Correlation between Social Media and Stock Market in Stock Price Prediction

Aishwarya Muralidharan Nair
Department of ECE
Stevens Institute of Technology
Hoboken, NJ, USA
anair9@stevens.edu

Udhayan Nateshan Ilangovan
Department of ECE
Stevens Institute of Technology
Hoboken, NJ, USA
unatesha@stevens.edu

Simrat Kaur Anand
Department of ECE
Stevens Institute of Technology)
Hoboken, NJ, USA
sanand5@stevens.edu

***Abstract*—This project aims to develop a stock price prediction model that includes a feature of news data along with the previous stock. Through this project, we hope to contribute to the development of new techniques for using news data along with trends in stock prices to predict future stock prices and improve the accuracy of stock price prediction models.**

I. Introduction

Stock price prediction is a challenging task that has gained considerable attention in recent years due to its importance in every field. The ability to predict stock prices accurately is critical for investors and traders as it allows them to make informed decisions about buying or selling stocks. However, this is a complex problem which is influenced by many factors, including but not limited to economic indicators, market trends, and social media/news articles. In recent years, there has been a growing interest in using social media and news data to predict stock prices. News articles and social media discussions can have a significant impact on the stock market, as they provide insights into a company's performance, prospects, and overall market sentiment.

The objective of this project is to develop a stock price prediction model that uses previous stock prices along with news data to predict the stock prices of a set of companies. Specifically, we will explore the correlation between news data and stock prices and use this information to train a machine learning model to predict future stock prices.

To achieve this objective, we have collected a dataset of news articles and stock prices for a set of eighty-two companies. We then preprocess the news data to extract relevant features such as sentiment, topic, and source of news. Next, we will explore the correlation between news data and stock prices using statistical analysis and visualization techniques. We then use different machine learning algorithms to predict the future stock prices based on this consolidated dataset which comprise of both news data and stock price data.

In summary, this project aims to develop a stock price prediction model that includes a feature of news data along with the previous stock. Through this project, we hope to contribute to the development of new techniques for using news data along with trends in stock prices to predict future stock prices and improve the accuracy of stock price prediction models.

II. Related Work

A major financial offense that has the potential to seriously affect stock markets and individual investors is illegal insider trading. Because the data is frequently complicated and challenging to evaluate, traditional methods for detecting insider trading have proved ineffective. In recent years, deep learning algorithms have been applied to create more precise and effective insider trading detection and prediction methods. This literature review gives a summary of current research in this field.

Many studies have been done on the application of deep learning methods for identifying and forecasting insider trading. For instance, Wang et al. (2019) created a model to analyze financial news items and forecast insider trading using a convolutional neural network (CNN). According to their findings, the algorithm was 89.3% accurate in identifying insider trading.

A long short-term memory (LSTM) network is used to evaluate social media data in another study by Xiong et al. (2020) that presented a deep learning-based insider trading detection approach. The precision and recall of the suggested approach were 84.6% and 87.9%, respectively.

Similar to this, Li et al. (2020) created a deep learning-based framework for insider trading prediction that incorporates a variety of data sources, such as stock prices, social media, and financial news. Their findings show that the suggested framework performs better than established techniques for insider trading prediction.

Together with the studies mentioned above, several other research papers have suggested other deep learning-based strategies for identifying and forecasting insider trading. As an illustration, Zhang et al. (2019) examined financial statements and found suspicious trading activity using a deep belief network (DBN). A graph convolutional network (GCN) is used in a model created by Liu et al. (2021) to evaluate stock trading networks and spot insider trading.

In conclusion, deep learning approaches have demonstrated significant promise for identifying and forecasting unauthorized insider trading in the stock market. The experiments analyzed in this literature review show that various deep learning models can be applied to the analysis of news articles, social media, financial statements, and stock trading networks, among other types of financial data. The efficiency and accuracy of insider trading detection and prediction algorithms based on deep learning still need to be improved, though.

III. Our Solution

The problem statement and solution for our project can be defined as: “To predict the future stock prices of a set of 82 companies using input features of sentiment based on news data along with the prior stock price for the respective company.”

A. Description of Dataset

Dataset is created using the data from the following two websites:

1) News API (https://newsapi.org/)
2) Yahoo finance

This dataset contains stock market data for a specific ticker symbol, acquired from Yahoo Finance. It covers the period from 2001 to February 2023, and includes daily records of stock prices and trading volume. The dataset is stored in a pickle (.pkl) format, which can be accessed using Python’s pandas library.

