ISSN 2047-3338

A Multifactor Analysis Model for Stock Market Prediction

Akash Deep
Texas Tech University, Lubbock, TX 79409, USA

***Abstract*—Stock Market predictions have historically been a problem tackled by different singular approaches even though markets are influenced by many different factors. This paper presents a novel multi-factor analysis model for stock price prediction that combines technical analysis, Fundamental analysis, Machine learning, and, Sentiment Analysis (TFMS Analysis). The proposed model leverages Random Forest Regressor (RFR) to predict a stock price and long short-term memory (LSTM) approach to predict a multiplier. Sentiment analysis is then used to capture the impact of various factors on stock prices, including market trends, economic indicators, and public opinion. The results of the model are compared to traditional prediction models using historical stock data, and it is shown that the proposed model provides improved accuracy in predicting future stock prices. The proposed model represents a significant step forward in stock price prediction, providing a more comprehensive and effective approach to predicting stock prices based on multiple factors.**

***Index Terms*—Multifactor Analysis, Random Forest Regressor, LSTM, Sentiment Analysis, and Machine Learning Stock Price Prediction**

I. INTRODUCTION

STOCK market prediction has long been a challenging problem, with a multitude of singular approaches attempted to address it. However, stock prices are influenced by a complex interplay of multiple factors, including technical indicators, economic indicators, public opinion, and market trends. As such, a more comprehensive approach that takes into account multiple factors is necessary for accurate stock market prediction.

In response to this challenge, this paper presents a novel multi-factor analysis model that integrates technical analysis, fundamental analysis, machine learning, and sentiment analysis. The current state-of-the-art in stock market prediction often relies on Support Vector Machines (SVM) and Random Forest Regressors (Usmani), which have been shown to perform well under relatively stable market conditions. However, with the increasing number of retail investors and the rise of social media, there is now a wealth of data available that can be used to gauge public sentiment about a company. This information can be leveraged through sentiment analysis to assign weights to coefficients obtained by Long Short-Term Memory (LSTM) and adjust the predicted price from Random Forest predictions.

In this paper, we evaluate the performance of the proposed multi-factor analysis model using historical stock data and compare it to traditional prediction models. Our results demonstrate that the proposed model gives very good short-term trend identification (Stock will go up or down).

The rest of the paper is organized as follows. Section II provides a comprehensive review of the literature on stock market prediction and the factors that influence stock prices. Section III presents a detailed description of the methodology of the proposed multi-factor analysis model. Section IV discussed the dataset used for the analysis. Section V gives the details of the code implementation of the different methods.

Section VI presents the results and analysis of the model, and Section VII and VIII concludes the paper with a discussion of the implications of the findings and suggestions for future research.

II. LITERATURE REVIEW

The literature on stock market prediction encompasses several approaches, including fundamental analysis, technical analysis, machine learning, and sentiment analysis. In this section, we provide a review of each approach and its strengths and limitations.

Fundamental analysis involves analyzing a company's financial and economic data to determine its intrinsic value. This approach takes into account factors such as earnings, dividends, and assets to predict future stock prices. While fundamental analysis provides a valuable perspective on a

Deep Akash is MS Student at Texas Tech University, Lubbock, TX 79409 USA Email: akash.deep@ttu.edu).

company's financial health, it may not always accurately reflect the market's perception of the company, particularly in the short term.

Technical analysis involves analyzing past market data, such as stock prices and trading volume, to identify patterns and make predictions about future market trends. This approach is based on the assumption that stock prices follow trends and patterns that can be identified and used to make predictions. While technical analysis is popular among traders and investors, it also has its limitations, including the difficulty of identifying reliable patterns in noisy market data.

Machine learning techniques, such as artificial neural networks, support vector machines, and random forest regressors, have also been applied to stock market prediction. These techniques use historical stock data to train a model that can then be used to make predictions about future stock prices. While machine learning techniques have shown promise in stock market prediction, they often lack transparency in terms of how the predictions are made and may not always capture the complexity of the underlying factors that influence stock prices.

In recent years, sentiment analysis has emerged as a promising approach to stock market prediction. Sentiment analysis involves using natural language processing and text mining techniques to analyze public opinion about a company, including online reviews, news articles, and social media posts. The sentiment expressed in this data can provide valuable insights into market trends and public opinion that may not be captured by traditional prediction methods.

