STOCK PRICE PREDICTION AND RECOMMENDATION USINGMACHINE LEARNING TECHNIQUES AND TWITTERSENTIMENT ANALYSIS

Kunal Gaur
*Department of Information TechnologyGuru Tegh Bahadur Institute Of Technology
Delhi, India*

Saud Akhtar
*Department of Information Technology
Guru Tegh Bahadur Institute Of Technology
Delhi, India*

Krishvi Srivastava
*Department of Information Technology
Guru Tegh Bahadur Institute of Technology
Delhi, India*

KamalJyot Singh
*Department of Information Technology
Guru Tegh Bahadur Institute Of Technology
Delhi, India*

Gaurav Sandhu
*Department of Information Technology
Guru Tegh Bahadur Institute Of Technology
Delhi, India*

***Abstract*— The prediction of the stock market has entered a technologically advanced era, redesigning the traditional concept of trade, thanks to technical wonders such as worldwide digitalization. Stock prices are difficult to predict due to their extreme volatility, which is influenced by a variety of political and economic issues, as well as changes in leadership, investor attitude, and a variety of other factors. Stock price predictions based solely on historical data or textual information have shown to be unsatisfactory, That's why we have used both for predicting the stock price and recommendation to buy or sell a particular stock. As stockprice prediction is a Time series problem we have used various machine learning and deep learning techniques such as LSTM, ARIMA & Linear Regression. Out of these ARIMA performedreally well with an accuracy of 83%. Existing studies in sentiment analysis also have found that there is a strong correlation between the movement of stock prices and the twitter tweets for a particular company. We have performed sentiment analysis on the latest tweets of the respective company and will provide a recommendation to buy or sell that particular stock of a company.**

Keywords—LSTM, ARIMA, Machine Learning, Sentiment Analysis, Trade Open, Trade Close.

I. INTRODUCTION

Stock market prediction refers to predicting a company's present developments and the value of its stocks, whether they are rising or falling. The stock market is where a company's stock is traded. A stock is a type of investment that reflects ownership in a corporation. The stock market is where such stocks are bought and sold. Buying a company's shares is like buying a small piece of an institution. There are many factors which can impact the price of the stock. A prediction model that just takes into account one component may not be reliable. As a result combining both the tweets and historical price of the stock might improve the accuracy. There are primarily two approaches for predicting market trends. Technical analysis and Fundamental analysis are two types of analysis. Fundamental analysis uses previous price and volume to forecast future trends, but technical analysis does not. Fundamental analysis of a firm, on the other hand, entails evaluating financial data to get insights. Theefficient-market theory, which holds that stock market prices are basically unpredictable, casts doubt on the usefulness of both technical and fundamental analysis. The goal of this research work is to build a model which predicts the stock trends. Three models are used as a part of this research work. The models are ARIMA, LSTM and Linear Regression. Sentiment Analysis is performed by using twitter data of the company.
LSTM model was made by Hochreiter & Schmidhuber [1] which was capable of learning long term dependencies. Later on, many researched improved this work in [2] [3] [4].

The rest of the paper is organized is as follows. Section 2 includes the research state of the stock price prediction. Section 3 includes the Data Collection & Preprocessing. Section 4 consists of the methodologies used. Section 5 includes the Experimental Results. Section 6 concludes the paper.

II. LITERATURE SURVEY

The Siliverstovs of Manh Ha Duong Boris [5], investigate the abstraction between equity prices and combined financesin key European countries such as the United Kingdom and Germany. Acceleration in European country investments is likely to result in a stronger link between European nation equity prices. If innovations in stock markets effect actual financial instruments like investment and consumption, this operation might lead to a merger in financial development across EU states. Fahad Almudhaf et al.[6], examine CIVETS' weak-form market efficiency from 2002 to 2012. CIVETS employs the random walk hypothesis procedure.
P. Bhat[7], employed convolution neural networks to forecast stock prices. In this model, learning is completed by computing the mean square error for each subsequent

observation, and the model with the least error and the highest predictive power is chosen. They are using CNN in this study to forecast stocks and incentives for the next day. For future stock price prediction, Mohammad Mekayel Anik et al[8], designed a linear regression technique. They met their objectives in that the model's accuracy is excellent, and it may be used to forecast stock prices. Xiao Ding et al.[9] created an intuitive and effective interface to include common sense information into the learning process.

