Opinion Mining for Social Networking Platform for Stock Price Prediction

Nitin Kumar[1], Shubham Agrawal[2] , Kumar Harsh[3], Dr. Ankush Jain[4]

nitin.kumar.co19@nsut.ac.in[1], shubham.agrawal.co19@nsut.ac.in[2], kumarharshnsut@gmail.com, ankush.jain@nsut.ac.in[4]

Department of Computer Engineering

Netaji Subhas University of Technology, India

Azad Hind Fauj Marg Sector - 3 , Dwarka New Delhi – 110078

Abstract :

Social media serves as both a communication tool and a valuable database for researchers and practitioners to gather information, share knowledge, and express opinions about stock performance. The sentiment expressed in social media content can be analyzed to predict stock performance. While many past studies have attempted to predict stock price movement using social media sentiment, some emerging analytical tools, such as existing lexicons, may require further testing and validation in a financial decision-making context. This study develops and tests predictive models for stock price and trend forecasting, using a large-scale sample of tweets related to four companies: Apple, Google, Microsoft, and Netflix. A novel decision tree approach is proposed for stock performance prediction, and theoretical and practical implications are provided based on the findings. Directions for future work are also discussed. In the field of stock price prediction, sentiment analysis using social media information has gained significant attention from researchers and practitioners across various disciplines, such as computer science, statistics, finance, and economics. Previous studies have shown that public opinions expressed and shared online, and their embedded sentiment, can greatly impact investment decisions.

Keywords : Stock market prediction, Machine learning, Long Short Term Memory(LSTM), Sentiment analysis, Historical analysis, Tweepy, Twitter, TextBlob, National Stock Exchange (NSE)

I. Introduction

The concept of "beating the market" has always intrigued economists, investors, and researchers [1]. Countless discussions and studies have been conducted to better understand how to profit from it. One area of interest is using public sentiment from social media. Mukhtar notes that "leveraging social media to obtain the fastest news affecting stock prices is no longer a theory but a reality" [2]. With over 2.3 billion active Facebook users recorded in 2021 [3], recent studies have demonstrated that sentiment data gathered from this vast pool of people can be used to make stock market decisions and even predictions [4]. The correlation between stock market performance and public sentiment is evident in specific situations, particularly when popular individuals share their views about certain stocks, which directly impacts the stock's performance [5]. However, despite ongoing research in this area, much of the work has produced mixed results, indicating the need for further exploration [6].

Efficiently extracting and calculating sentiment from vast amounts of social media data is a primary challenge in this field of work [7]. Currently, the most effective method to accomplish this is through the use of computational power and Machine Learning [8]. Machine Learning has become pervasive in our lives and is present in speech recognition technologies, image classification tools, autonomous driving, and numerous other areas [9]. Sentiment Analysis (SA) is a specific area of Machine Learning that can be utilized to analyze sentiment from social media efficiently [10]. SA is a type of data intelligence that is used to identify the emotion behind text [11]. It can be implemented through Natural Language Processing (NLP), and for it to function effectively, it must be configured for its specific purpose, which necessitates an understanding of the right datasets and types of algorithms to use [12]. As a result, selecting the appropriate SA algorithm for predicting social media sentiment is critical to our objective of assisting users in making investment decisions [13].

The initial objective of this study is to offer an understanding of the field of study and describe current research and technologies. Firstly, it will investigate the correlation between public sentiment and stock market performance, followed by an examination of sentiment from social media, and ultimately the topic of Sentiment Analysis (SA) [14]. Additionally, it will feature a technical review component, in which it will present current methods for creating SA models, extracting social media data, and assessing outcomes [15].

The second objective of this study is to demonstrate the potential for utilizing public sentiment from social media to assist users in making investment decisions. This will be accomplished by developing and deploying a prototype that will monitor the stock market performance of specific S&P 500 stocks, leveraging the sentiment data obtained from social media. Additionally, a new method of incorporating social media metrics into the sentiment score will be introduced, which will take into account posts like weighting in the SA algorithm sentiment score. The purpose of this is to attempt to accurately measure the overall sentiment of the data [16].

The presented article is organized as follows :

Section 2 mentions the research work carried out by fellow researchers. Section 3 discusses the methodology which includes dataset requirements, libraries used, 2-stage architecture (Tweepy and LSTM) and their working, System architecture, and Project File structure. Section 4 discusses the result analysis and it includes Performance Parameters, Training steps, Testing steps, Testing results. Section 5 Observation and comparison of models and the flow chart showing the working of the model. Section 6 is about the conclusion of the project. Section 7 discusses the future scope and applications of this project. Section 8 mentions the references that have been used to create this research paper.

II. Literature Work

The research papers used as reference for developing the Opinion Mining Model are as follows :

Forecasting stock market index daily direction: A Bayesian Network approach by Malagrino, L. S. , Roman, N. T. & Monteiro [1]

This study examines the use of Bayesian networks for stock market forecasting, with a focus on the paper by Vellido et al. (2012) which applies a Bayesian network approach to forecast the direction of the daily movements of the Dow Jones Industrial Average index. The study discusses previous research that has utilized Bayesian networks for stock market forecasting and highlights the advantages of this method, such as its ability to model complex relationships between variables. While the Vellido et al. study achieved a higher accuracy than several benchmark methods, the study suggests that there is still room for improvement and future research could explore more sophisticated network structures and the integration of additional data sources.

A novel data-driven stock price trend prediction system. Expert Systems with Applications by Zhang, J., Cui, S., Xu, Y., Li, Q., & Li, T [2]

This paper examines data-driven approaches for predicting stock price trends. It was published in the journal Expert Systems with Applications. The paper discusses previous research that has utilized machine learning algorithms, such as deep belief networks and support vector machines, for stock price trend prediction with promising results. Zhang et al. (2018) proposed a novel system that combines multiple machine learning algorithms and feature selection techniques to achieve an accuracy of 63.9%, outperforming several benchmark methods. While data-driven systems show potential, there is still room for improvement in terms of accuracy and feature selection. Future research could explore more advanced machine learning algorithms and the integration of additional data sources for improved predictions.