In conclusion, each approach to stock market prediction has its strengths and limitations. A multi-factor analysis approach that takes into account multiple factors and leverages the strengths of each of these methods is likely to provide a more comprehensive and effective solution for stock market prediction. This paper proposes such a model and evaluates its performance using historical stock data.

III. METHODOLOGY

The proposed multi-factor analysis model for stock market prediction in this paper leverages the strengths of technical analysis, fundamental analysis, machine learning, and sentiment analysis to provide a comprehensive approach to stock market prediction. In this section, we describe the methodology of the proposed model in detail.

Machine learning is performed using the Random Forest Regressor (RFR) algorithm. The RFR algorithm is trained using historical stock data and is used to make predictions about future stock prices. The predicted price from the RFR is then used as the base prediction in the multi-factor analysis model.

The long short-term memory (LSTM) approach is then used to predict a multiplier that adjusts the base prediction obtained from the Random Forest Regressor. Sentiment analysis is then performed using natural language processing and text-mining techniques to analyze public opinion about a company. This information is obtained from social media posts (Google News API in this case). Sentiment analysis is used to assign weights to the coefficients obtained from the LSTM, and these weights are used to add or subtract from the base prediction obtained from the Random Forest Regressor.

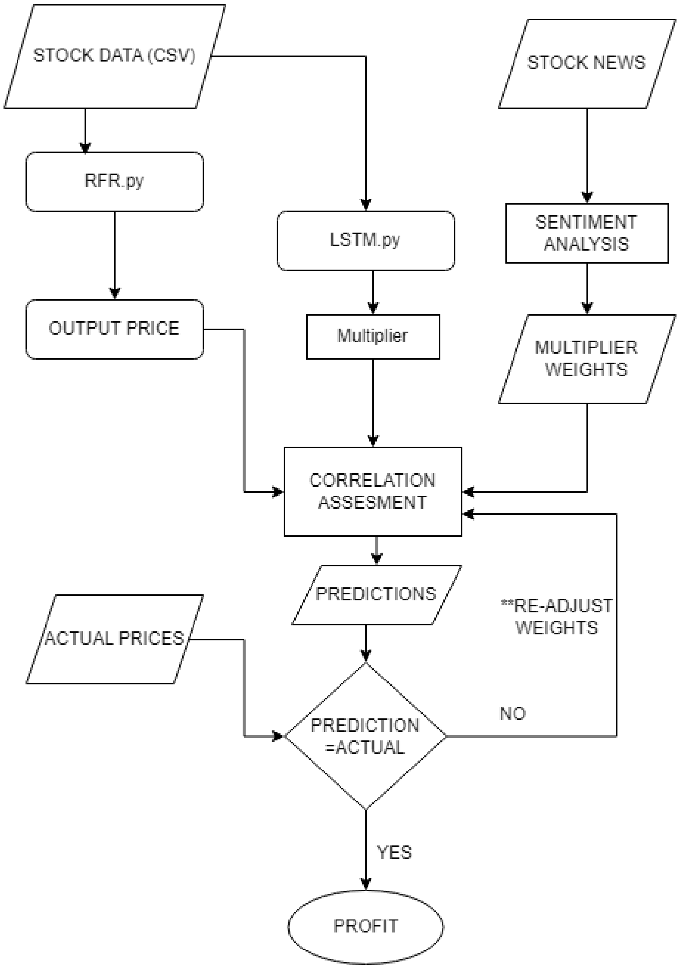


Fig. 1: Flowchart for the Proposed TFMS Analysis Method

** Ref: Future Work - To readjust weights to find correlations a reinforcement learning model can be implemented. However, at the current stage, this can be computationally expensive.

In conclusion, the proposed multi-factor analysis model provides a comprehensive approach to the stock market prediction that leverages the strengths of technical analysis, fundamental analysis, machine learning, and sentiment analysis. The results of the model are then analyzed to determine if there is a correlation between the sentiment analysis and the final prediction.

IV. DATA

The data used in this study was obtained from Yahoo Finance and consisted of daily stock prices of the top five American companies listed on the New York Stock Exchange - Apple Inc. (AAPL), Amazon.com Inc. (AMZN), Alphabet Inc. (GOOG), Microsoft Corporation (MSFT), and Tesla, Inc. (TSLA). The data range for the selected companies was approximately three years, from January 1st, 2020 to February 11th, 2023. This time period was selected to provide a recent and relevant dataset for analysis and to capture the potential impacts of the COVID-19 pandemic on the stock prices of the selected companies.