Alostad and Davulcu [10] used hourly stock prices for 30 stocks as well as NASDAQ online stock news items. For six months, they gathered tweets linked to those 30 stocks. Li etal. [11] gathered data from the Hong Kong stock market for five years. They collected financial news stories from the same time period to see if there was a link between the articles and stock market changes. They gathered the open, high, close, and low stock prices for each firm on a given trading day. The stock price trend prediction issue was addressed as a classification problem by Alostad and Davulcu [12]. They used logistic regression on the n-gram document matrix, hourly stock price direction, and document weight. They then performed the classification using SVM. Experiments also revealed that extracting document-level sentiment does not improve prediction accuracy considerably. Random forest, naïve Bayesian, and evolutionary algorithms have all been used to forecast stock price and direction in earlier studies.

Stock prediction model employing logistic regression with feature index variables has been established by Gong and Son [13]. They claim that logistic regression outperforms other approaches such as the RBF – ANN prediction model for daily stock trading prediction.

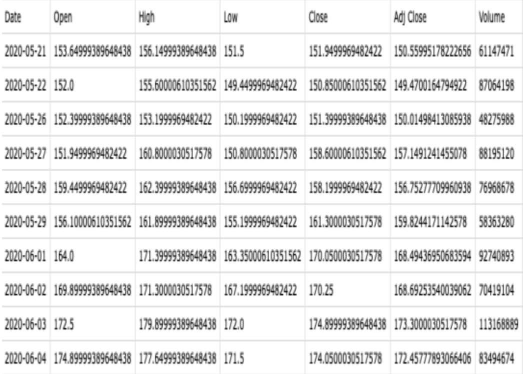
J. Bean [14] utilises keyword tagging to measure polarity and emotion in Twitter feeds concerning airline satisfaction. This might give you a fast overview of how people feel about airlines and their customer service scores.

From all above researches done we were not able to find anything which have used more than 1 models and based on Polarity of tweets a recommendation of buy/sell is given to the user.

III. DATA COLLECTION AND PREPROCESSING

A. *Data Collection*

- For getting the historical data, we downloaded the last 2 years dataset of a particular company by using **Yahoo Finance** API (https://finance.yahoo.com/).
- For Sentiment Analysis Part, firstly we are reading the company name from a csv file named **Yahoo- Finance-Ticker-Symbols.csv** and then fetching the latest tweets of that particular company by using the library named **Tweepy**.



Date	Open	High	Low	Close	Adj Close	Volume
2020-05-21	153.64999389648438	156.14999389648438	151.5	151.9499969482422	150.55995178222656	61147471
2020-05-22	152.0	155.60000610351562	149.4499969482422	150.85000610351562	149.4700164794922	87064198
2020-05-26	152.39999389648438	153.1999969482422	150.1999969482422	151.39999389648438	150.01498413085938	48275988
2020-05-27	151.9499969482422	160.8000030517578	150.8000030517578	158.60000610351562	157.1491241455078	88195120
2020-05-28	159.4499969482422	162.39999389648438	156.6999969482422	158.1999969482422	156.7527770996093 8	76968678
2020-05-29	156.10000610351562	161.89999389648438	155.1999969482422	161.3000030517578	159.8244171142578	58363280
2020-06-01	164.0	171.39999389648438	163.35000610351562	170.0500030517578	168.49436950683594	92740893
2020-06-02	169.89999389648438	171.3000030517578	167.1999969482422	170.25	168.69253540039062	70419104
2020-06-03	172.5	179.89999389648438	172.0	174.89999389648438	173.3000030517578	113168889
2020-06-04	174.89999389648438	177.64999389648438	171.5	174.0500030517578	172.4577789306640 6	83494674

Table III.1 **Sample Data Input of SBIN.NS**

B. *Data Preprocessing*

To make the data from the mode of entry appropriate for trustworthy analysis, it has to be pre-processed.

- We preprocessed the historical data in the following manner :-

S.No.	Techniques
1.	Dropping Null Values present in the dataset if any.
2.	Appending the stock symbol at the end of a respective company.
3.	Used normalization techniques to get data in same range.