Predicting the effects of news sentiments on the stock market by Shah, D., Isah, H., & Zulkernine, F. [3]

The paper by Shah et al. (2018) presents a system for predicting the effects of news sentiments on the stock market, which combines machine learning and statistical techniques to analyze sentiment in news articles and predict its impact on stock prices. The system achieved an accuracy of 57.14% in predicting the effects of news sentiment on the New York Stock Exchange. While news sentiment analysis shows promise for predicting stock market movements, there is still room for improvement in terms of accuracy and the selection of relevant news sources.

A decision support approach for online stock forum sentiment analysis by Wu, D. D., Zheng, L., & Olson, D. L. [4]

This study examines previous research on sentiment analysis in online stock forums and focuses on the paper by Wu et al. (2014) titled "A decision support approach for online stock forum sentiment analysis." The authors proposed a system that combines sentiment analysis, opinion mining, and visualization techniques to extract useful information from user-generated content in online stock forums. The system achieved an accuracy of 78.5% in predicting the sentiment of the posts. While sentiment analysis techniques have shown promise in predicting stock prices and supporting investment decisions, there is still room for improvement in accuracy and the selection of relevant user-generated content. Future research could explore more advanced natural language processing techniques and the integration of additional data sources. The paper by Wu et al. (2014) provides an example of a decision support approach that combines sentiment analysis, opinion mining, and visualization techniques for online stock forum sentiment analysis with high accuracy.

III. Proposed Methodology

A. Dataset Requirements :

A stock price prediction dataset would typically include historical data of stock prices, including features such as opening price, closing price, and other relevant financial indicators, recorded on a daily basis. However, some commonly used financial indicators for stock price prediction include :

1. **Opening price :** The price at which a stock opens for trading at the beginning of the trading day.
2. **Closing price :** The price at which a stock closes for trading at the end of the trading day.
3. **High price :** The highest price at which a stock was traded during the trading day.
4. **Low price :** The lowest price at which a stock was traded during the trading day.
5. **Volume :** The total number of shares of the stock that were traded during the trading day.
6. **Market capitalization :** The total market value of the company's outstanding shares of stock.
7. **Dividend yield :** The ratio of the annual dividend payment to the current stock price.
8. **Moving averages :** A commonly used technical indicator that smooths out price fluctuations and helps identify trends in the stock's price movement.

About Dataset :

Name of dataset : AMZN.csv
Source of dataset : Yahoo Finance
Description of dataset : opening & closing price of stocks for a particular date or a day.

B. Libraries Used :

Tensorflow : An open-source toolkit for deep learning that works with Python and other frameworks.

Keras : An open-source Python package that allows deep learning models to be evaluated.

Numpy : The Numpy library is used to work with arrays.

Matplotlib : A Python library for creating static and animated graphics.

Tqdm : A Python package that allows you to create Progress Meters or Progress Bars.

C. Architecture

1. Mathematical Analysis on Historical data

Long Short Term Memory (LSTM)

Long Short Term Memory (LSTM) [57] is a machine learning technique that was developed by Hochreiter and Schmid- huber in 1997. It was introduced to overcome the flaws of its predecessor RNN (Recurrent Neural Network). The biggest drawback of RNN is that it lacks learning property when the extent in between the requirement and previous information increases, this is known as the long-term dependency problem. LSTM overcomes the long-term dependency problem as it has the ability to hold the values for both long and short duration of time.
LSTM has a totally different architecture as compared to other neural network models. Considering RNN, it has a very simple feedback loop neural network design whereas LSTM consists of a memory block or a cell positioned inside a single neural network layer. Since LSTM is good at remembering information for quite a long time, it becomes the first choice for use as it enhances the accuracy of the prediction models.

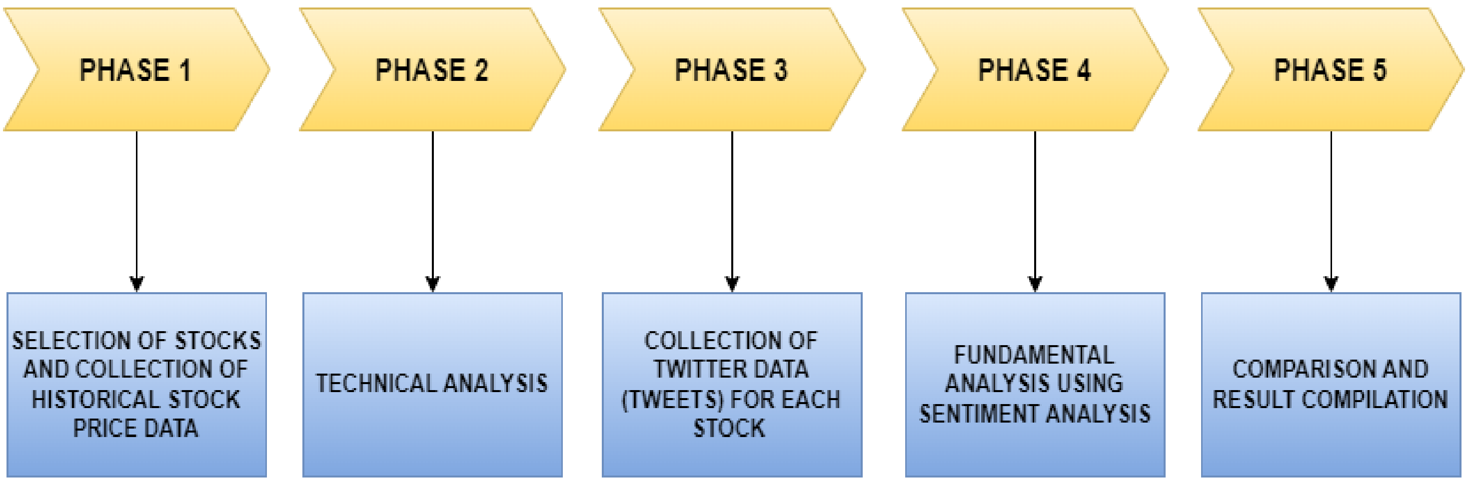


Figure 1: A flowchart of the workflow of the project and the key objectives and challenges