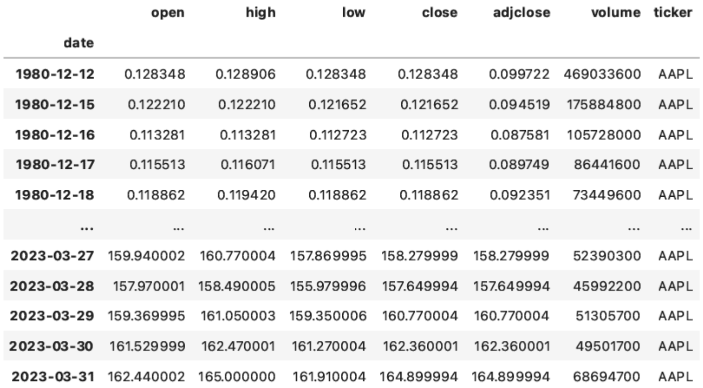
The news data used for sentiment analysis, only available for the last 30 days, is collected from a free news API, and includes articles related to the company associated with the ticker symbol. This data was stored as a csv containing the title of the news article, the source, and the text content of the news article. This was done for every company in the list and for as many days in the period of 30 days, when some news was available.

The dataset contains the following variables:

1) open - the opening price of the stock on a given day
2) high - the highest price of the stock on a given day
3) low - the lowest price of the stock on a given day
4) close - the closing price of the stock on a given day
5) adjclose - the adjusted closing price of the stock on a given day
6) volume - the trading volume of the stock on a given day
7) ticker - the ticker symbol of the stock

B. Machine Learning Algorithms

We have used a two-step approach to work on the problem statement of predicting stock prices. We have used NLP for analyzing News articles pertaining to the companies shortlisted and implemented different machine-learning algorithms for predicting stock prices. In the end, we have compared the performance results of the respective models and suggested future methods in which, the stock prediction problem can be addressed using different datasets.



	open	high	low	close	adjclose	volume	ticker
date							
1980-12-12	0.128348	0.128906	0.128348	0.128348	0.099722	469033600	AAPL
1980-12-15	0.122210	0.122210	0.121652	0.121652	0.094519	175884800	AAPL
1980-12-16	0.113281	0.113281	0.112723	0.112723	0.087581	105728000	AAPL
1980-12-17	0.115513	0.116071	0.115513	0.115513	0.089749	86441600	AAPL
1980-12-18	0.118862	0.119420	0.118862	0.118862	0.092351	73449600	AAPL
...	...	...	...	...	...	...	...
2023-03-27	159.940002	160.770004	157.869995	158.279999	158.279999	52390300	AAPL
2023-03-28	157.970001	158.490005	155.979996	157.649994	157.649994	45992200	AAPL
2023-03-29	159.369995	161.050003	159.350006	160.770004	160.770004	51305700	AAPL
2023-03-30	161.529999	162.470001	161.270004	162.360001	162.360001	49501700	AAPL
2023-03-31	162.440002	165.000000	161.910004	164.899994	164.899994	68694700	AAPL

Fig. 1. Sample Dataset of Stock Prices - Model 2, Model 3

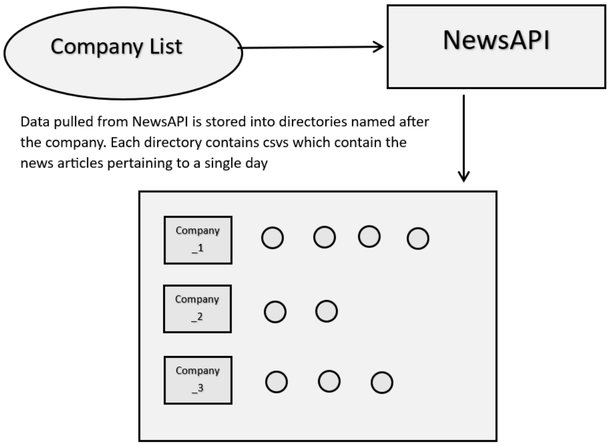

Company List
NewsAPI
Data pulled from NewsAPI is stored into directories named after the company. Each directory contains csvs which contain the news articles pertaining to a single day
Company _1
Company _2
Company _3

Fig. 2. Dataset Retrieval and Storage Architecture for News Data - Model 1

C. Implementation Details

1) Model 1 - NLP based Sentiment Analyzer: Author 1 works on analyzing News articles pertaining to the respective company for understanding the sentiment of the article. This is to determine what effect the article has on the price of the stock after the article was released. Using a certain set of companies (which is common to all three members), the author gathers the news articles for the last 30 days and uses this to predict the sentiment of the news for the particular company for the particular day. The author also worked on exploring other algorithms to draw better inferences from the data and how this can be combined with the stock price to better aid prediction.

After collecting the dataset, as mentioned in the subsection above, the news articles were analyzed for one such company - Apple, ticker - AAPL. Within the 30 days’ of data, the author implements data processing techniques to clean the available data. NLTK Sentiment Analyzer is used on the concatenated input which provides the sentiment score for the particular article. This is done for every article and the scores are aggregated per day. The mean score determines how positive or negative the news article is perceived.

The obtained sentiment scores were concatenated with the stock price data from yahoo finance for the said period. This dataset was used to further correlate and analyze the plots of both the sentiment score as well as the stock closing price as per the date. The said plots for Apple, Amazon and Bank of America are depicted in Fig.5, Fig.4 and Fig.6 for reference.