- Data Preprocessing For ARIMA Model :-

S.No.	Techniques
1.	Parsing the date and time in format '%Y-%m-%d'.
2.	Filling the null values with backward fill method.
3.	Taken 80% of the data for training and 20% data for testing.

- Data Preprocessing for LSTM Model :-

S.No.	Techniques
1.	Scaled the values using Min- Max Scaler
2.	Storing trends of a particular company from 7 days before current day to predict 1 next output and storing them to training part
3.	Converting training list into numpy arrays
4.	Adding 3rd Dimension to training part.

- Data Preprocessing for Linear Reg Algorithm :-

S.No.	Techniques
1.	Declaring number of days (n) to be forecasted in future.
2.	Declaring new dataframe with relevant data.

- Data Preprocessing for Tweets :-

S.No.	Techniques
1.	Cleaning up the tweets.
2.	Passing the tweets to TextBlob for calculating the Polarity.

IV. METHODOLOGIES

A. ARIMA (Auto Regressive Integrating Moving Average):-

ARIMA stands for auto regressive integrated moving average. It is a statistical analysis model that uses time series data to better understand the data set or anticipate future trends.

If a statistical model predicts future values based on previous values, it is called autoregressive. For example, an ARIMA model may try to anticipate a company's earnings based on prior periods or predict a stock's future pricing based on historical performance.

The model's final goal is to forecast future time series movement by looking at disparities between values in the series rather than actual values. When there is evidence of non-stationarity in the data, ARIMA models are used. Non-stationary data are always turned into stationary data in timeseries analysis.

We can break down the model into smaller components based on the name:

- The AR , which stands for Autoregressive Model, shows a random process. The output of the modelis linearly dependent on its prior value, such as the number of lagged data points or previous observations.
- Integrative (I) :- denotes the separating of raw observations so that the time series can become stationary (i.e., data values are replaced by the difference between the data values and the previous values).
- Moving Average (MA):- It takes into account the relationship between an observation and a residual error from a lagged moving average model.

B. LSTM (Long Short Term Memory Network) :-

It's a unique type of recurrent neural network that can learn long-term data relationships. This is possible because the model's recurring module is made up of four layers that interact with one another. An LSTM module has a cell state and three gates, giving it the ability to learn, unlearn, or retain information from each of the units selectively. By permitting only a few linear interactions, the cell state in LSTM allows information to travel across the units without being altered. Each unit contains an input, output, and a forget gate that adds or removes data from the cell state. The forget gate utilises a sigmoid function to determine which information from the previous cell state should be ignored. The input gate uses a point-wise multiplication operation of 'sigmoid' and 'tanh' to control the information flow to the current cell state. Finally, the output gate determines which data should be transmitted on the next hidden state. LSTM can be used in many applications such as for weather forecasting, NLP, speech recognition, handwritingrecognition, time-series prediction, etc The cell state is represented by the horizontal line that runs across the top of the figure. The condition of the cell is similar to a conveyor belt. This flows straight down the chain with just minimal linear interactions. The ability of LSTM to add or delete information from the cell state is controlled by gates. Gates are used to allow information to pass through if desired. A sigmoid neural net layer plus a point wise multiplication operation make up gates. The sigmoid layer produces values ranging from 0 to 1, indicating how much of each component should be allowed to pass. Let nothing through with a value of 0, and everything through with a value of 1! To safeguard and govern the cell state, an LSTM contains three of these gates.

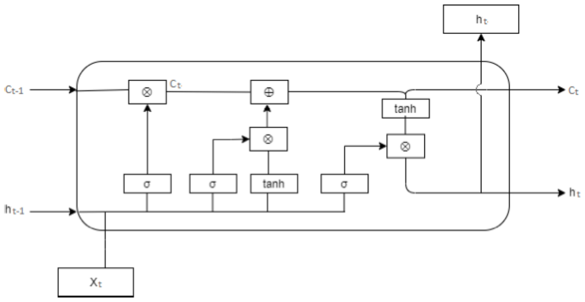


Figure IV.1 **LSTM Architecture**

The prior hidden state (h_t-1), previous cell state (C_t-1) and present input are the inputs to the current cell state (C_t), as illustrated in Fig. IV.1. (X_t). The forget gate, input gate, and output gate are the three gates that make up the cell.