In our work, we proposed a model that uses an LSTM network for prediction. Our model incorporates historical as well as sentiment data to predict future stock prices. Fig. 1 shows how the LSTM Network works, here Ct − 1 is the old cell state, Ct is present cell state, ht − 1 is an output of the previous cell, ht is an output of the present cell, it is input gate layer, ft is forget gate layer and it is output sigmoid gate layer.

2. Sentimental Analysis using Twitter's data

Twitter API – Tweepy

Twitter, one of the most popular microblogging websites, plays a vital role in sentiment analysis for various fields like election results prediction [58], cryptocurrency price prediction [59], etc. Twitter provides the feature to mine their tweets data for research purposes using a Twitter API or Tweepy. The user needs to register first as a developer to gain access, then a set of keys are generated including consumer key, consumer secret, access key, and access secret. These keys are essential to link the code with the Twitter server and fetch the required data as legitimate users.

Working

The proposed model extracts the features for training the prediction model by analyzing both historical as well as sentiment data. The tweets from Twitter are extracted through Tweepy API and then processed for sentiment analysis using Text Blob. After this, we extract the historical data from NSE India. Then a model is trained for stock market prediction using stock price data and sentiment score to predict the change in the stock market. Further, the stock price prediction is performed using LSTM by adding new features into the historical dataset. The features added in the dataset are the sentiment data classified into positive and negative classes and their respective percentages.

The proposed methodology for predicting the stock market movement through sentiment analysis is carried out in 5 phases as shown in Fig. 3 that describes the working of the proposed model. The 1st phase gathers the data for historical and sentiment analysis from NSE India and Twitter respectively. In the 2nd phase, the historical data is pre-processed. The data obtained from NSE and Twitter can't be used in the prediction model as it is, hence few amendments are essential to incorporate the data. The historical data needs to be normalized into a certain range which would maintain the trend of the market as there may be instances where the data outlines a certain boundary and might totally change the outcome. Also, the data may contain null or missing values, therefore it is important to rectify them by using the mean of the data so the trend is not hampered. Regarding the sentiment data, it is essential to use the percentage values of the obtained tweets. Once the data is classified using TextBlob, the percentage of each class is computed and added to the data. The 3rd phase classifies the data collected from Twitter using Tweepy into 3 classes, positive, negative, and neutral using TextBlob. In the 4th phase, both sentiment data and historical data are merged into one common dataset for better calculations and enhancing efficiency. The 5th phase predicts the output (stock prices) with the help of the LSTM network.

