Research Article

Stock Prediction by using NLP and Deep Learning Approach

Rushali Deshmukh[1], Yogesh Bodkhe[2]

[1]Savitribai Phule Pune Univresity, faculty, Maharhatra/India
[2]Savitribai Phule Pune Univresity, PG research Scholar, Maharhatra/India

Article History: **Received**: 11 november 2020; Accepted: 27 December 2020; Published online: 23 January 2021; Published online: 05 April 2021;

Abstract: People have a tendency to analyze existing strategies and so planned new strategies for inventory prediction. We have used Sentiment evaluation and Technical evaluation through NLP and Deep mastering approach. In order to exploit benefits of sentiment analysis on enterprise associated inventory, we have proposed a model that will use the sentiment analysis on twits associated with special sectors that are Information Technology sector, Banking sector, Pharmaceutical sector, Automobile sector, Infrastructure sector which are extracted from twitter. These twits are extracted from twitter for calculating polarity. The rating of sentiment analysis is calculated here by using Natural Language Processing's method. According to sector we've taken five groups. Top four performer businesses of every sector. Using polarity score we got finalized pinnacle ten groups with great sentiment rating. We then downloaded the CSV facts of historical share charge of top ten organizations that we've selected. Then downloaded CSV records are used to build a CNN version to predict in addition stock movement of these pinnacle ten companies.

Keywords: Stock prediction, Natural Language processing, Deep Learning, Price forecasting, sentiment analysis, CNN, SVM, ANN, NSE, BSE, financial market, Data mining.

1. Introduction

Financial analysts investing in stocks usually, but they are not aware about the inventory market place conduct. They are going through the problem of trading as they do not properly recognize which shares to shop for or which shares to promote with a purpose to get greater profits. In today's world, all the information pertaining to inventory market is available. Analyzing all these records in my opinion or manually is pretty much difficult. As such, automation of the method is required. This is where Data mining techniques help. Understanding that analysis of numerical time series offers close results, wise traders use system learning techniques in predicting the inventory market conduct. This will allow financial analysts to foresee the conduct of the inventory that they may be interested by and consequently act accordingly. The input to our gadget will be historical data. Appropriate data could be carried out to discover the stock fee trends. Hence the prediction value will notify the up or down of the stock price movement for the buying and selling of stocks and traders can act upon it so one can maximize their chances of gaining a profit.

1.1. Problem Statement

- Deep learning is an AI work that mirrors the activities of the human mind in preparing information for use in basic leadership.
- Deep learning can learn from the data that is from both unstructured and unlabeled source.
- Deep learning is machine learning's subset that can be used to help to detect fraud or money laundering.

2. Literature Survey

Rakhi Batra et al.[1] had used a technique of sentiment analysis for stock tweets, which was related to a different type of Apple product; for this, she had extracted stocks related twits from different social networking sources for eight years of time duration. Apart from shares data, these people had decided to use stocks related data from the Yahoo Finance source for that time duration. They had used the SVM technique to find out the polarity of those tweets. So because of this, they were able to differentiate the tweets as Positive or Negative. After that, they had used polarity results, which was based on sentiment analysis and stock data, to create an SVM model and to forecast a subsequent day's share price movements.

Yaojun Wang et al. [2] used social media sites to gather data for their research. In this research paper, their focus was on the share price movements in the market. For the forecasting of the stocks along with mining techniques, they had used other relevant information. Their result showed that they had calculated the stocks' polarity for better prediction of the stock's price.

Ashish Sharma et al. [3] had gone through the stock market data in regressive manner. So that they got a good amount of stock data for their research from the share market. The motto of their research study was to help the stockbrokers and investors for investing money in the stock market.

Ze Zhang et al. [4] had used one system to find out the opening value of stocks in the financial market. However, their developed system was self-learner so that they were able to predict the opening value of the market. They had given the stocks data to their developed system to find out the forecasted value. Last, they developed another network system and compared both the system with each other to predict the starting day value of the stock.

Dev Shah et al. [5] had studied the news and on the basis of that news they had do the sentiment analysis. In this paper they had find out the polarity for pharma sector. Mainly their focus was on the stocks prices movement which was based on the polarity dictionary.

Du Peng [6] had mainly studied the market volatility and the worked on the sentiments of the people to find out the relation between share price and the sentiments of the traders. For this he had studied different indexes for the news.

Muhammad Firdaus et al. [7] had used ANN algorithm for the share market prediction. On the basis of their studies of ANN they had claimed that they had achieved the high percentage of accuracy while predicting the values in the stock market. For this they had studied different methods and after studying that they had find out the accurate and proper results.

Research work of Nonita Sharma et al. [8] had given focus on to do prediction of the share prices by studying the historical data. For that they had taken decade data from the two well know indexes like NSE and BSE. For this they had developed model by using SVM algorithm. For the predicted value they had considered the closing value of the share. Also, they had forecasted the share values around more than 35 days.