C. Linear Regression :-

Linear regression is a supervised learning machine learning algorithm. It carries out a regression task. Based on independent variables, regression models a goal prediction value. It is mostly utilised in forecasting and determining the link between variables. Different regression models differ in terms of the type of relationship they evaluate between dependent and independent variables, as well as the amount of independent variables they employ. Linear regression is used to predict the value of a dependent variable (y) given an independent variable (x). As a result of this regression technique, a linear relationship between x (input) and y (output) is discovered (output). Linear Regression gets its name from this.

$$\mathbf{Y = (W^{i} * x + b)}$$

D.Research Methodology :-

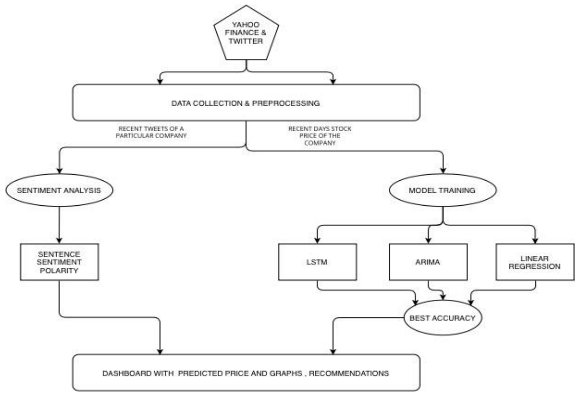

YAHOO FINANCE & TWITTER
DATA COLLECTION & PREPROCESSING
RECENT TWEETS OF A PARTICULAR COMPANY
RECENT DAYS STOCK PRICE OF THE COMPANY
SENTIMENT ANALYSIS
MODEL TRAINING
SENTENCE SENTIMENT POLARITY
LSTM
ARIMA
LINEAR REGRESSION
BEST ACCURACY
DASHBOARD WITH PREDICTED PRICE AND GRAPHS , RECOMMENDATIONS

Figure IV.2 System Architecture

A dashboard was created where a user can enter any stock symbol for predicting its future price up to 7 days. After entering the stock symbol user will be migrated to another webpage where he/she can find predictions given from different models with their RMSE and polarity from latest tweets. A recommendation will be given to user based on theoverall Polarity whether to buy the stock or not.

V. EXPERIMENTAL RESULTS

The models ran for various stock. Some illustrations are given below. We'll be seeing the RMSE (Root Mean Squared Error) for various stocks.

A. *AAPL (Apple Inc.)*

- ARIMA :-
  - ARIMA model was ran for Apple stock. We got an RMSE value of 3.54 which is pretty good as we can see in the following graph plotted below

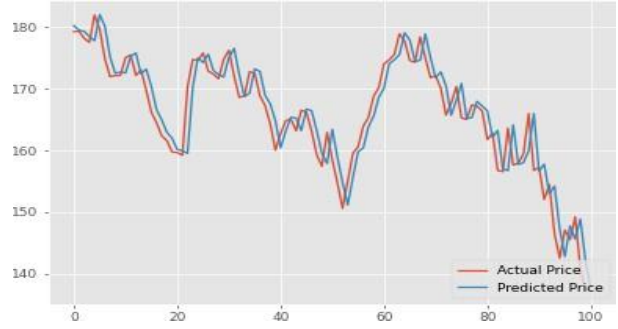


Figure V.1 :- ARIMA Model Predictions

- LSTM :-
  - We ran LSTM model for various epochsas depicted in the table below. After 30 epochs we got a good RMSE value.

Table V.1 (LSTM On Various Epochs)

Epochs	RMSE
10	7.49
20	6.93
30	5.64
40	6.74

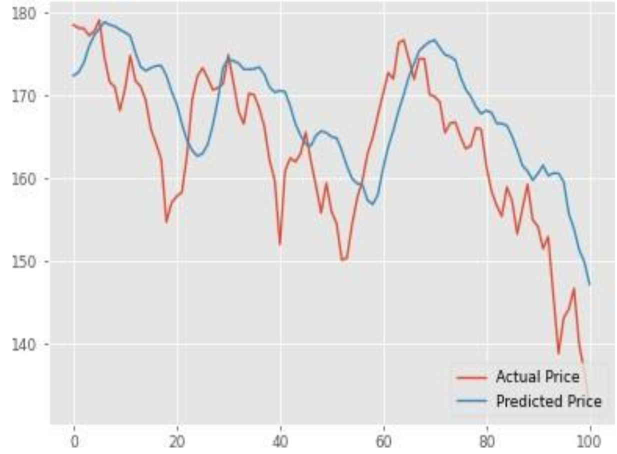


Figure V.2. :- LSTM Model Accuracy

- Linear Regression :-
  - Linear Regression model was ran and got an RMSE of 12.84. Following is the graph plotted for Linear Regression Model.