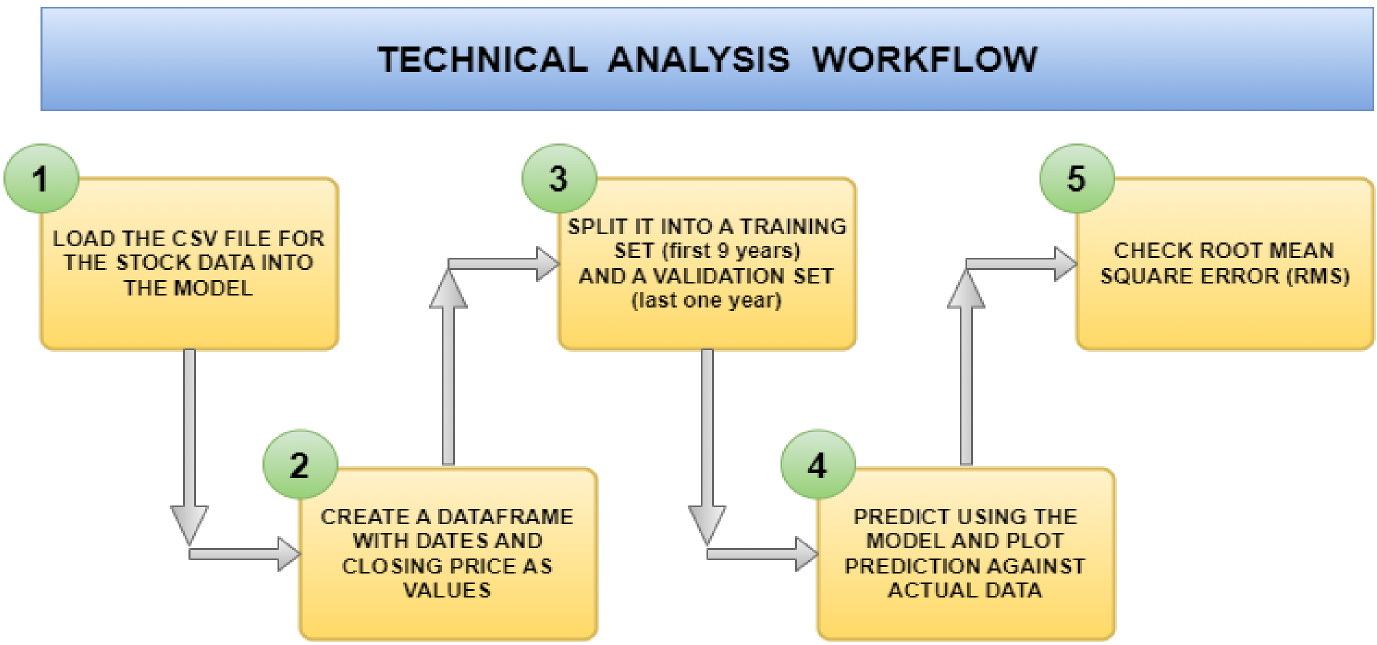


Figure 2 : Technical Analysis Workflow

IV. Results Analysis

1. We compared the various methods of technical analyses to obtain the most accurate one.
2. We compared the two fundamental analyses cases to obtain which was more accurate.
3. We then compare the accuracy of the best fundamental analysis and technical analysis techniques with the real world data to see which gives more accurate results.
4. We then combine the two to generate a more accurate model for stock prediction.

4.1 TECHNICAL ANALYSIS

In the following pages, we have shown the plots for the various models for 4 of the stocks we selected –

- **Ford**
- **Goldman Sachs**
- **JP Morgan Chase**
- **Amazon**
- **Facebook**

This is done to compare the prediction capabilities of the models across stocks.

Moving Average Model (MA) : A time series model that uses the average of past observations to forecast future values. It is a simple and effective method for smoothing out the noise in a time series.

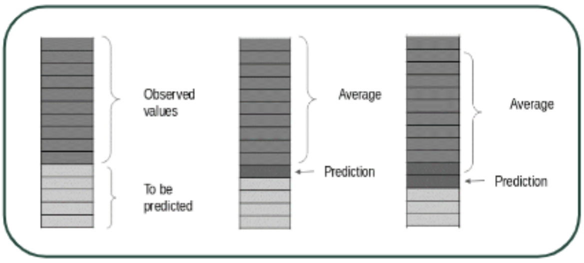


Figure 3 : Moving Average Model

Linear Regression Model : A statistical method that uses a linear equation to model the relationship between a dependent variable and one or more independent variables. It is a widely used technique in machine learning for predicting numeric outcomes.

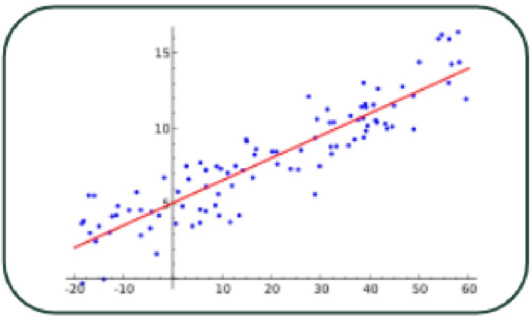


Figure 4 : Linear Regression Model

ARIMA Model : ARIMA stands for AutoRegressive Integrated Moving Average. It is a statistical model for forecasting time series data. It is also often used to understand the time series data better by fitting the time series data to it. It has three parts to it -

1. AR (Autoregression)
2. I (Integrated)
3. MA (Moving Average)
4.

LSTM Model : LSTMs stand for Long Short Term Memories, and are improvements over standard recurrent neural networks.

To understand how LSTMs can be used for stock price prediction, we see the factors that the stock price of today will depend upon :

- The trend that the stock has been following in the previous days, maybe a downtrend or an uptrend.
- The price of the stock on the previous day, because many traders compare the stock's previous day price before buying it.
- The factors that can affect the price of the stock for today. This can be a new company policy that is being criticized widely, or a drop in the company's profit, or maybe an unexpected change in the senior leadership of the company.

These dependencies can be generalized as:

- The previous **cell state** *(i.e. the information that was present in the memory after the previous time step*).
- The previous **hidden state** (*i.e. this is the same as the output of the previous cell*).
- The input at the **current time** step (*i.e. the new information that is being fed in at that moment*).