3. Proposed Methodology

Here we tend to area unit planned System that is working with Improved level of recommendation. System is developed with Natural Language Processing (NLP) technique of computer science and Convolutional Neural Network (CNN) of Deep Leaning. Natural Language Processing technology is used facilitate system to search out companies with excellent news in terms of live performance in market. That helped to facilitate to create selection of best performer in market. NLP is used to classify news in positive and negative sets and to provide performance graph of selected organization. Supported to that we got to know the best performing company. Natural Language Processing provides to system NLP (Natural Language Processing) that worked on our twits for detection merchandise and unhealthy of its impact.

3.1 Architecture

The proposed system has the advantage of multiple platforms for input files for model development. The system has been developed with over one algorithmic rule; thence Prediction guarantees are magnified. Live updates area unit concerned in prediction thence it is often used for live recommendation.

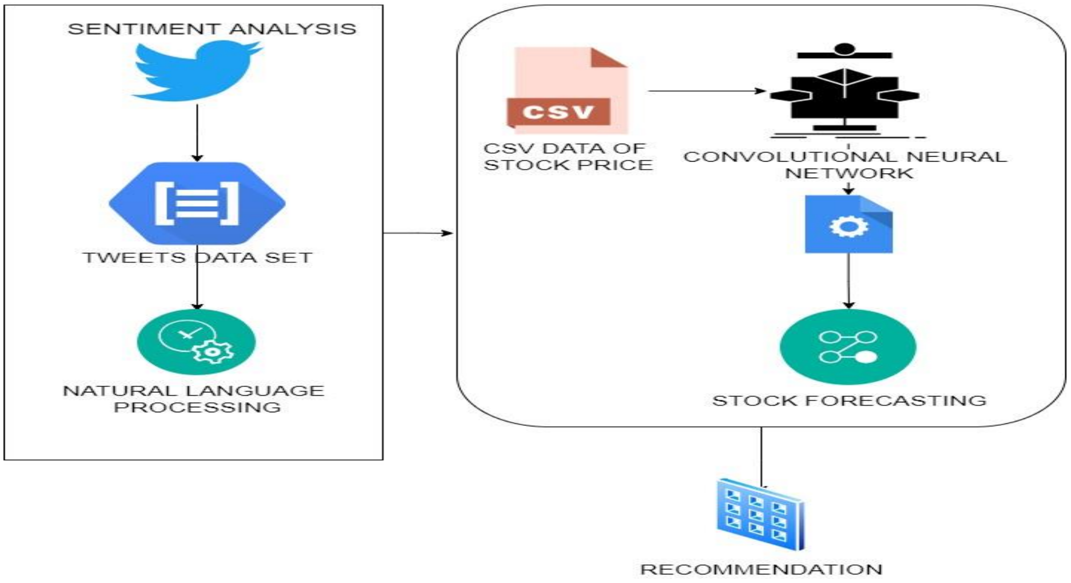


Figure 1. Proposed System Architecture

3.2 Algorithms

1D CNN and VADAR Sentiment analysis algorithms are used.

3.2.1 CNN algorithm

First 1D CNN layer

The primary layer defines a filter (or conjointly known as feature detector) of height ten (also known as kernel size). Solely shaping one filter would enable the neural network to be told one feature within the initial layer. This won't be comfortable so that we will outline N variety of filters. This permits North American country to coach N completely different options on the primary layer of the network. The output of the first neural network layer is in a very somatic cell-matrix type.

Second 1D CNN layer

The outcome from the first CNN will be sustained into the second CNN layer. We will again characterize 100 unique channels to be prepared on this level. Following a similar rationale as the primary layer, the yield grid will be of size 62 x 100.

Max pooling layer

A pooling layer is frequently utilized after a CNN layer so as to lessen the unpredictability of the yield and forestall over fitting of the information. In our model, we picked a size of three. This implies the size of the yield lattice of this layer is just 33% of the info network. Third and fourth 1D CNN layer: Another arrangement of 1D CNN layers follows to learn more significant level highlights. The yield lattice after those two layers is a 2 x 160 grid.

Average pooling layer

One all the more pooling layer to additionally abstain from over fitting. This time not the most extreme worth is taken but rather the normal estimation of two loads inside the neural system. The yield lattice has a size of 1 x 160 neurons. Per include identifier, there is just one weight staying in the neural system on this layer.

Dropout layer

The dropout layer will haphazardly dole out 0 loads to the neurons in the system. Since we picked a pace of 0.5, half of the neurons will get a zero weight. With this activity, the system turns out to be less touchy to respond to littler varieties in the information. Along these lines, it should additionally expand our precision on concealed information. The yield of this layer is as yet a 1 x 160 network of neurons. Fully connected layer with SoftMax activation: The last layer will decrease the vector of stature 160 to a vector of six since we need to anticipate ("Jogging", "Sitting", "Strolling", "Standing", "Upstairs", "Ground floor") because we have six classes. Another lattice augmentation finishes this decrease. Softmax is utilized as the initiation work. It powers every one of the six yields of the neural system to summarize to one. The yield worth will hence speak to the likelihood for every one of the six classes.

Sentiment Intensity analyzer

Business: In advancing field firms use it to build up their strategies, to know clients' sentiments towards product or entire, people answer their battles or item dispatches and why customers don't get some product [1] [5].