This sentiment is used in two ways. One, it is used to correlate with the stock price and understand if the observed sentiment is matching with that of the difference in the stock price. Second, it is used in addition to the stock price data to train and test a clustering model, which tells us if the stock price have any effect due to the sentiment based on the obtained clusters. All the experiments conducted are on a smaller scale due to the unavailability of a larger corpus of news data. The same can be implemented on different companies data to compare and understand their influence.

Natural Language Processing (NLP) tools used were NLTK (Natural Language Toolkit) and VADER (Valence Aware Dictionary and sEntiment Reasoner) Sentiment Analysis Tool. VADER is a sentiment analysis tool designed for analyzing emotions in text, particularly in social media. It uses rules to determine sentiment polarity and computes sentiment scores based on word context and order. Scores range from -1 to 1, representing negative to positive sentiment. VADER's simplicity and effectiveness make it popular for sentiment analysis in short, informal texts like social media posts.

The figure Fig.3 and Fig.2 shows the structure of the dataset and the process of creation of the dataset. The news data pulled from NewsAPI is stored in folders pertaining to the different companies considered. Every folder contains multiple csvs with respect to the date. Each csv contains news headline, news body and the news source.

To further understand this in-depth, a K-Means Clustering Model was fitted to the dataset. The dataset obtained above was further added with columns pertaining to difference in opening and closing price, difference in opening price and the low price for the day, the difference between the opening price and the high price of the day and also the difference betweeen the high and the low price of the day, depicted as Volatility. These features were then added to the exisiting datatset with the sentiment scores to perform the K-Means Clustering.

The figures show the clusters for different combinations of:

1) Volatility vs Mean Sentiment Score Fig.7
2) Maximum Peak Difference vs Mean Sentiment Score Fig.9
3) Maximum Dip Difference vs Mean Sentiment Score Fig.8
4) Difference between Opening and Closing Prices vs Mean Sentiment Score Fig.11
5) Normalized Trade Volume vs Mean Sentiment Score Fig.10

The feature of volume was normalized using a MinMaxScaler before being used in the dataset.

We can infer the following from this clustering model:

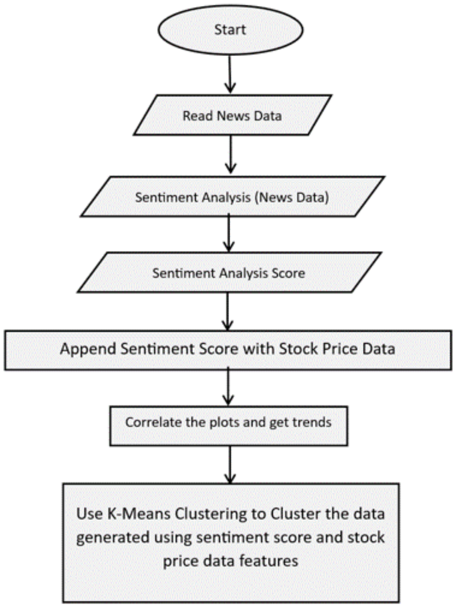

Start
Read News Data
Sentiment Analysis (News Data)
Sentiment Analysis Score
Append Sentiment Score with Stock Price Data
Correlate the plots and get trends
Use K-Means Clustering to Cluster the data generated using sentiment score and stock price data features