Note that, for each graph –

A. **Blue** : Past training data

B. **Orange** : Testing data

C. **Green** : Our prediction

A. MOVING AVERAGE MODEL - AMAZON

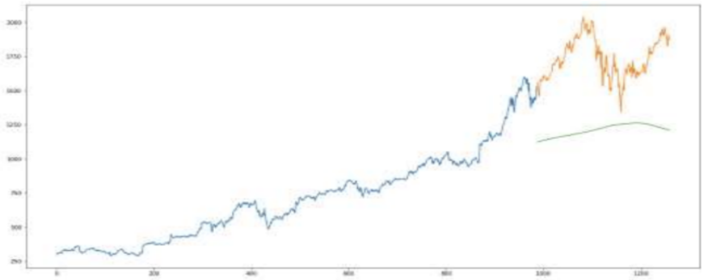


Figure 5 : Moving Average Graph for Amazon

B. LINEAR REGRESSION MODEL - AMAZON

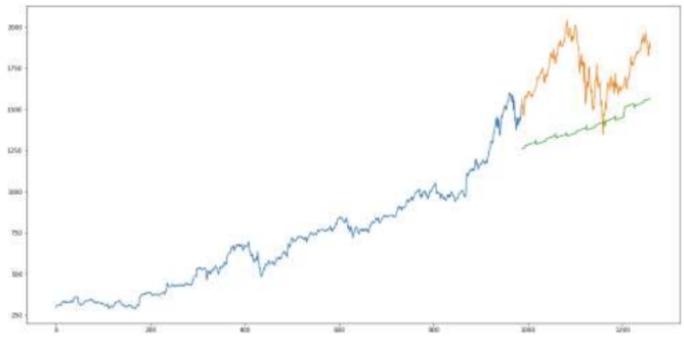


Figure 6 : Linear Regression Graph for Amazon

C. ARIMA MODEL - AMAZON

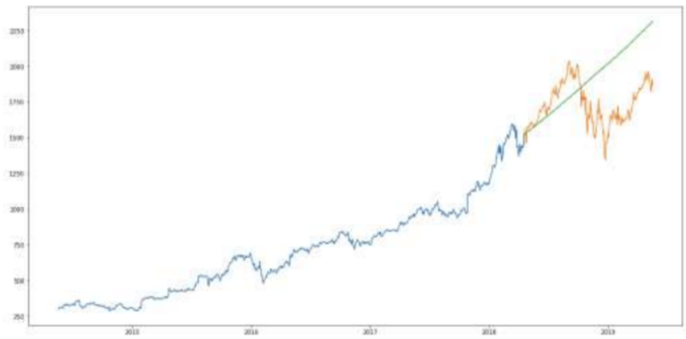


Figure 7 : ARIMA Graph for Amazon

D. LSTM Model - AMAZON

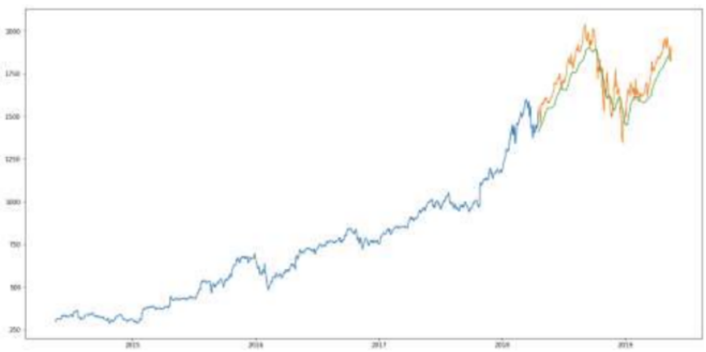


Figure 8 : LSTM Graph for Amazon

4.2 COMPARING VARIOUS TECHNICAL ANALYSES

The table highlights that LSTMs have the lowest percentage RMS error out of all the models tested. This is in line with the graphical observations and proves that LSTMs are the best method for technical analysis of stocks amongst the selected methods, and we can safely move forward with LSTMs for further analysis.

Stock	Technical Analysis RMS error (in %)			
	MA	LR	ARIMA	LSTM
Amazon	0.2954239	0.20107157	0.165251376	0.04892843
AMD	0.39364303	0.28850856	0.388753056	0.06601467
Adobe	0.26205784	0.20964627	0.18778132	0.0641479
Alphabet	0.111965	0.06382518	0.092819343	0.02794847
Facebook	0.11704104	0.2004729	0.118729946	0.04070258
Apple	0.17386243	0.15126984	0.140634921	0.05407407
JP Morgan	0.07683488	0.07168006	0.167734946	0.03430532
Walmart	0.06860995	0.12095975	0.036585366	0.01804481
McDonalds	0.08422849	0.04035739	0.036592712	0.02198574
Coca Cola	0.04695122	0.03861789	0.051626016	0.01544715
General Electric	0.805	1.127	0.656	0.076
General Motors	0.10621622	0.08594595	0.121351351	0.04216216
Goldman Sachs	0.20376842	0.24742947	0.325279846	0.03981158
Ford Motors	0.21185617	0.07774538	0.142857143	0.04567541
Schlumberger	0.44566889	0.25550999	0.527421835	0.09584828
Nike	0.22052737	0.17050964	0.139056403	0.0224666
Procter & Gamble	0.10525826	0.09297348	0.200651466	0.01610051
HP Inc.	0.1256572	0.16508938	0.227129338	0.04100946
Pfizer	0.1478177	0.1294912	0.110682421	0.01736195
Johnson & Johnson	0.05014068	0.07329918	0.049779958	0.02135488

Table 1: Comparing percentage RMS error for tech. analyses (Green indicates minimum error).

Stock	Symbol	Stock sentiment value	Symbol sentiment value	Actual Percentage change (in %)	Accuracy with stock name	Accuracy with stock symbol
Amazon	$AMZN	4.68	6.021	0.43	Yes	Yes
AMD	$AMD	1.91	18.75	0.30	Yes	Yes
Adobe	$ADBE	4.05	9.91	0.19	Yes	Yes
Alphabet	$GOOG	1.07	9.5	-0.59	No	No
Facebook	$FB	1.78	4.57	0.11	Yes	Yes
Apple	$AAPL	4.2	29.18	-0.38	No	No
JP Morgan	$JPM	14.23	11.64	0.98	Yes	Yes
Walmart	$WMT	-1.43	23.81	0.8	No	Yes
McDonalds	$MCD	-0.63	-7.37	-0.6	Yes	Yes
Coca Cola	$KO	-2.29	1.85	-0.48	Yes	No
General Electric	$GE	-11.3	10.4	-1.15	Yes	No
General Motors	$GM	2.4	-10.86	-0.028	No	Yes
Goldman Sachs	$GS	3.65	27.3	0.54	Yes	Yes
Ford Motors	$F	11.91	58.03	-0.2	No	No
Schlumberger	$SLB	12.15	6.65	-1.31	No	No
Nike	$NKE	4.69	5.86	-0.58	No	No
Procter & Gamble	$PG	22.78	10.52	-0.07	No	No
HP Inc.	$HPQ	16.69	7.78	4.38	Yes	Yes
Pfizer	$PFE	4.16	15.22	0.07	Yes	Yes
Johnson & Johnson	$JNJ	-4.46	20.58	0.01	No	Yes