1D CNN Algorithm

The Algorithm of a 1D-CNN is formed through the following important steps:

```
Input: Dataframe (train_data , test_data)
Process: Build 1D CNN Model
def Model():
        define model
        add filter (kernel) size to each layer
        model.add(layers)
        model.add(kernel size)
        add pooling layer
        add dropout value
        model activation layer
        fit model with training and testing data
 Model summary
Output: Prediction = model.predict(test data)
Accuracy  = (accuracy_score(Y_test,Y_pred)*100)
```

VADER Sentiment Analysis

VADER content opinion investigation utilizes a human-driven methodology, consolidating substance examination, and exact approval by exploiting human raters and hence the information on the gathering.

Five Easy Heuristics

1. Lexical alternatives aren't the sole things inside the sentence that affect the opinion. Their territory unit elective talk segments, similar to accentuation, capitalization, and modifiers that conjointly give feeling. VADER's conclusion examination thinks about these by thinking about five simple heuristics. The after effect of those heuristics zone unit, once more, measured exploitation human raters.

For Ex.

[1] I Like that. [2] I Like that!!!

2. VADER assumption examination mulls over this by enhancing the sentence's slant score relative to the quantity of shout focuses and question marks finishing the sentence. VADER first figures the estimation score of the sentence. If the score is certain, VADER includes a specific experimentally acquired sum for every accentuation mark (0.292) and accentuation (0.18). On the off chance that the score is negative, VADER subtracts.

The second heuristic is capitalization.

[1] amazing work.

[2] AMAZING work.

Thus, VADER takes this under consideration by incrementing or decrementing the word's sentiment score by zero.733, betting on whether or not the word is positive or negative, severally.

3. The third heuristic is that the use of degree modifiers. View example “effing cute” and “sort of cute”. The modifier's result within the 1st sentence is to extend the intensity of cute, whereas within the second sentence, it's to decrease the intensity. VADER maintains a booster wordbook that contains a collection of boosters and

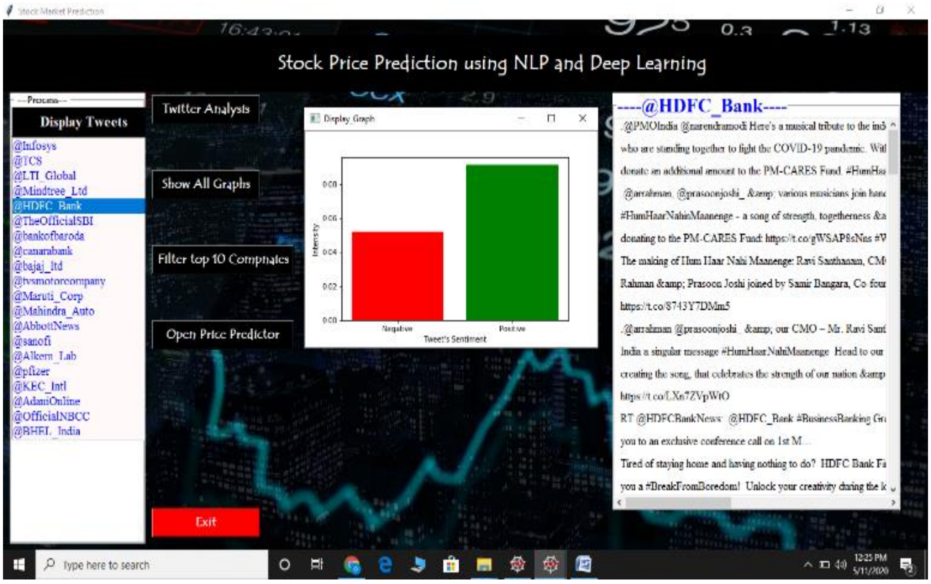
dampeners. The result of the degree modifier conjointly depends on its distance to the word it’s modifying. Farther words have a comparatively smaller exacerbating result on the bottom word. One modifier adjacent to the base word adds or subtracts zero.293 to the slant score of the sentence, wagering whether the base word is sure. A second modifier from the base word includes/subtract ninety-fifth of zero.293, and a third includes/subtracts ninetieth.

4. The fourth heuristic is that the shift in polarity thanks to “but”. In many cases, "but" interfaces 2 provisions with contrastive conclusions. The prevailing opinion, in any case, is that the last one. for example, " I like it, but I don’t wish to use that anymore " the essential provision "I like it" is sure, the other VADER actualizes a "but" checker. Fundamentally, all conclusion bearing words before the "but" have their valence decreased to five hundredth of their qualities, though those when the "but" increment to a hundred and fiftieth of their qualities.

5. The fifth heuristic is looking at the tri-gram before a feeling loaded lexical component to get extremity nullification. Here, a tri-gram alludes to an assortment of 3 lexical choices. VADER keeps up a stock of useless words. Refutation is caught by increasing the opinion score of the assessment loaded lexical component by partner degree experimentally decided cost - 0.74.

4. Results and Discussions

Figure 2 displays the base window of proposed system, where it contains control panel with buttons which are bind with events to load tweets from twitter according to selected company. Scraped tweets are displayed in canvas window at the right side of base window. Scraped tweets are being displayed in text form in the canvas frame. So here we can say that user will be able to display all live tweets of selected company.