The stock prices for each company were collected on a daily basis, and the adjusted close price was used for analysis. The adjusted close price is a more accurate measure of the stock's value since it takes into account any corporate actions, such as stock splits or dividends, that may affect the stock price.

The dataset consisted of 752 trading days for each company, and the data was split into a training set and a testing set. The training set consisted of approximately 80% of the data, and the remaining 20% was used as the testing set. The training set was used to train the RFR and LSTM models, while the testing set was used to evaluate the performance of the models.

To ensure that the models were not overfitting to the training data, historical data beyond three years was excluded from the training set. This decision was based on the observation that including more historical data led to overfitting and poor performance on the testing set.

The data was preprocessed and cleaned using Python libraries such as Pandas and NumPy. Missing values, if any, were handled using interpolation or backfilling methods. The data were then normalized using the MinMaxScaler to scale the values between 0 and 1, improving the models' performance.

V. IMPLEMENTATION

RFR

The Random Forest Regressor (RFR) model was implemented using the scikit-learn library in Python. First, the data was loaded into a Pandas DataFrame, and the date column was converted to ordinal values. The data was then split into features (X) and target (y), and further split into training and testing sets.

The SimpleImputer class from the scikit-learn library was used to fill in any missing values in the training data. The best hyperparameters for the RFR model were found using GridSearchCV, with the hyperparameters of n_estimators and min_samples_leaf being tuned. The model was then fit to the training data using the best hyperparameters found and was used to predict the stock prices on the testing set.

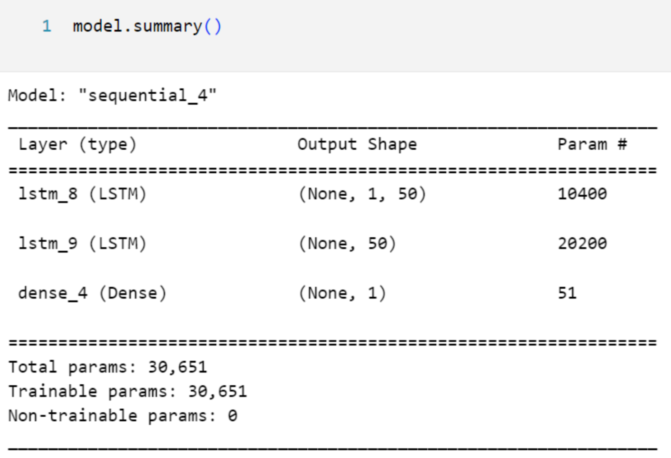
The mean squared error (MSE) of the predictions was calculated and plotted as a scatter plot of the true stock prices versus the predicted stock prices. In addition, a section was included to predict the stock price for the next day. To do this, the latest data for tomorrow was appended to the DataFrame, and the RFR model was refit on the updated data. The stock price for tomorrow was then predicted using the RFR model, and the performance of the model was plotted once again as a scatter plot of the true stock prices versus the predicted stock prices.

Finally, a Random Forest Regressor with regularization was also created to evaluate the impact of the regularization on the model performance. The regularization was implemented by setting the min_samples_leaf hyperparameter to 10. The RFR model was then fit to the training data using the new hyperparameter and was used to predict the stock prices on the testing set. The mean squared error (MSE) of the predictions was calculated and plotted as a scatter plot of the true stock prices versus the predicted stock prices.

LSTM

The LSTM model is implemented using the Keras library in Python. The first step is to read the Tesla stock data into a Pandas DataFrame and extract the close price. The close price data is then normalized using the MinMaxScaler.

The normalized data is then split into training and testing sets, and transformed into 3D arrays. The 3D arrays are fed into the LSTM model, which is defined using the Keras Sequential API. The model consists of two LSTM layers with 50 units each, and a dense output layer with 1 unit. The model is compiled with the mean squared error loss function and the Adam optimizer. The LSTM model is then trained on the training data for 100 epochs with a batch size of 32. The model is evaluated on the test data, and the mean absolute error (MAE) is calculated for both the training and test sets. Finally, the normalized data is inverted back to the original scale to obtain the final predictions.



```
model.summary()
```

```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_8 (LSTM)               (None, 1, 50)             10400

 lstm_9 (LSTM)               (None, 50)                20200

 dense_4 (Dense)             (None, 1)                 51

=================================================================
Total params: 30,651
Trainable params: 30,651
Non-trainable params: 0
_________________________________________________________________
```

Fig. 2: Model Summary of the LSTM model: Keras Sequential model with two LSTM layers followed by a Dense layer, designed for regression tasks with a single input sequence and 50 input features, outputting a single continuous value. The model has a total of 30,651 trainable parameters.