Table 2: Results of sentiment analysis with Tweepy and VADER

Stock	Symbol	Stock sentiment value	Symbol sentiment value	Actual Percentage change (in %)	Accuracy with stock name	Accuracy with stock symbol
Amazon	$AMZN	331.9	277.03	0.43	Yes	Yes
AMD	$AMD	89.5	250.2	0.30	Yes	Yes
Adobe	$ADBE	272.33	381.2	0.19	Yes	Yes
Alphabet	$GOOG	10.23	-23.22	-0.59	No	Yes
Facebook	$FB	234.13	104.22	0.11	Yes	Yes
Apple	$AAPL	341.2	12.17	-0.38	No	No
JP Morgan	$JPM	331.06	139.76	0.98	Yes	Yes
Walmart	$WMT	-131.67	354.43	0.8	No	Yes
McDonalds	$MCD	-24.5	-32.43	-0.6	Yes	Yes
Coca Cola	$KO	-112.43	-46.3	-0.48	Yes	Yes
General Electric	$GE	320.4	1.27	-1.15	No	No
General Motors	$GM	-28.77	-38.21	-0.028	Yes	Yes
Goldman Sachs	$GS	158.89	139.97	0.54	Yes	Yes
Ford Motors	$F	122.67	102.4	-0.2	No	No
Schlumberger	$SLB	-91.12	-23.22	-1.31	Yes	Yes
Nike	$NKE	133.4	-12.34	-0.58	No	Yes
Procter & Gamble	$PG	234.22	-21.3	-0.07	No	Yes
HP Inc.	$HPQ	152.28	239.03	4.38	Yes	Yes
Pfizer	$PFE	7.13	134	0.07	Yes	Yes
Johnson & Johnson	$JNJ	-7.23	157.85	0.01	No	Yes

Table 3: Results of sentiment analysis with web scraping and VADER

S.no	Attributes	Values
1	Accuracy - company names are used to extract tweets	
For stocks with positive movement –		
A.	Accurate predictions	8
B.	Inaccurate predictions	2
C.	Total predictions	10
D.	Accuracy of predictions	80%
For stocks with negative movement –		
A.	Accurate predictions	3
B.	Inaccurate predictions	7
C.	Total predictions	10
D.	Accuracy of predictions	30%
Total –		
A.	Accurate predictions	11
B.	Inaccurate predictions	9
C.	Total predictions	20
D.	Accuracy of predictions	55%
2	Accuracy - stock symbol is used to extract tweets	
For stocks with positive movement –		
A.	Accurate predictions	10
B.	Inaccurate predictions	0
C.	Total predictions	10
D.	Accuracy of predictions	100%
For stocks with negative movement –		
A.	Accurate predictions	2
B.	Inaccurate predictions	8
C.	Total predictions	10
D.	Accuracy of predictions	20%
Total –		
A.	Accurate predictions	12
B.	Inaccurate predictions	8
C.	Total predictions	20
D.	Accuracy of predictions	60%

Table 4 : Accuracy Comparison (a)

S.no	Attributes	Values
1	Accuracy - company names are used to extract tweets	
For stocks with positive movement –		
A.	Accurate predictions	8
B.	Inaccurate predictions	2
C.	Total predictions	10
D.	Accuracy of predictions	80%
For stocks with negative movement –		
A.	Accurate predictions	4
B.	Inaccurate predictions	6
C.	Total predictions	10
D.	Accuracy of predictions	40%
Total –		
A.	Accurate predictions	12
B.	Inaccurate predictions	8
C.	Total predictions	20
D.	Accuracy of predictions	60%
2	Accuracy - stock symbol is used to extract tweets	
For stocks with positive movement –		
A.	Accurate predictions	10
B.	Inaccurate predictions	0
C.	Total predictions	10
D.	Accuracy of predictions	100%
For stocks with negative movement –		
A.	Accurate predictions	7
B.	Inaccurate predictions	3
C.	Total predictions	10
D.	Accuracy of predictions	70%
Total –		
A.	Accurate predictions	17
B.	Inaccurate predictions	3
C.	Total predictions	20
D.	Accuracy of predictions	85%

Table 5 : Accuracy Comparison (b)

V. Observation and Comparison of Models

The values in the above matrix are chosen after observing the trends of predicted vs real data and are subject to change pending further research.

For T>0, we predict upwards movement of stock

For T<0, we predict downwards movement of stock

For T=0, prediction is not clear

5.1 DEVELOPING A REINFORCED MODEL

For our improved model, we utilize both LSTM and sentimental analysis for a more comprehensive analysis of the price prediction. Please note we have just laid a basic foundation for further research, this model is not complete and reliable on its own and needs to include multiple factors and be tested on a large dataset to be deemed reliable.

For this purpose, we assign two scores to the stock –

LSTM score (L) and Sentiment score (S)

Now, let us assume that LSTM score and Sentiment score are related to the total stock score linearly –

$$T = W1.\,L + W2.\,S + c \qquad \text{Eq.(1)}$$

Where W1 and W2 are weights assigned to L and S respectively, and c is a constant

Let W1=W2=1 (Sentiment analysis and technical analysis are given equal weightage in terms of their importance)

Let us assume c = 0

LSTM score is derived from the LSTM model as the difference between todays predicted value and yesterday's value, divided by the yesterday's value (or the predicted percentage change)

If the difference is negative, LSTM has predicted the movement to be downwards. If the difference is positive, LSTM has predicted the movement to be upwards The sentiment score is derived from symbol sentiment value that our model calculated.