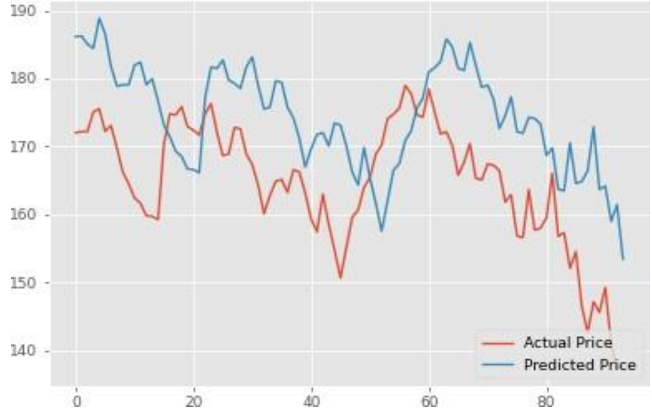


Figure V.3. :- Linear Regression Model Accuracy

- SENTIMENT ANALYSIS :-
  - Taken the latest 20 tweets based on apple company.
  - Calculated Polarity by using TextBlob and got Overall Polarity as Positive.

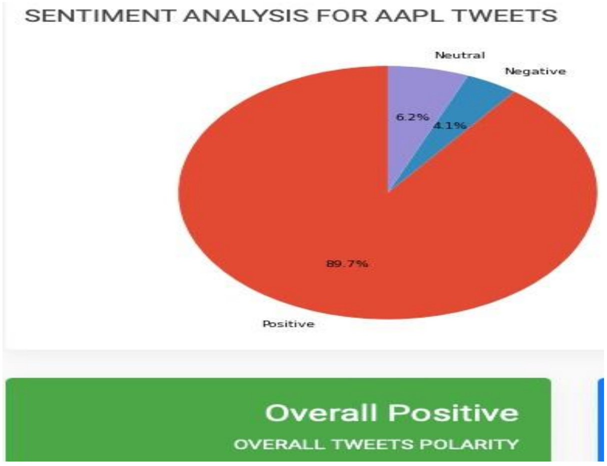
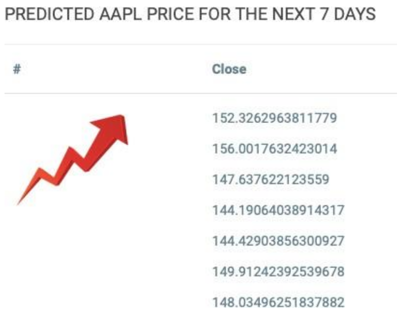


Figure V.4 :- Polarity From Tweets

- Predicted Price for Next 7 days :-
- We can see that Price is increasing.

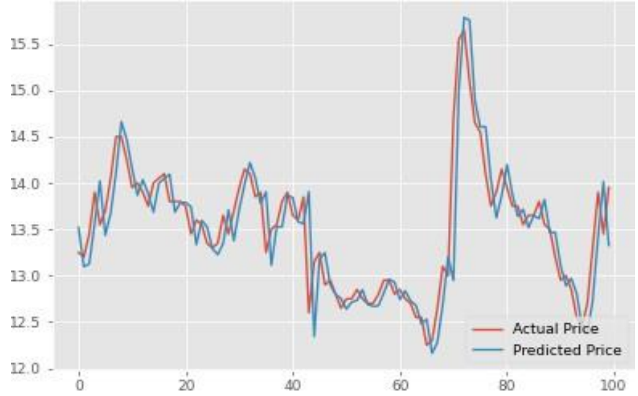


PREDICTED AAPL PRICE FOR THE NEXT 7 DAYS

#	Close
	152.3262963811779
	156.0017632423014
	147.637622123559
	144.19064038914317
	144.42903856300927
	149.91242392539678
	148.03496251837882

YESBANK.NS (Yes Bank Ltd.)