Fig. 3. Process Flow Chart - Model 1

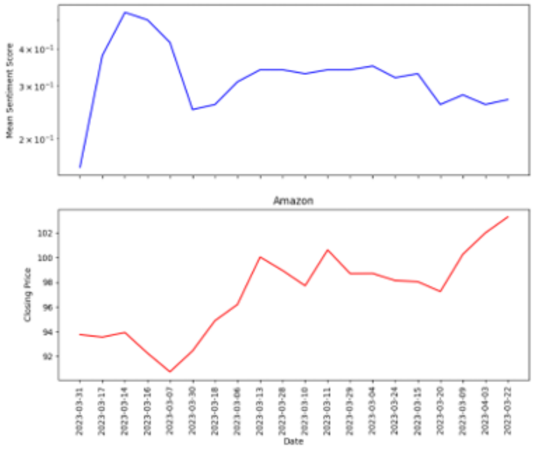


Fig. 4. Amazon Stock Price and Mean Sentiment Score - Model 1

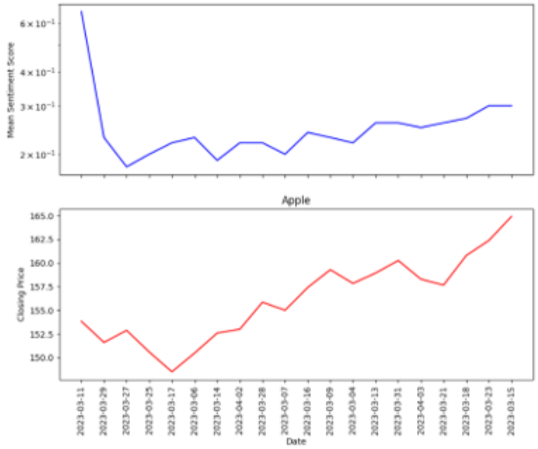


Fig. 5. Apple Stock Price and Mean Sentiment Score - Model 1

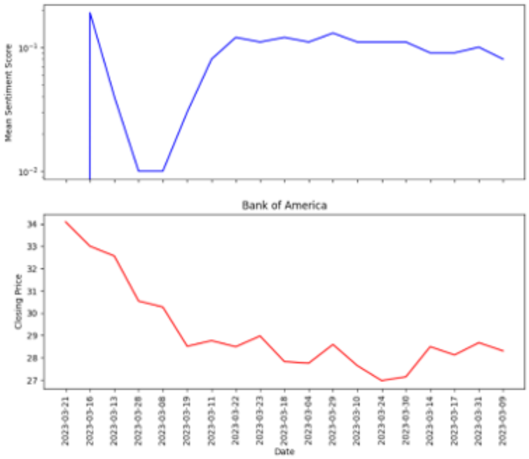


Fig. 6. Bank of America Stock Price and Mean Sentiment Score - Model 1

1) Volatility vs Mean Sentiment Score shows that there is a clear cluster of blue data points which do not show a highly volatile market with a neutral sentiment score. And also that another cluster of orange data points is present which can be said to be highly influenced by the sentiment score and has highly volatile market.
2) Maximum Peak Difference vs Mean Sentiment Score shows that most of the clusters here are concentrated around a neutral to positive sentiment score. Wherein there is one particular cluster, orange, which again shows highly varying maximum stock price for the day, given the sentiment of the news.
3) Maximum Dip Difference vs Mean Sentiment Score shows a similar trend in clustering as with the previous one. Here too, the orange cluster shows reactive nature to the sentiment score.
4) Difference between Opening and Closing Prices vs Mean Sentiment Score has all the clusters concentrated around 0 difference and neutral to positive sentiment score. Here again, the blue cluster seems to not be affected by the sentiment score, whereas the orange, green and red clusters behave almost similarly.
5) Normalized Trade Volume vs Mean Sentiment Score is the most different out of the other four trends so far observed. Here, we can see that despite a neutral sentiment score, the blue cluster has a significantly higher trade volume as compared to the others. And again, the green cluster is divided into two regions, one showing no reaction in trade volume to the positive sentiment, whereas the other shows somewhat of a meagre movement of trades.

Correlation analysis helps understand the impact of news sentiment on intra-day and next-day stock prices. The use of K-Means Clustering model helps us to deep dive into the inferences due to the inclusion of a sentiment score, or the effect of news on the value of stock prices. This helps to inherently understand which cluster of company or what features show the maximum effect on the value of a stock.