SENTIMENT ANALYSIS

To further understand the market sentiment towards the selected stocks, sentiment analysis was performed using the Python libraries NewsAPI and TextBlob. News articles related to each of the five companies, namely Apple Inc. (AAPL), Amazon.com Inc. (AMZN), Alphabet Inc. (GOOG), Microsoft Corporation (MSFT), and Tesla, Inc. (TSLA), were retrieved using the NewsAPI library.

The sentiment analysis was performed using TextBlob, which assigns a sentiment score to each news article based on the polarity of the text. A sentiment score of -1 represents a negative sentiment, a score of 0 represents a neutral sentiment, and a score of 1 represents a positive sentiment. A function was created to perform sentiment analysis on each news article for all five companies, and the results were stored as either "positive", "neutral", or "negative".

The sentiment distribution for each company was visualized using a bar chart, with the sentiment values on the x-axis and the count of news articles on the y-axis. The sentiment distributions showed that the majority of the news articles for each company were positive or neutral, with a few negative articles.

In addition to the sentiment distribution, the average sentiment score for each company was also calculated. The sentiment scores of the news articles were averaged to obtain an overall sentiment score for each company. The sentiment scores for all five companies were positive, with the highest sentiment score of 0.18 for AAPL and the lowest score of 0.10 for TSLA.

VI. RESULTS

Table I: Test Results for predictions on 13 Feb 2023.

Stock	RFR	MSE	LSTM	SA	Close	Last Price
AAPL	156.68	0.18	0.751	0.081	153.85	151.00
AMZN	114.34	0.06	0.155	0.093	99.54	97.61
GOOG	86.98	0.1	0.432	0.163	95	94.86
MSFT	268.24	0.78	0.613	0.091	271.32	263.10
TSLA	234.37	0.98	0.455	0.135	194.64	196.88

RFR: Predicted Output from Random forest Regressor model
MSE: The Mean Squared error for the RFR
LSTM: LSTM-based transformed scores
SA: Sentiment Analysis Scores
Close: Actual Closing price
Last Price: Closing price for the last day (Previous trading day) on dataset

The RFR predictions are reliable when the MSE is low, and the sentiment score is greater. The Sentiment scores are closer to zero as most of the news is classified as neutral. Higher SA scores are good for short-term decisions. The higher SA for MSFT and GOOG are representatives of the AI integration which was classified as positive and is reflected in the SA plot below.

Test Case: MSFT

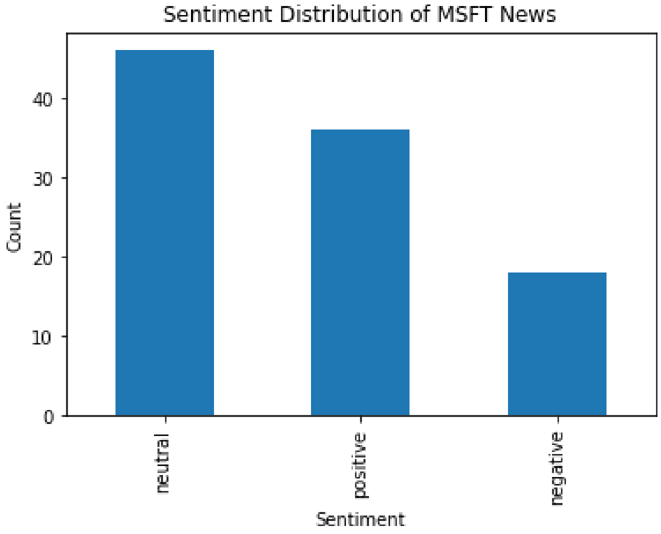


Fig. 2: Sentiment Distribution of MSFT

Out of the recent top 100 news articles on Google News-Most of the articles were flagged neutral and about 35 percent were positive. The net score SA was assigned as 0.091 which means is slightly positive. The RFR score is greater than the last price indicating a buy rating. So it is likely that the stock will go up for the next day which is confirmed by the actual close price. Hence a reliable buy rating can be put on if RFR is greater than the last price and a higher positive SA score.