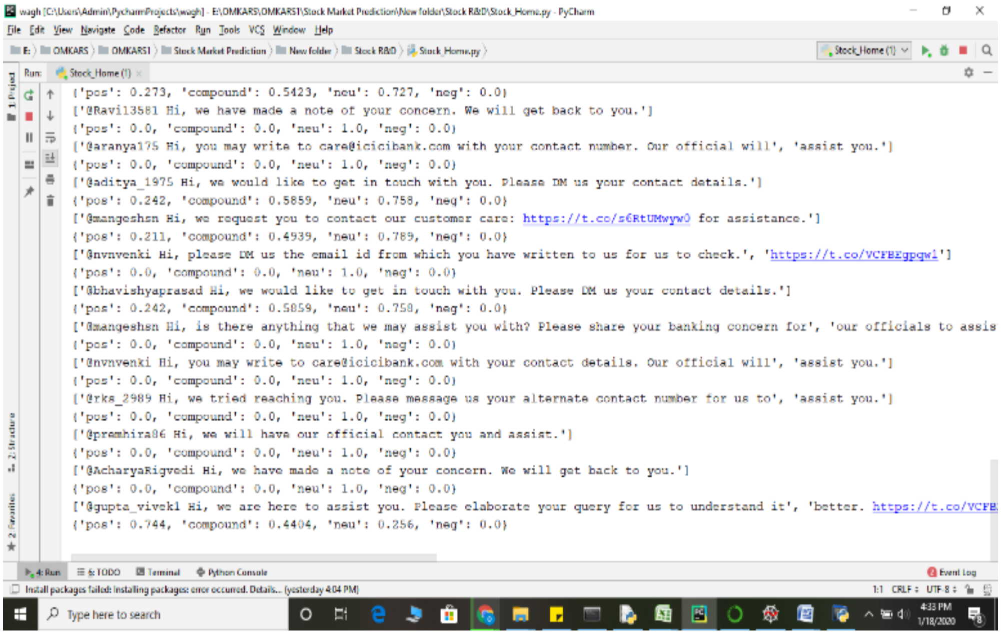


----@HDFC_Bank----

.@PMOIndia @narendramodi Here's a musical tribute to the ind
who are standing together to fight the COVID-19 pandemic. Wil
donate an additional amount to the PM-CARES Fund. #HumHa
@arrahman, @prasoonjoshi_ &amp; various musicians join hand
#HumHaarNahinMaanenge - a song of strength, togetherness &a
donating to the PM-CARES Fund: https://t.co/gWSAP8sNns #V
The making of Hum Haar Nahi Maanenge: Ravi Santhanam, CM
Rahman &amp; Prasoon Joshi joined by Samir Bangara, Co-fou
https://t.co/8743Y7DMm5
.@arrahman @prasoonjoshi_ &amp; our CMO – Mr. Ravi Sant
India a singular message #HumHaarNahiMaanenge. Head to our
creating the song, that celebrates the strength of our nation &amp
https://t.co/LXn7ZVpWtO
RT @HDFCBankNews: .@HDFC_Bank #BusinessBanking Gr
you to an exclusive conference call on 1st M…
Tired of staying home and having nothing to do? HDFC Bank Fa
you a #BreakFromBoredom! Unlock your creativity during the k

Figure 2. Display Tweet window

So, for our project we have taken HDFC bank’s official twitter handler’s tweets as an example.


wagh [C:\Users\Admin\PycharmProjects\wagh] - E:\OMKARS\OMKARS1\Stock Market Prediction\New folder\Stock R&D\Stock_Home.py - PyCharm
File Edit View Navigate Code Refactor Run Tools VCS Window Help
E: OMKARS OMKARS1 Stock Market Prediction New folder Stock R&D Stock_Home.py
Stock_Home (1)
Run: Stock_Home (1)
1: Project
2: Structure
2: Favorites
4: Run
6: TODO
Terminal
Python Console
Event Log
1:1 CRLF UTF-8
Type here to search
4:33 PM
1/18/2020

```
{'pos': 0.273, 'compound': 0.5423, 'neu': 0.727, 'neg': 0.0}
['@Ravi13581 Hi, we have made a note of your concern. We will get back to you.']
{'pos': 0.0, 'compound': 0.0, 'neu': 1.0, 'neg': 0.0}
['@aranya175 Hi, you may write to care@icicibank.com with your contact number. Our official will', 'assist you.']
{'pos': 0.0, 'compound': 0.0, 'neu': 1.0, 'neg': 0.0}
['@aditya_1975 Hi, we would like to get in touch with you. Please DM us your contact details.']
{'pos': 0.242, 'compound': 0.5859, 'neu': 0.758, 'neg': 0.0}
['@mangeshsn Hi, we request you to contact our customer care: https://t.co/s6RtUMwyw0 for assistance.']
{'pos': 0.211, 'compound': 0.4939, 'neu': 0.789, 'neg': 0.0}
['@nvnvenki Hi, please DM us the email id from which you have written to us for us to check.', 'https://t.co/VCFBEgpqwl']
{'pos': 0.0, 'compound': 0.0, 'neu': 1.0, 'neg': 0.0}
['@bhavishyaprasad Hi, we would like to get in touch with you. Please DM us your contact details.']
{'pos': 0.242, 'compound': 0.5859, 'neu': 0.758, 'neg': 0.0}
['@mangeshsn Hi, is there anything that we may assist you with? Please share your banking concern for', 'our officials to assis
{'pos': 0.0, 'compound': 0.0, 'neu': 1.0, 'neg': 0.0}
['@nvnvenki Hi, you may write to care@icicibank.com with your contact details. Our official will', 'assist you.']
{'pos': 0.0, 'compound': 0.0, 'neu': 1.0, 'neg': 0.0}
['@rks_2989 Hi, we tried reaching you. Please message us your alternate contact number for us to', 'assist you.']
{'pos': 0.0, 'compound': 0.0, 'neu': 1.0, 'neg': 0.0}
['@premhira86 Hi, we will have our official contact you and assist.']
{'pos': 0.0, 'compound': 0.0, 'neu': 1.0, 'neg': 0.0}
['@AcharyaRigvedi Hi, we have made a note of your concern. We will get back to you.']
{'pos': 0.0, 'compound': 0.0, 'neu': 1.0, 'neg': 0.0}
['@gupta_vivek1 Hi, we are here to assist you. Please elaborate your query for us to understand it', 'better. https://t.co/VCFB
{'pos': 0.744, 'compound': 0.4404, 'neu': 0.256, 'neg': 0.0}
```

```
Install packages failed: Installing packages: error occurred. Details... (yesterday 4:04 PM)
```