Conclusion: Analyzing news sentiment can provide insights

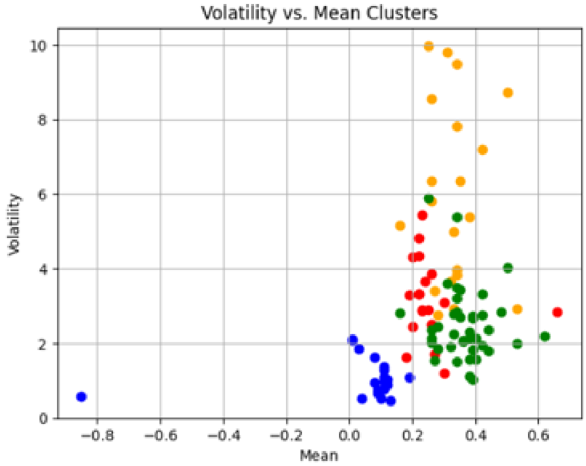


Fig. 7. Volatility vs Mean Sentiment Score - Model 1

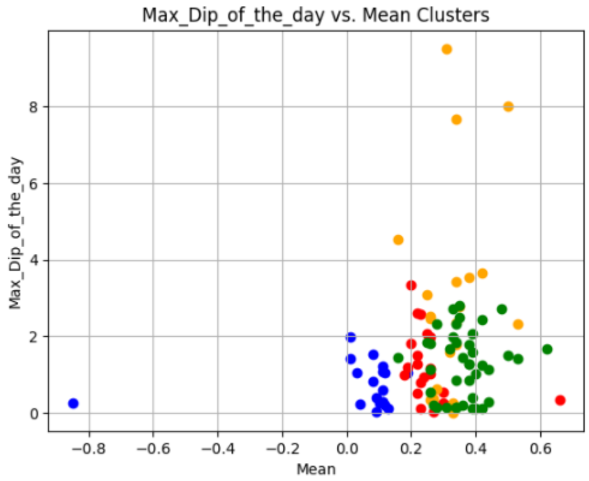


Fig. 8. Maximum Dip Difference vs Mean Sentiment Score - Model 1

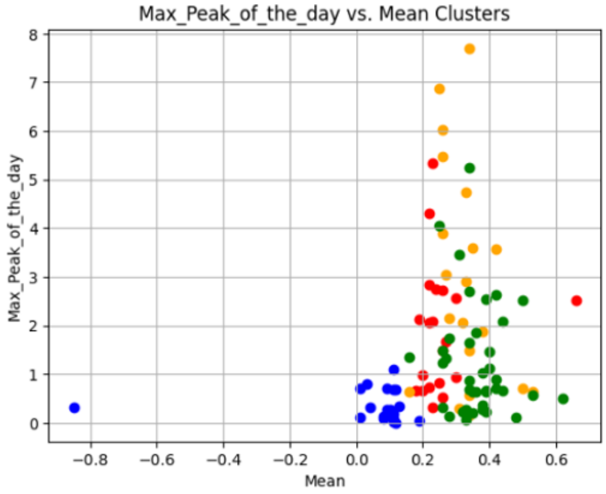


Fig. 9. Maximum Peak Difference vs Mean Sentiment Score - Model 1

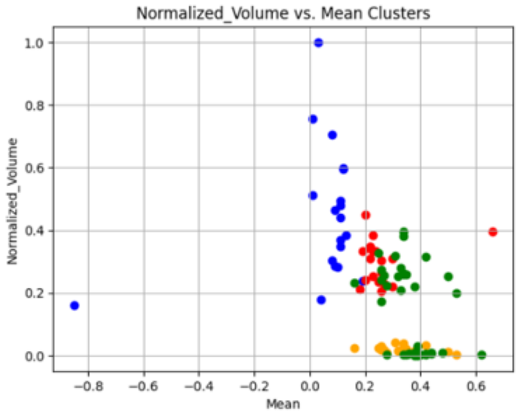


Fig. 10. Normalized Trade Volume vs Mean Sentiment Score - Model 1

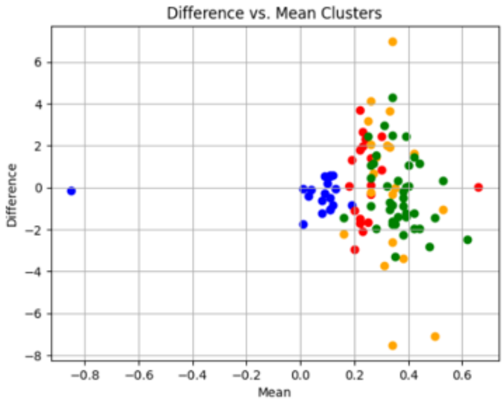


Fig. 11. Difference between Opening and Closing Prices vs Mean Sentiment Score - Model 1

into the effect of external factors on stock prices This understanding can help investors make informed decisions based on the sentiment reflected in the news

2) Model 2 - LSTM for Stock Price Prediction: Author 2 explores the use of LSTM for predicting future stock prices based on historical data. He has compared the performance of four different machine learning models, namely Long Short-Term Memory (LSTM) Networks, Gated Recurrent Units (GRU), Convolutional Neural Networks (CNN), and Random Forest, in predicting stock prices using time series data

LSTM - Long Short-Term Memory Networks is a type of recurrent neural network (RNN). It captures long-term dependencies in time series data so well-suited for stock prediction tasks involving sequential data. The best LSTM configuration had the following parameters:

1) Best LSTM RMSE: 0.02322727933710564
2) Best Sequence Length: 15
3) Best Epochs: 200

GRU - Gated Recurrent Units is another type of RNN and requires fewer parameters than LSTM networks. I captures temporal dependencies which make them suitable for predicting stock prices. The best GRU configuration had the following parameters:

1) Best GRU RMSE: 0.02308415804854275

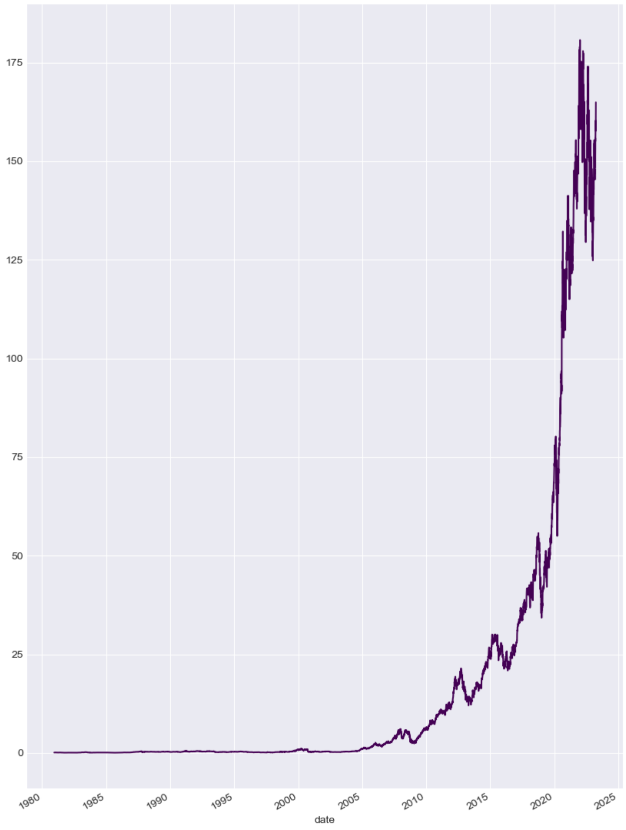


Fig. 12. AAPL stock data - Model 2

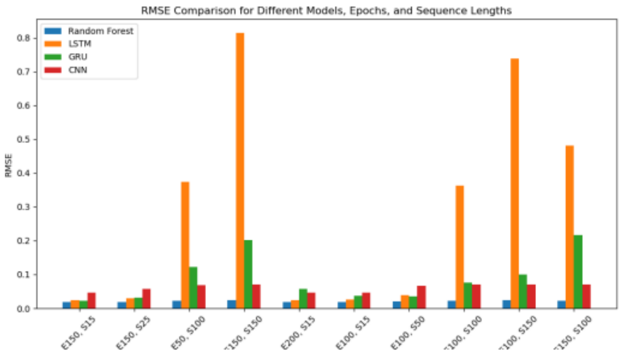


Fig. 13. RMSE Comparison for Different Models, Epochs, and Sequence Lengths - Model 2

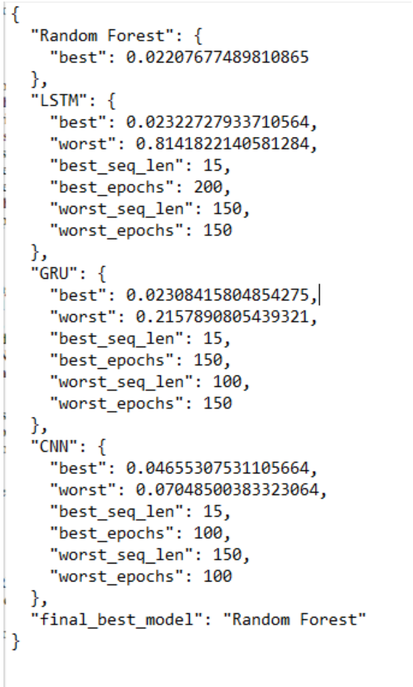
2) Best Sequence Length: 15
3) Best Epochs: 150

CNN - Convolutional Neural Networks excel at detecting local patterns in data. It can be adapted for time series analysis by treating temporal sequences as 1D data. It is a good choice for capturing short-term dependencies in stock data. The best CNN configuration had the following parameters:

1) Best LSTM RMSE: 0.04681689789910629
2) Best Sequence Length: 15
3) Best Epochs: 150

Random Forest - It is a powerful ensemble method that constructs multiple decision trees. It aggregates their results, providing robust and stable predictions for stock prices. It is good when dealing with noisy or missing data. The Random Forest scores:

1) Best RMSE: 0.01920465563497243
2) Best MAE: 0.11809418378518596



```
{
  "Random Forest": {
    "best": 0.02207677489810865
  },
  "LSTM": {
    "best": 0.02322727933710564,
    "worst": 0.8141822140581284,
    "best_seq_len": 15,
    "best_epochs": 200,
    "worst_seq_len": 150,
    "worst_epochs": 150
  },
  "GRU": {
    "best": 0.02308415804854275,
    "worst": 0.2157890805439321,
    "best_seq_len": 15,
    "best_epochs": 150,
    "worst_seq_len": 100,
    "worst_epochs": 150
  },
  "CNN": {
    "best": 0.04655307531105664,
    "worst": 0.07048500383323064,
    "best_seq_len": 15,
    "best_epochs": 100,
    "worst_seq_len": 150,
    "worst_epochs": 100
  },
  "final_best_model": "Random Forest"
}
```

Fig. 14. Best and Worst RMSE Final Model - Model 2

Best Model: Random Forest Best RMSE: 0.01920465563497243 Best MAE: 0.11809418378518596

Based on the results obtained, the Random Forest model outperformed the other models with the lowest RMSE of 0.01920465563497243 and MAE of 0.11809418378518596, indicating its robustness and stability when dealing with noisy or missing data.

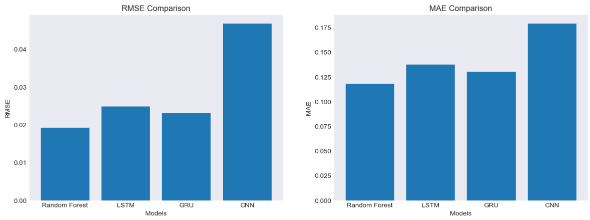


Fig. 15. All model error metric comparison - Model 2

GRU came in second, with an RMSE of 0.02308415804854275, followed by LSTM with an RMSE of 0.02322727933710564.

Both LSTM and GRU models demonstrated their ability to capture long-term dependencies in time series data.

