easy computation.
- **Yfinance**: A Python library providing easy access to historical market data from Yahoo Finance, allowing the retrieval of stock prices, volume, and other important financial metrics.
- **Scikit-learn (Sklearn)**: Employed for implementing machine learning models, particularly for preprocessing data, training models, and evaluating their performance.

3.2.2 Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable (in this case, stock prices) and one or more independent variables (such as historical prices and sentiment scores). This approach was employed to predict future stock prices by identifying the linear relationship between these variables. The model calculates a best-fit line that minimizes the difference between predicted and actual stock prices [7].

3.2.3 Mean Absolute Error

Mean Absolute Error (MAE) is a metric used to evaluate the accuracy of regression models. It measures the average magnitude of errors between predicted and actual values, providing an indication of how close the predictions are to the real stock prices. In this study, MAE was utilized to assess the performance of the linear regression model, serving as a key metric to determine the model's accuracy and effectiveness [8].

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} | x_i - x |$$

Fig. 3 Mean Absolute Error Formula

3.2.4 Sentiment Methods

To incorporate sentiment analysis into the stock price predictions, several tools and methodologies were utilized:

- **VADER Sentiment Analysis**: Used to analyze the sentiment of text data from social media posts and news articles. VADER is particularly well-suited for analyzing short, informal text from sources like Twitter, making it ideal for this project.
- **BeautifulSoup**: A Python library used for web scraping, which allows the collection of news articles from various financial news websites. This helped create a comprehensive dataset of news sentiments to be analyzed alongside stock prices.
- **Finviz**: An online stock screener that was used as the main source of additional sentiment data, news articles, and financial statements.

4 Results

To evaluate whether sentiment analysis had an affect on the model's accuracy, the mean absolute error was calculated on both the training and testing data sets for the eight technology stocks. Through the tables and column charts below, a comparison can be drawn between the MAE of the training and testing sets with and without sentiment analysis.

4.1 Model Evaluation

Stock	Training Dataset (MAE)	Testing Dataset (MAE)
APPL	1.72	1.64
MSFT	2.6	2.4
GOOG	1.15	1.15
AMZN	1.88	1.65
META	4.15	3.9
NVDA	0.66	0.64
INTC	0.63	0.63
CSCO	0.46	0.45

Fig. 4 MAE without Sentiment Analysis

Stock	Training Dataset (MAE)	Testing Dataset (MAE)
APPL	1.65	1.63
MSFT	2.2	2.5
GOOG	0.95	1.12
AMZN	1.73	1.54
META	3.78	3.71
NVDA	0.59	0.62
INTC	0.6	0.62
CSCO	0.41	0.43