Figure 3. Polarity Result

Figure 3. displays score of polarity on tweets which were scraped from twitter. VADER library that supported our proposed system in terms of getting sentiment of tweets has given the polarity report with Positive, Compound, Neutral and Negative in the form of percentage of each sentiment. So according to polarity score we will get an idea about the positive and negative tweets.

Figure 4. and Figure 5. are screen shot of Sentiment Intensity of tweets per company (Red bar – Negative Tweets, Green bar – Positive Tweets). All tweet data is passed through Sentiment intensity analyzer and Positive and Negative values are calculated. Summing all positive and negative values bar graph is plotted for both values. To filter out top ten best performing company's mathematical logic used is ratio of positive value sum with negative value sum is calculated. As per ratio companies are arranged in descending order to get top performers

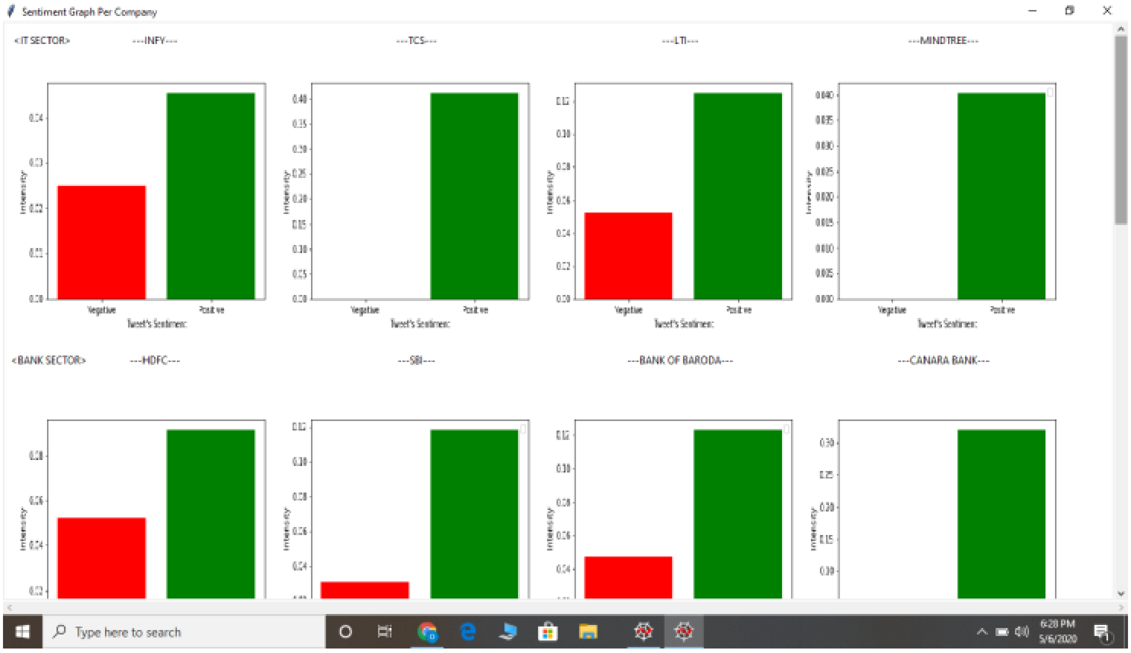


Figure 4. Polarity Graph (a)