Score	Score = -1	Score = 0.5	Score = 0	Score = +0.5	Score = +1
Symbol Sentiment Value	<-10	-10< and <-0.01	-0.01< and <10	10< and <100	<100
LSTM Percentage change	<-0.5	-0.5< and <0.001	0.001< and <0.001	0.001< and <0.5	<0.5

Table 6: Scorecard for reinforced model

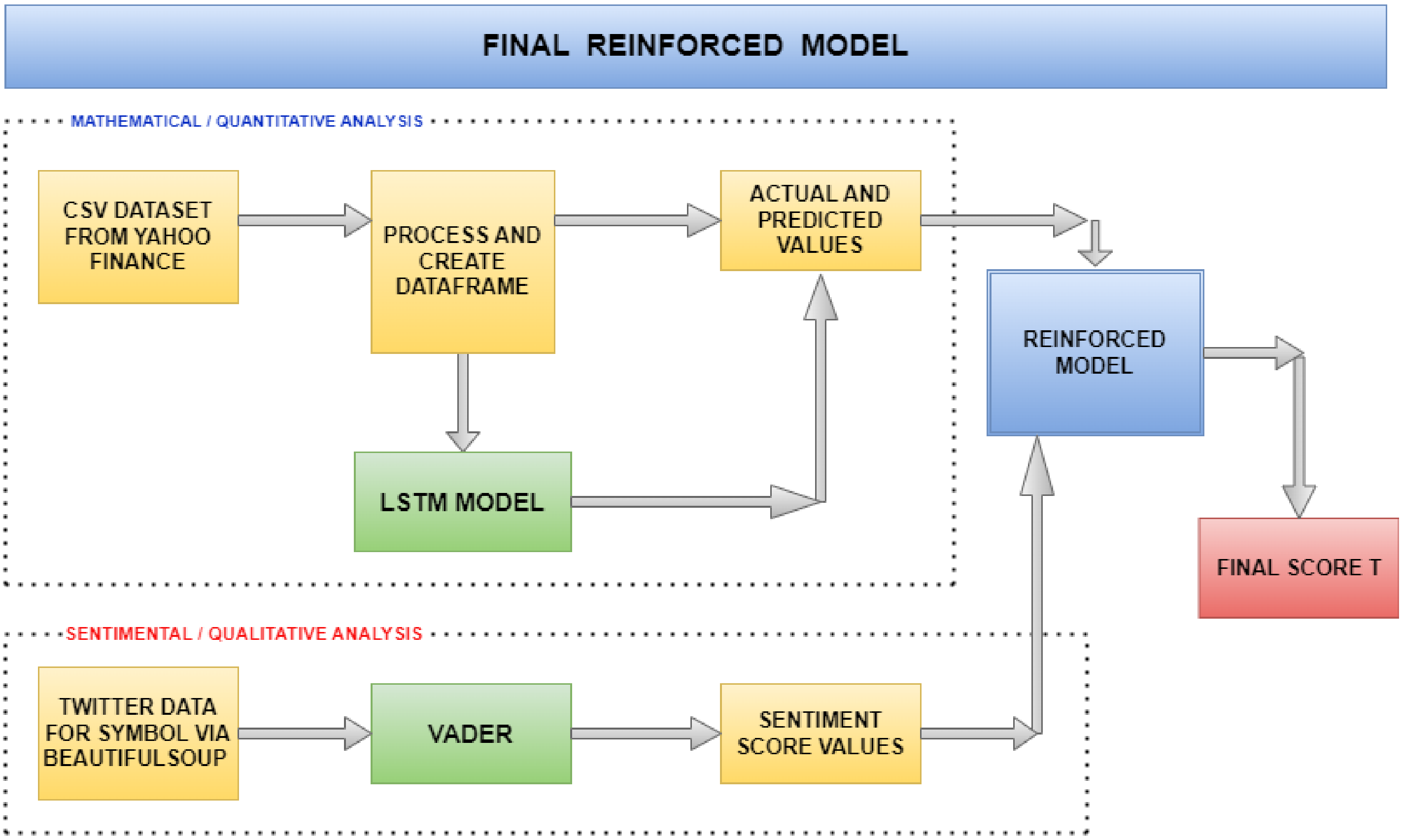
Figure 9 : Plot for Reinforced Model

Stock	Symbol	Symbol sentiment value	S	LSTM percentage difference	L	Actual Percentage change (in %)
Apple	$AAPL	12.17	0.5	-0.084	-0.5	-0.38
General Electric	$GE	1.27	0	-0.51	-0.5	-1.15
Ford Motors	$F	102.4	1	-0.007	0	-0.2
Amazon	$AMZN	277.03	1	-0.036	0.5	0.43

Table 7 : Reinforced model score calculation

Stock	S	L	T	Actual Percentage change (in %)
Apple	0.5	-0.5	0	-0.38
General Electric	0	-0.5	-0.5	-1.15
Ford Motors	1	-0.5	0.5	-0.2
Amazon	1	-0.5	0.5	0.43

Table 8 : Reinforced model testing results



As we can see from the two tables, this new scorecard based reinforced model accurately predicted the movement for one out of the three stocks whose movements were being inaccurately predicted by regular models, and gave an inconclusive result for another one, leading to just one incorrect prediction instead of the previous 3.

VI. Conclusion

Through the course of this project, various technical analyses and sentimental analysis models were compared and a reinforced model was developed. Four technical analyses were compared – Moving Averages method, Linear Regression method, ARIMA and LSTM based method, and the most accurate out of the four, that is LSTM, was chosen for further comparison and development. For sentiment analysis, two means were used to extract data from Twitter – API based method using Tweepy, which gives us a maximum of 3200 latest tweets, and the web scraping based method using BeautifulSoup, which lets us gather a higher number of tweets (in our case, ~20,000). The correlation of the number of tweets analyzed to obtain sentiment value, with the actual stock price movement was obtained. The stocks with the company name being used for tweet extraction and 60% when the company's stock symbol was used. The BeautifulSoup extraction based model gave us 60% accuracy when the company name was being used for tweet extraction, and 85% when the company's stock symbol was used. This result indicates that the market isn't extremely volatile and takes into consideration the entire day's events more than the most recent events, which is a fair result. The results also indicate that the stock symbol provides more relevant tweets. To draw further from these results, groundwork for a new stock market prediction model using LSTM and sentiment analysis was developed, where a stock was given a technical and a fundamental score and its movement was predicted using the sum of those scores.This model gave us slightly higher accuracy as compared to purely. The results indicate that the market considers the entire day's events and the stock symbol is more relevant.