Fig. 5 MAE with Sentiment Analysis

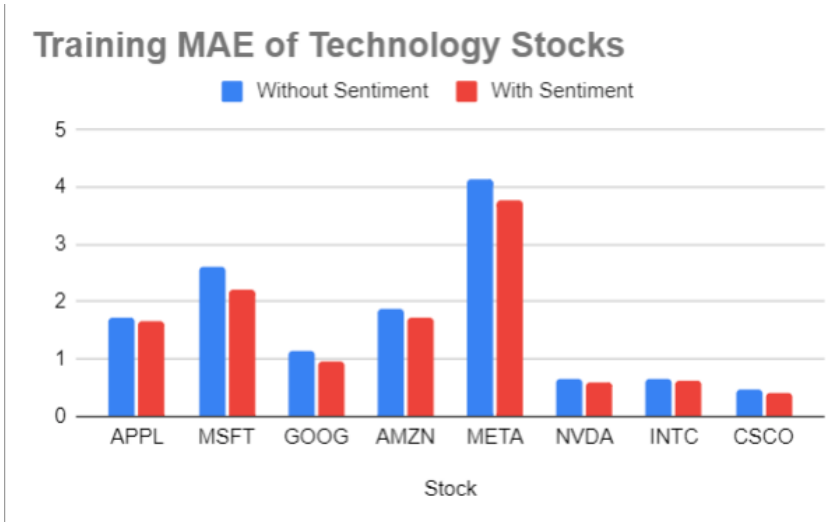

Fig. 6 MAE in Training Set

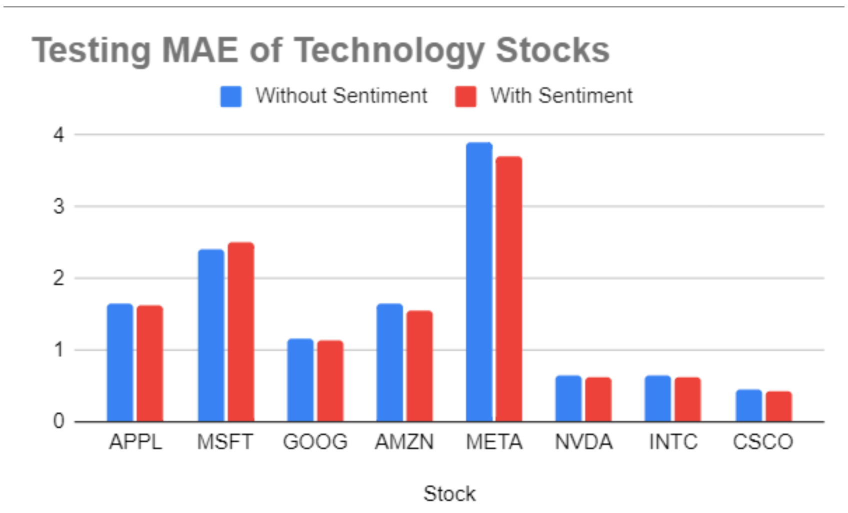

Fig. 7 MAE in Testing Set

Through observing the results, the integration of sentiment analysis led to a 3.2% reduction in Mean Absolute Error, enhancing the model's accuracy. Despite the challenges posed by the limited historical sentiment data, the inclusion of sentiment metrics provided valuable insights, particularly for stocks with high public engagement.

5 Conclusion

The research presented demonstrates that the integration of machine learning techniques with sentiment analysis can enhance the predictive accuracy of stock prices. The model developed showed a Mean Absolute Error reduction of 3.2% when sentiment analysis was utilized, indicating a concrete improvement in prediction accuracy. However, it was observed that the model performed better on the training set than on the testing set, suggesting potential overfitting. This discrepancy points to the need for further refinement in the model to ensure that it generalizes well to unseen data.

5.1 Applications

The methodology and findings from this research have several potential applications in finance. For example, investment firms can utilize the enhanced model to make more informed trading decisions by integrating real-time sentiment data with historical financial metrics [9]. Additionally, the approach could be adapted for use in predicting price movements in other financial markets, such as commodities or foreign exchange. The integration of sentiment analysis can also be extended to other domains where understanding the influence of public sentiment is critical, such as in consumer behavior analysis or political forecasting.

5.2 Concerns and Next Steps

Several concerns were identified during the research process. The limited availability of historical sentiment data restricted the analysis primarily to recent trends, which may have affected the model's overall performance. This is due to the fact that the model relied on historical stock price data from the past ten years, and the sentiment data was not available for every single one of those dates. Additionally, the model was only trained on stocks from the technology sector, which are inherently volatile and may not represent broader market dynamics [10]. Future work should focus on expanding the dataset to include a wider range of industries and a more extensive historical sentiment dataset. This would help limit overfitting and improve the model's generalization to unseen data. Additionally, exploring more sophisticated machine learning algorithms and refining sentiment analysis techniques could be a next step in enhancing the model's robustness and accuracy.

Acknowledgements We would like to thank Professor Christof Teuscher and the altREU faculty mentors for their helpful guidance and feedback throughout this research project.

References

[1] Y. Chen, S. H. Sim, S. Yan, and M. Dong, "Stock market prediction using artificial neural networks and sentiment analysis," *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1061-1066, 2013, doi: 10.1109/SMC.2013.187.

[2] T. M. Dang et al., "A hybrid deep learning model for stock market prediction," *Procedia Computer Science*, vol. 207, pp. 1230-1242, 2022, doi: 10.1016/j.procs.2022.09.121.

[3] M. S. (2024, July 29). Stock price prediction using LSTM and its implementation. *Analytics Vidhya*.

[4] Ghorbani, M., & Chong, E. K. P. (2020, March 20). Stock price prediction using principal components. *National Institutes of Health*.

[5] Xiao, Q., & Ihnaini, B. (2023, March 20). Stock trend prediction using sentiment analysis. *National Institutes of Health*.

[6] Cristescu, M. P., Nerisanu, R. A., Mara, D. A., & Oprea, S.-V. (2022, November 14). Using market news sentiment analysis for stock market prediction. *MDPI*.

[7] Mo, H. (2023, December). Comparative analysis of linear regression, polynomial regression, and ARIMA model for short-term stock price forecasting. *ResearchGate*.

[8] Hodson, T. O. (2022, July 19). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*.

[9] Chen, Y., Zhao, H., Li, Z., & Lu, J. (2020, December 4). A dynamic analysis of the relationship between investor sentiment and stock market realized volatility: Evidence from China. *National Institutes of Health*.

[10] Rasiova, B., & Arendas, P. (2022, December 1). Copula approach to market volatility and technology stocks dependence. *Finance Research Letters*.