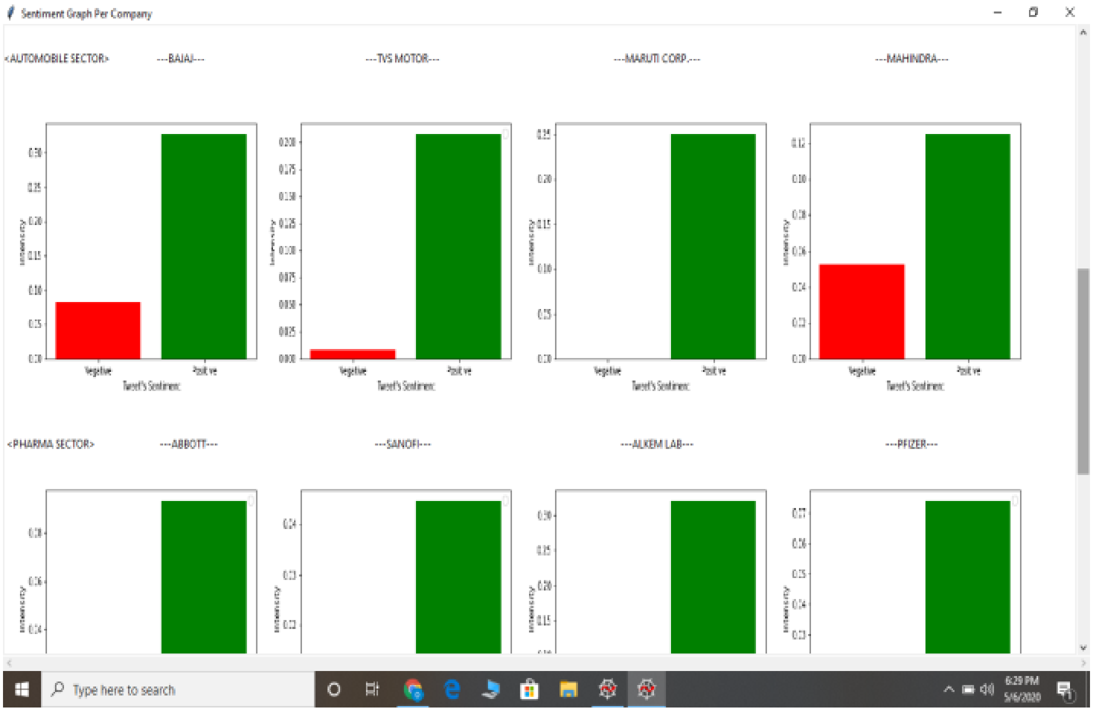


Figure 5. Polarity Graph (b)

Figure 6. plotted graph with Actual prices and Predicted prices by algorithm (Green line – Actual Prices, Blue line – Predicted Prices).

Actual closing price and predicted closing price by proposed mathematical algorithmic model are plotted on single graph to visualize the comparison of model prediction with actual value data. As result graph reflects that both predicted and actual values i.e. lines are moving in parallel manner which means prediction is correct with approx. accuracy. Data of fifty is plotted on graph in which initial section shows that both predicted and actual values are closer to each other. As lines move further difference between lines is increased but predicted value line is following in parallel manner to actual value line.

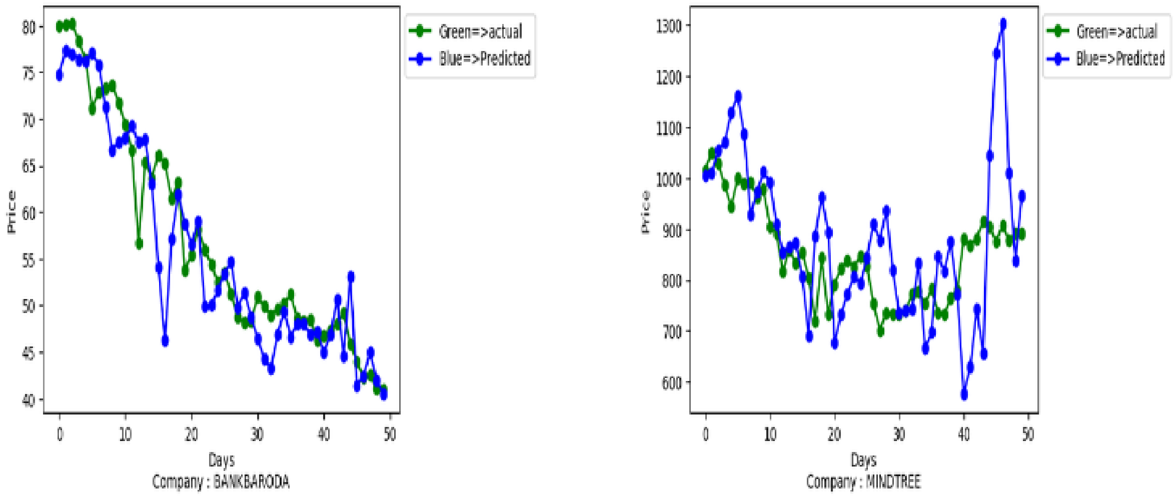


Figure 6. Actual vs. Predicted value graph

Figure 7. is the image of graphs of top first company of each sector with forecasted prices by algorithm. It shows the image of top five companies of each Information Technology, Banking, Automobile, Pharmaceutical and Infrastructure sectors. Each company is filtered out from total four companies on the basis of tweet sentiment of each company and forecasting of stock prices of each company.