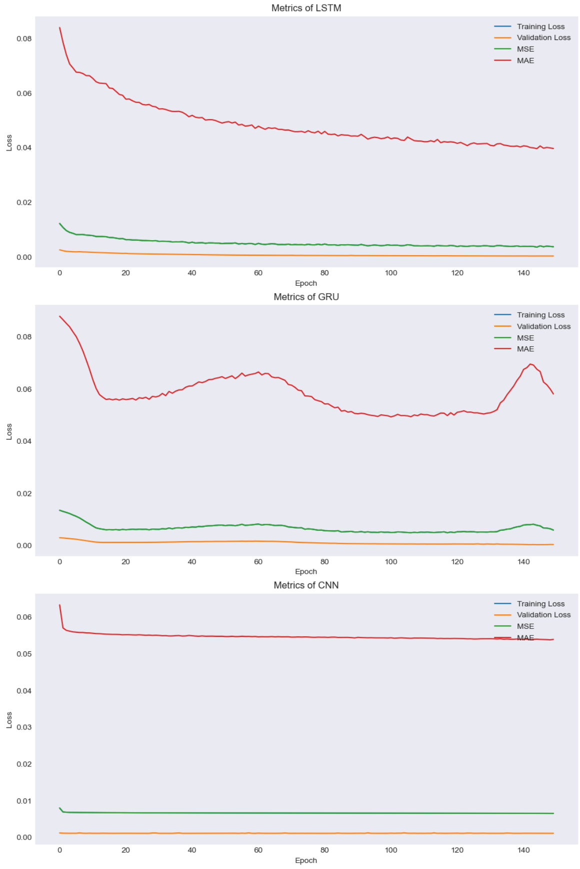


Fig. 16. All model loss curves - Model 2

CNN had the highest RMSE of 0.04681689789910629, highlighting its suitability for capturing short-term dependencies in stock data.

The results can be summed up as in Fig.14

In conclusion, the Random Forest model was found to be the best model for stock price prediction among the four models compared in this study. The output charts are shown in the figures.

3) Model 3 - Machine Learning Algorithms: Author 3 compares three machine learning models -

1) Logistic Regression
2) Support Vector Machines (SVM)
3) Long Short-Term Memory Networks (LSTM)

Models are chosen for their ability to :

1) Perform classification and regression tasks (Logistic Regression, SVM),
2) Data preprocessing and feature engineering
3) Model training and parameter tuning
4) Model evaluation and selection
5) Parameter tuning performed manually using performance metrics (MSE, R2) to minimize computational cost

Logistic Regression/SVM: Simple and interpretable models and effective in predicting stock prices LSTM: Most accurate model, captures complex patterns in data, low loss, indicating well-suited model for the task