- ARIMA :-
- ARIMA model was ran for Apple stock. We got an RMSE value of 0.35 which is pretty good as we can see in the following graph plotted below



B. Figure V.5 :- ARIMA Model Accuracy

- LSTM :-
- We ran LSTM model for various epochs as depicted in the table below. After 30 epochs we got a good RMSE value.

- Model was tested on test data and a graph is plotted between Actual Price and LSTM's predicted price.

Table V.1 (LSTM On Various Epochs)

Epochs	RMSE
10	0.60
20	0.58
30	0.57
40	0.63

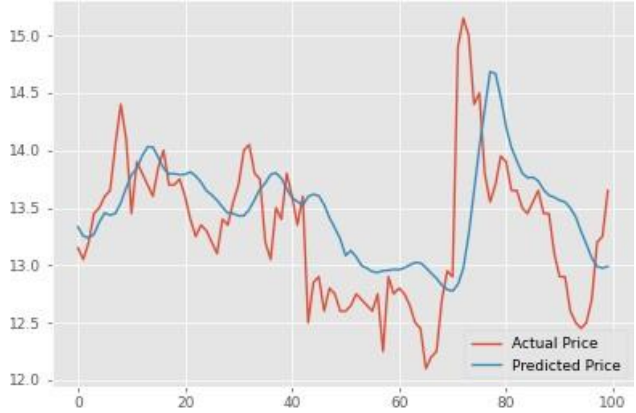


Figure V.6 :- LSTM Model Accuracy

- Linear Regression :-
- Linear Regression model was ran and got an RMSE of 1.04. Following is the graph plotted for Linear Regression Model

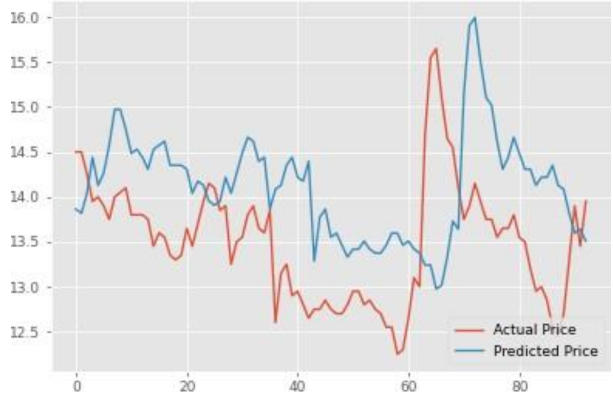


Figure V.7. : - Linear Regression Model Accuracy

- SENTIMENT ANALYSIS :-
- Taken the latest 20 tweets based on Apple Company.
- Calculated Polarity by using TextBlob and got Overall Polarity as Neutral.

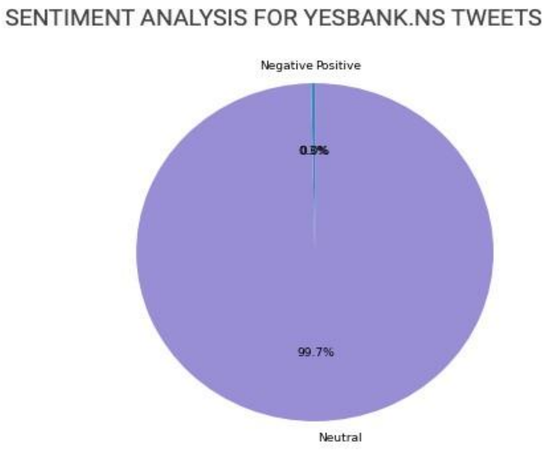
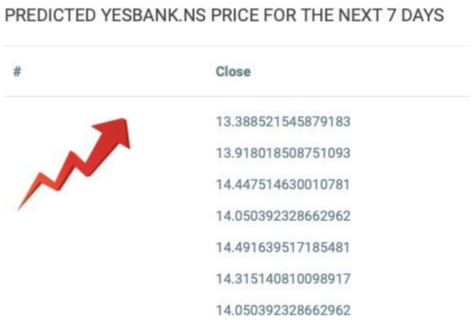