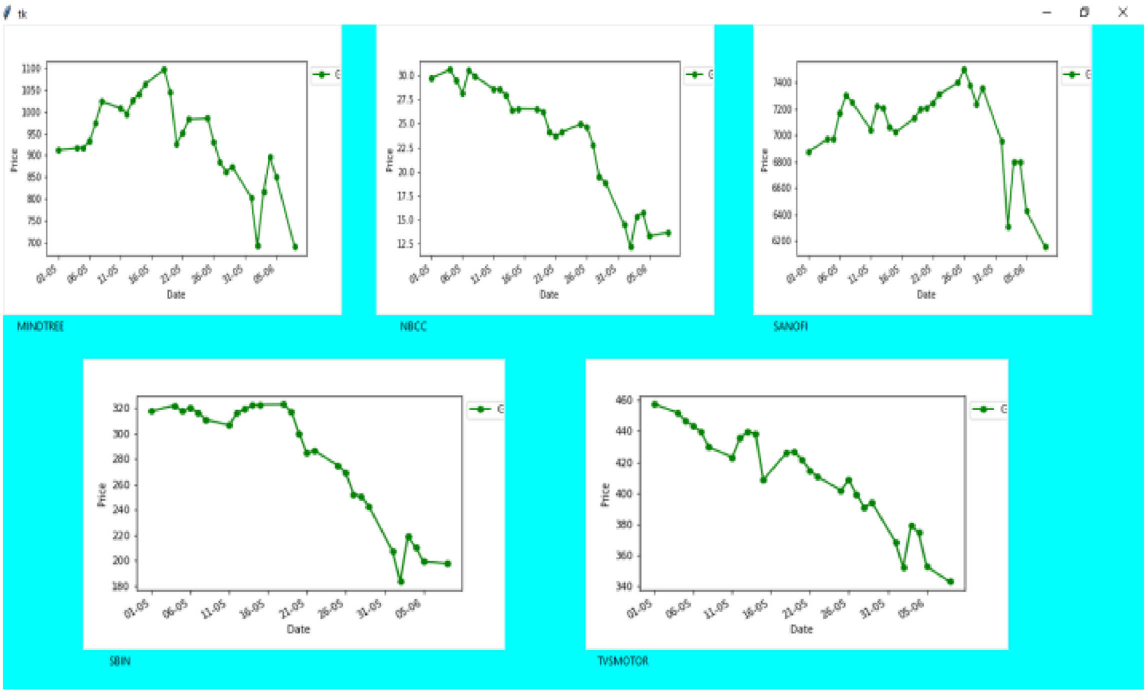


Figure 7. Forecasted value graph of top five companies of each sector

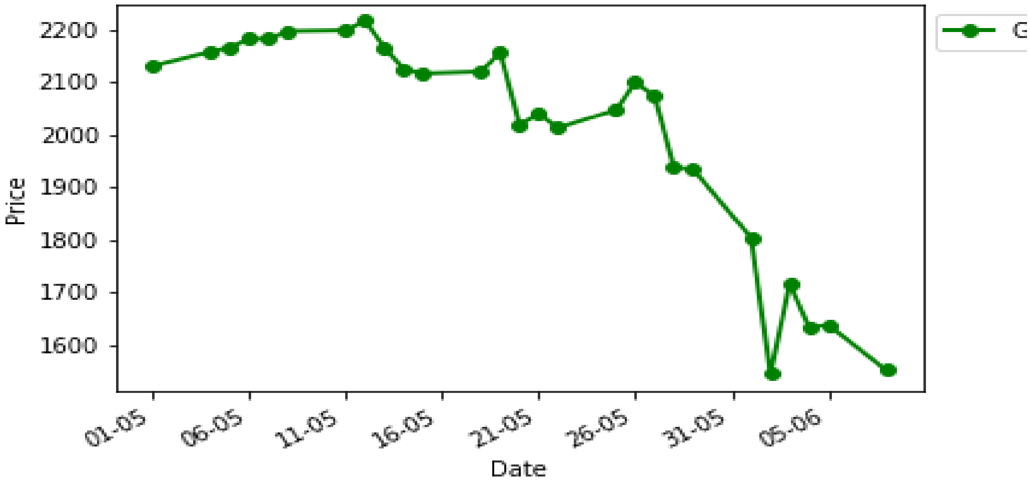


Figure 8. Forecasted value graph of single company

Figure 8 is the graph of forecasted values of single company. On Y axis prices are plotted and on X axis dates are plotted. For securing forecasted value data from misleading depending on dates. As share market remains closed on Saturday and Sunday dates of such days are excluded while plotting forecasted values.