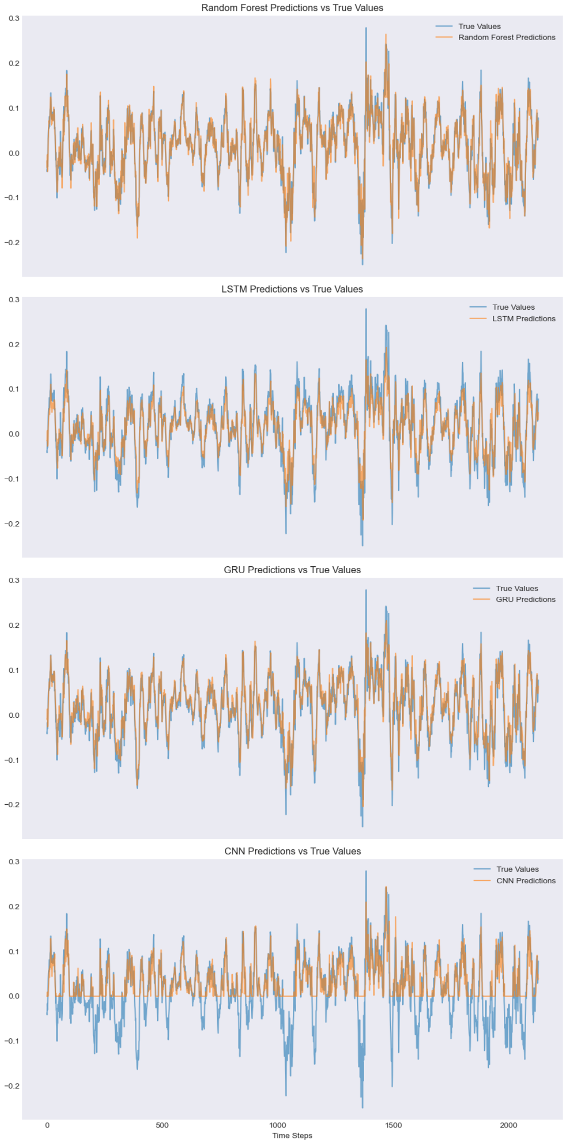


Fig. 17. All model predictions vs true values - Model 2

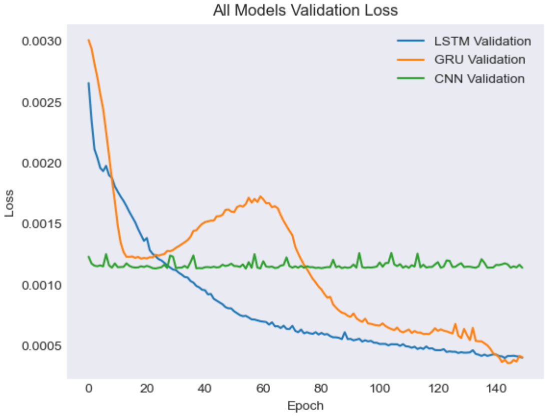


Fig. 18. All model validation loss curve - Model 2

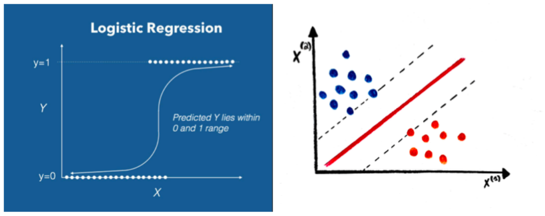


Fig. 19. Logistic Regression - Model 3

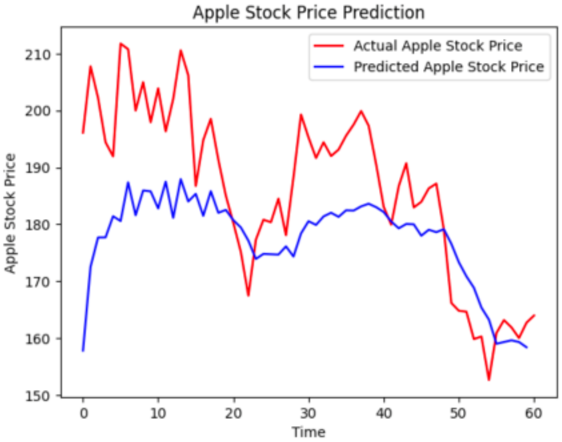


Fig. 20. Apple Stock Price Prediction - Model 3

Conclusion: Both Logistic Regression/SVM and LSTM are effective in predicting stock prices LSTM is particularly well-suited for time-series data analysis, demonstrating the potential of machine learning in the financial industry

IV. Comparison

After the individual models were tested for stock price prediction using historical stock data, it was observed that Random Forest performs better with the said Apple Stock Data. It had the best performance out of the others and outperformed the other models with the lowest RMSE of 0.01920465563497243 and MAE of 0.11809418378518596, indicating its robustness and stability when dealing with noisy or missing data.

With respect to Sentiment Analysis, we found that the inclusion of sentiment scores can significantly help in the future. Stock price prediction could involve sentiment scores of news or other social media data to provide a different feature and another angle of insight into improving the prediction process. The inferred correlations from the K-Means clustering model shows how different features are affected differently by the presence of a sentiment score.

Possible improvements to the algorithms are suggested as follows:

1) Incorporating more sources of data, such as social media sentiment
2) Exploring other machine learning and deep learning techniques, such as transformers or reinforcement learning

3) Enhancing feature engineering and selection processes
4) Combining different models for an ensemble approach

V. Future Directions

Integrating Sentiment Score and Adj Close for Prediction Combining sentiment scores and adjusted closing prices for enhanced stock price prediction Develop a model that takes the following inputs: Sentiment scores for the past 15 days Adjusted closing prices for the past 15 days Predict the adjusted closing price for the next day based on these inputs Such a model could provide a more comprehensive understanding of the factors affecting stock prices, including both historical price data and news sentiment This approach may result in improved prediction accuracy and better insights for investors making informed decisions

VI. Conclusion

Our project successfully demonstrated the power of machine learning and natural language processing in predicting stock prices and analyzing news sentiment We found that Random Forest and GRU are effective models for stock price prediction Analyzing news sentiment can provide insights into the effect of external factors on stock prices, helping investors make informed decisions Future research directions can further improve our models and broaden their applicability to the financial industry

VII. References

Wang, J., Wu, J., & Wang, Y. (2019). A Deep Learning-based Insider Trading Detection Approach Using Financial News. In Proceedings of the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 2019, pp. 969-974. doi: 10.1109ITNEC.2019.8721053.

Xiong, Y., Zhu, Q., & Zeng, Y. (2020). A Deep Learning-based Insider Trading Detection Approach Using LSTM Network. In Proceedings of the 2020 IEEE 2nd Conference on Advances in Artificial Intelligence (ICAAI), Harbin, China, 2020, pp. 272-276. doi: 10.1109/ICAAI50182.2020.00058.

Li, Z., Li, H., & Zhao, L. (2020). Deep Learning-based Insider Trading Prediction with Multiple Data Sources. In Proceedings of the 2020 IEEE International Conference on Information and Automation (ICIA), Shenyang, China, 2020, pp. 1027-1032. doi: 10.1109/ICInfA51470.2020.9281301.

Zhang, Q., Zhang, Y.,& Wang, X. (2019). A Deep Belief Network-based Insider Trading Detection Method Using Financial Statements. In Proceedings of the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 2019, pp. 975-979. doi: 10.1109/ITNEC.2019.8721042.

Liu, K., Yin, X., & He, Y. (2021). Insider Trading Detection in Stock Trading Networks using Graph Convolutional Networks. In Proceedings of the 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA), Chengdu, China, 2021, pp. 212-216. doi: 10.1109/ICBDA51650.2021.00051.