Figure V.8 :- Polarity From Tweets

- Predicted Price for Next 7 days :-

We can see that Price is decreasing



PREDICTED YESBANK.NS PRICE FOR THE NEXT 7 DAYS

#	Close
	13.388521545879183
	13.918018508751093
	14.447514630010781
	14.050392328662962
	14.491639517185481
	14.315140810098917
	14.050392328662962

VI. CONCLUSION AND FUTURE WORK

In recent years, most people have been seen investing in the stock market in order to make quick money. At the same time, an investor stands a good probability of losing all of his or her money. To comprehend future market trends, the user will need an effective predictive model.

Many prediction models exist that can anticipate whether the market is going up or down, but they are inaccurate. A model for predicting the stock market movement for the next day has been attempted. A model has been constructed and evaluated using diverse stock market data accessible open source, taking into account numerous patterns such as continuous up/down, volume traded each day, and also includes corporate sentiment.

On the considered dataset, LSTM and ARIMA model are performing best.

We have also performed sentiment analysis on twitter data to detect polarity of that particular tweet. Recommendation System is running well with the help of polarity of each tweet.

In Future we can make this research broader by predicting cryptocurrency prices. As cryptocurrency trading is the most volatile trading so, we can build a much more concentrated model which can focus more in depth for prediction.

VII. REFERENCES

[1]S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," in Advancesin neural information processing systems, NIPS, 1997, pp. 473—479.

[2]Y. Bengio, P. Simard, P. Frasconi and others, "Learning long-term dependencies with gradient descent is difficult," IEEE transactions on neural networks, vol. 5, no. 2, pp. 157-166, 1994.

[3]S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," in Advancesin neural information processing systems, NIPS, 1997, pp. 473--479.

[4]S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 6, no. 2, pp. 107-116, 1998.

[5]Manh Ha Duong Boriss Siliverstovs June 2006 - The Stock Market and Investment.

[6]Fahad Almudhaf, Yaser A. Alkulaib from Kuwait University - Are Civets Stock Markets Predictable?

[7]Pranav Bhat Electronics and Telecommunication Department, Maharashtra Institute of Technology, Pune.

[8]Savitribai Phule Pune University - A Machine Learning Model for Stock Market Prediction.

Mohammad Mekayel Anik, Mohammad Shamsul Arefin and M. Ali Akber Dewan, Department of Computer Science and Engineering -An Intelligent Technique for Stock Market Prediction

[9]Xiao Ding, Kuo Liao, Ting Liu, Zhongyang Li, Junwen Duan Research Centre for Social Computing and Information Retrieval Harbin Institute of Technology, China - Event Representation Learning Enhanced with External Common-sense Knowledge.

[10]H. Alostad and H. Davulcu, "Directional prediction of stock prices using breaking news on twitter," in 2015

[11]X. Li, H. Xie, Y. Song, S. Zhu, Q. Li, and F. L. Wang, "Does summarization help stock prediction? a news impact analysis," IEEE Intelligent Systems, vol. 30, no. 3, pp. 26–34, May 2015.

[12]J. Gong and S. Sun, A New Approach of Stock Price Prediction Based on Logistic Regression Model, In 2009. NISS '09. International Conference on New Trends in Information and Service Science, pp. 1366–1371, June (2009).

[13]J. Bean, R by example: Mining Twitter for consumer attitudes towards airlines, In Boston Predictive Analytics Meetup Presentation

[14]R. Kotikalapudi, "Keras Visualization Toolkit," [Online]. Available: https: //raghakot.github.io/keras-vis. [Accessed 31 May 2019].

BIOGRAPHIES

KUNAL GAUR
Department of Information Technology Guru Tegh Bahadur Institute Of Technology
Delhi, India

SAUD AKHTAR
Department of Information Technology Guru Tegh Bahadur Institute Of Technology
Delhi, India

KRISHVI SRIVASTAVA
Department of Information Technology Guru Tegh Bahadur Institute Of Technology
Delhi, India

KAMALJYOT SINGH
Department of Information Technology Guru Tegh Bahadur Institute Of Technology
Delhi, India

GAURAV SANDHU
Department of Information Technology Guru Tegh Bahadur Institute Of Technology
Delhi, India