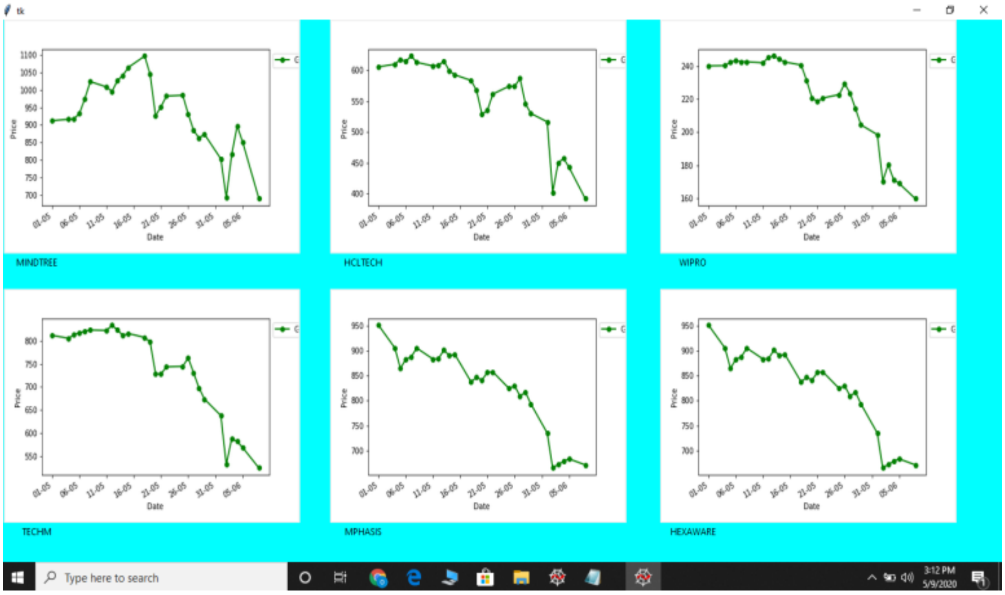


Figure 9. Forecasted value graph of other five IT sector companies which are not previous twenty company list

In Figure 9 we have extended the forecasting facility towards IT sector in terms of adding contribution in sector wise stock analysis. We collected price value data of other five companies of IT sector which companies are not included in previous twenty company list. As we got top performer of each sector, top performing company of IT sector is shown in comparison with newly selected five companies. In practical this will add more benefit to a stock investor to analyze stock performance on focused sector.

5. Conclusion

In an early research on stock prediction were totally based on random walks, machine learning, numerical prediction and support vector machine but with the introduction of behavioral finance, the people's literacy about market were considered while predicted about stock movement. Making it more efficient we used the idea of sentiment analysis of Stock Tweets through NLP technology. We implemented the idea by collecting sentiment data and stock price data and built a CNN model for prediction and forecasting, also in the last we measured the prediction accuracy. Results showed that we have achieved a polarity and based on this polarity we measured top ten well performing companies in given sectors.

Also putting focus on IT sector, we downloaded the data of other five top performing companies in same sector which are not included in our twenty-company list. Price values data of all five companies again passed through model to compare the top performer of IT sector as per our model with other companies. This is to make more contribution in benefit of sector wise stock analysis. Practically investors can visualize the forecasting and future performance of higher end IT sector in stock market. In future we will attempt to execute more calculations and all the newer methods planning to give live proposal to securities exchange financial specialists. Additionally, our emphasis will be on entire securities exchange for forecasting.

References

1. Batra, Rakhi, and Sher Muhammad Daudpota. (2018) "Integrating StockTwits with sentiment analysis for better prediction of stock price movement." In 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1-5. IEEE.
2. Wang, Yaojun, and Yaoqing Wang. (2016) "Using social media mining technology to assist in price prediction of stock market." In 2016 IEEE International Conference on Big Data Analysis (ICBDA), pp. 1-4. IEEE.
3. Sharma, Ashish, Dinesh Bhuriya, and Upendra Singh. (2017) "Survey of stock market prediction using machine learning approach." In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), vol. 2, pp. 506-509. IEEE.
4. Zhang, Ze, Yongjun Shen, Guidong Zhang, Yongqiang Song, and Yan Zhu. (2017) "Short-term prediction for opening price of stock market based on self-adapting variant PSO-Elman neural network." In 2017 8th IEEE International Conference on Software Engineering and Service

Science (ICSESS), pp. 225-228. IEEE.

5. Shah, Dev, Haruna Isah, and Farhana Zulkernine. (2018) "Predicting the Effects of News Sentiments on the Stock Market." In 2018 IEEE International Conference on Big Data (Big Data), pp. 4705-4708. IEEE.
6. Peng, Du. (2019) "Analysis of Investor Sentiment and Stock Market Volatility Trend Based on Big Data Strategy." In 2019 International Conference on Robots & Intelligent System (ICRIS), pp. 269-272. IEEE.
7. Firdaus, Muhammad, SwelandiahEndahPratiwi, Dionysia Kowanda, and Anacostia Kowanda. "Literature review on Artificial Neural Networks Techniques Application for Stock Market Prediction and as Decision Support Tools." In 2018 Third International Conference on Informatics and Computing (ICIC), pp. 1-4. IEEE, 2018.
8. Sharma, Nonita, and Akanksha Juneja. (2017). "Combining of random forest estimates using LSboost for stock market index prediction." In 2017 2nd International Conference for Convergence in Technology (I2CT), pp. 1199-1202. IEEE.