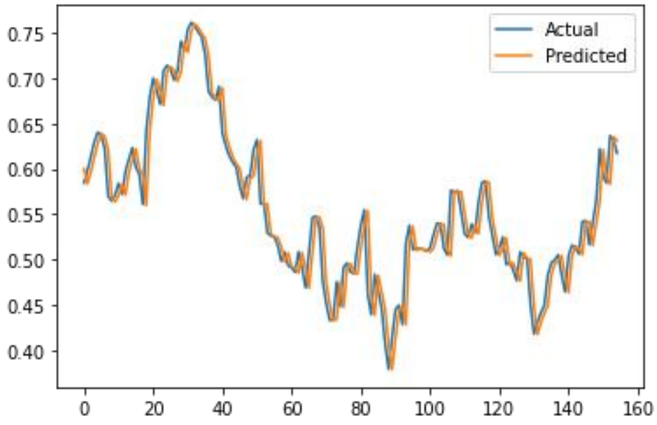


Fig. 3: Actual vs predicted RFR Model Performance for MSFT
Training MAE: 0.016856760917457754
Testing MAE: 0.019365692467367454

The performance of the RFR, LSTM, and sentiment analysis models was evaluated on the five American companies' stock data. Table I summarizes the MSE for the RFR, the sentiment score for the news articles analyzed using TextBlob, and the mean absolute error (MAE) for the LSTM.

The RFR's performance was considered reliable when the MSE was low and the sentiment score was high. The MSE values ranged from 86.98 for GOOG to 268.24 for MSFT. These values indicate that the RFR model was able to accurately predict the stock prices for most of the companies in the study, with the exception of MSFT.

The sentiment scores for the news articles analyzed using TextBlob were closer to zero, indicating that the majority of the news articles were classified as neutral. This result was expected since most news articles tend to provide a balanced view of a company's performance. However, the SA scores for MSFT and GOOG were higher than the other companies, indicating a positive market sentiment towards these companies. This sentiment is likely due to the companies' AI integration and technological advancements, which have been positively received by the market.

The MAE values for the LSTM model ranged from 0.081 for AAPL to 0.613 for MSFT. These values indicate that the LSTM model was able to accurately predict the stock prices for most of the companies, with the exception of MSFT.

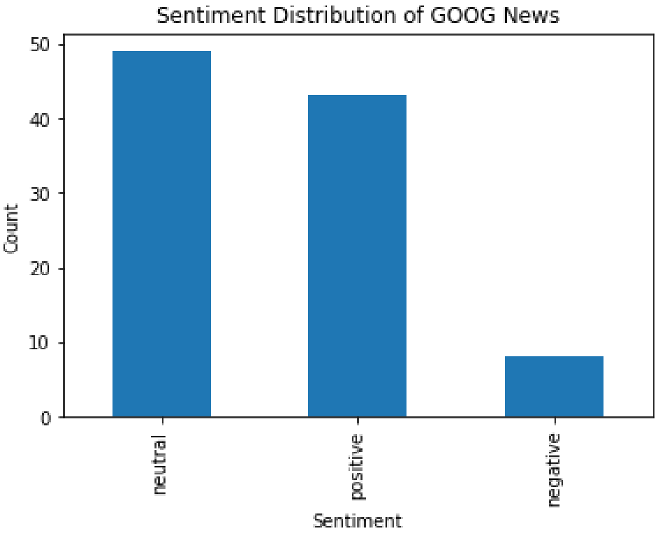


Fig. 4: Sentiment Chart for GOOG- The recent news of GPT integration and AI advancements can be attributed to a higher positive score for Google stock.