VII. Future Work

1. Use ensemble models: Ensemble models involve combining the predictions of multiple models to make a more accurate prediction. Consider using ensemble models to combine the predictions of different types of models, such as deep learning models and traditional statistical models, to improve the accuracy of the predictor.

2. Implement real-time prediction: Real-time prediction involves predicting stock prices in real-time as new data becomes available. Consider implementing real-time prediction to make the predictor more useful for traders and investors who need up-to-date information.

3. Use deep learning techniques: Deep learning techniques such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) have been shown to be effective in predicting stock prices. Consider using these techniques to improve the accuracy of the predictor. purely Technical or Fundamental analysis based models.

VIII. References

[1] Malagrino, L. S., Roman, N. T., & Monteiro, A. M. (2018). Forecasting stock market index daily direction: A Bayesian Network approach. Expert Systems with Applications, 105, 11-22.

[2] Zhang, J., Cui, S., Xu, Y., Li, Q., & Li, T. (2018). A novel data-driven stock price trend prediction system. Expert Systems with Applications, 97, 60-69.

[3] Shah, D., Isah, H., & Zulkernine, F. (2018, December). Predicting the effects of news sentiments on the stock Page 4858 market. In 2018 IEEE International Conference on Big Data, 4705-4708.

[4] Wu, D. D., Zheng, L., & Olson, D. L. (2014). A decision support approach for online stock forum sentiment analysis. IEEE transactions on systems, man, and cybernetics: systems, 44(8), 1077-1087.

[5] Das, S., Behera, R. K., & Rath, S. K. (2018). Real-time sentiment analysis of twitter streaming data for stock prediction. Procedia computer science, 132, 956-964.

[6] Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. Expert Systems with Applications, 73, 125-144.

[7] Yang, S. Y., Mo, S. Y. K., & Liu, A. (2015). Twitter financial community sentiment and its predictive relationship to stock market movement. Quantitative Finance, 15(10), 1637-1656.

[8] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business horizons, 53(1), 59-68.

[9] Java, A., Song, X., Finin, T., & Tseng, B. (2007, August). Why we twitter: understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, 56-65.

[10] Thomson Reuters. (2017). More financial executives forgo business decisions as personal liability seems more likely, Thomson Reuters Culture and Conduct Risk 2017
Survey finds.

[11] Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., & Jaimes, A. (2012, February). Correlating financial time series with micro-blogging activity. In Proceedings of the fifth ACM international conference on Web search and data mining, 513-522.

[12] Bollen, J., Mao, H., & Pepe, A. (2011, July). Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 5, No. 1).

[13] Chen, M. P., Chen, P. F., & Lee, C. C. (2013). Asymmetric effects of investor sentiment on industry stock returns: Panel data evidence. Emerging Markets Review, 14, 35-54.

[14] Chen, K., Li, X., Luo, P., & Zhao, J. L. (2021). Newsinduced dynamic networks for market signaling: understanding the impact of news on firm equity value. Information Systems Research, 32(2), 356-377.

[15] He, W., Guo, L., Shen, J., & Akula, V. (2016). Social media-based forecasting: A case study of tweets and stock prices in the financial services industry. Journal of Organizational and End User Computing (JOEUC), 28(2), 74-91.

[16] Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. Marketing Science, 31(2), 198-215.

[17] Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management science, 53(9), 1375-1388.

[18] Bukovina, J. (2016). Social media, big data and capital markets – An overview. Journal of Behavioral and Experimental Finance, 11, 18-26. [19] Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data, 2(1), 1-14.

[20] Godsay, M. (2015). The process of sentiment analysis: a study. International Journal of Computer Applications, 126(7).

[21] Bhuriya, D., Kaushal, G., Sharma, A., & Singh, U. (2017, April). Stock market prediction using a linear regression. In 2017 international conference of electronics, communication and aerospace technology (ICECA) (Vol. 2), 510-513.

[22] Singh, S., & Kaur, A. (2022). Twitter sentiment analysis for stock prediction. Available at SSRN 4157658.

[23] Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: A review and taxonomy of prediction techniques. International Journal of Financial Studies, 7(2), 26.

[24] Annett, M., & Kondrak, G. (2008, May). A comparison of sentiment analysis techniques: Polarizing movie blogs. In Conference of the Canadian Society for Computational Studies of Intelligence, 25-35. Springer, Berlin, Heidelberg.

[25] Pang, B., & Lee, L. (2008, August). Using very simple statistics for review search: An exploration. In Coling 2008: Companion volume: Posters, 75-78.

[26] Russell, M. A. (2011). Mining the social web: Analyzing data from Facebook, Twitter, LinkedIn, and other social media sites. " O'Reilly Media, Inc.

[27] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 168-177.

[28] Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv preprint arXiv:1103.2903.

[29] Mohammad, S., & Turney, P. (2010, June). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text, 26-34.

[30] Das, S., Behera, R. K., & Rath, S. K. (2018). Real-time sentiment analysis of twitter streaming data for stock prediction. Procedia computer science, 132, 956-964.

[31] Batra, R., & Daudpota, S. M. (2018, March). Integrating StockTwits with sentiment analysis for better prediction of stock price movement. In 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 1-5.

[32] Vu, T. T., Chang, S., Ha, Q. T., & Collier, N. (2012, December). An experiment in integrating sentiment features for tech stock prediction in twitter. In Proceedings of the workshop on information extraction and entity analytics on social media data, 23-38

[Github code link] : https://github.com/Nitinkumar3399/opinion-mining-stock-price-prediction