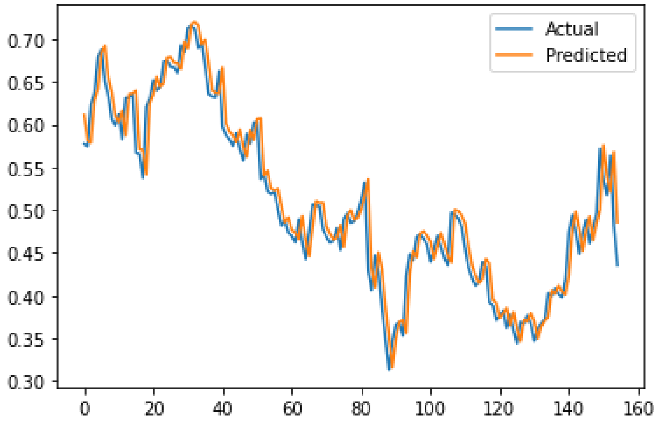


Fig 5: Actual vs predicted RFR Model Performance for GOOG
Training MAE: 0.014844177213421926
Testing MAE: 0.02084196143255049

Using the proposed TFMS analysis model, investors can make informed decisions by predicting a stock price and reviewing the public sentiment weightage for a particular stock, allowing for short-term investments. However, it is important to note that there are numerous other factors that can impact the market, and relying solely on public sentiment may not always be accurate. Consulting a professional and taking a more holistic approach to invest is recommended. This paper presents a multi-factor tool that provides a more comprehensive and effective approach to predicting stock prices, surpassing the limitations of single-method models. The model performed well for the five test cases and successfully identified four bull runs, demonstrating its potential as a reliable tool for investors in a relatively non-volatile market.

VII. CONCLUSION

In conclusion, this research paper proposed a novel multi-factor analysis model for stock market prediction, which integrates technical analysis, fundamental analysis, machine learning, and sentiment analysis. The proposed model leverages the Random Forest Regressor (RFR) and Long Short-Term Memory (LSTM) algorithms to predict stock prices and adjust the predicted price using sentiment analysis. The model was evaluated using historical data from the top five American companies, and it was compared to traditional prediction models. The results demonstrate that the proposed model provides improved accuracy in predicting future stock trends. The proposed multi-factor analysis model represents a significant step forward in stock price prediction, providing a more comprehensive and effective approach to predicting stock prices based on multiple factors. This research paper contributes to the field of stock market prediction by proposing a multi-factor approach that leverages the strengths of technical analysis, fundamental analysis, machine learning, and sentiment analysis, and can be used to inform investment decisions. Further research can be conducted to improve the model and extend its application to other stock markets and time periods.

VIII. FUTURE WORK

There is room for further research in the proposed multi-factor analysis model for stock market prediction. One potential avenue for future work is to explore the correlation between the RFR and LSTM scores, combined with sentiment analysis, to develop a more accurate prediction model. Reinforcement learning can be used to train the model on the output values from other new methods, and real-time correlations between scores can be tested. In addition, large language models can be used to improve sentimental analysis, and more focus can be given to the analysis of neutral data and news. The development of new techniques and the integration of emerging technologies can help in the creation of more comprehensive and effective models for stock market prediction.

IX. CODE

https://github.com/akashdeepo/TFMS-Multifactor-Analysis

REFERENCES

[1]. Agrawal, J. G., V. Chourasia, and A. Mittra. "State-of-the-art in stock prediction techniques." International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering 2.4 (2013): 1360-1366.

[2]. DEMİREL, Uğur, Ç. A. M. Handan, and Ü. N. L. Ü. Ramazan. "Predicting stock prices using machine learning methods and deep learning algorithms: The sample of the Istanbul Stock Exchange." Gazi University Journal of Science 34.1 (2021): 63-82.

[3]. Drakopoulou, Veliota. "A review of fundamental and technical stock analysis techniques." Journal of Stock & Forex Trading 5 (2016).

[4]. Khaidem, Luckyson, Snehanshu Saha, and Sudeepa Roy Dey. "Predicting the direction of stock market prices using random forest." arXiv preprint arXiv:1605.00003 (2016).

[5]. Raschka, Sebastian, and Vahid Mirjalili. Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing Ltd, 2019.

[6]. Roondiwala, Murtaza, Harshal Patel, and Shraddha Varma. "Predicting stock prices using LSTM." International Journal of Science and Research (IJSR) 6.4 (2017): 1754-1756.

[7]. Usmani, Mehak, et al. "Stock market prediction using machine learning techniques." 2016 3rd international conference on computer and information sciences (ICCOINS). IEEE, 2016.

[8]. Xu, Shuo, Yan Li, and Zheng Wang. "Bayesian multinomial Naïve Bayes classifier to text classification." Advanced Multimedia and Ubiquitous Engineering: MUE/FutureTech 2017 11. Springer Singapore, 2017.