International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Hybrid Deep Learning Based Stock Market Prediction with both Sentiment and Historic Trend Data

Guruprasad S, Sahilverma, H Chandramouli

Abstract: ***Stock market is highly volatile and it is necessary for investors to have an accurate prediction of stock prices for a better profitability. Towards this need many methods have been proposed for stock market prediction with aim to provide a higher prediction accuracy. Current methods for stock market prediction are in two categories of machine learning and statistics based. Considering the need for accurate prediction in short term and long term, the merits of both methods must be combined for accurate prediction. This work proposes a hybrid deep learning approach for stock market prediction which combines the historic price-based trend forecasting along with stock market sentiments expressed in twitter to predict the stock price trend.***

Keywords: machine learning, statistics based.

I. INTRODUCTION

Stock exchanges are the financial institutions which allow the people to sell and buy goods (monetary values, actions, precious metals) between stockbroker components. With the trade of getting more than billions of dollars, it attracts the people to invest and trade with the amount of money an individual is holding. The goods are traded on the market, following their subsequent values the profit or gain is defined. In general, more the goods get traded in profit volume, more the investors make money out of it. And the purpose of any investor is to grasp the opportunity for maximum profit. But, if the trade is made in negative, this generates a loss, to the company listed and to the investors.

Thus, it becomes necessary to predict the stock prices to maximize the profitability. The volatile moment of the market and the stock can be predicted considering the main elements through which it works, which can be determined as

1) The past performance of a company.
2) The sentiment of the people, and
3) The business happening in the market.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Mr. Guruprasad S*, Assistant Professor, Department of Computer Science and Engineering. BMS Institute of Technology &management, Bangalore, India. Email: guruprasad@bmsit.in
Mr. Sahilverma, Software Development Engineer Trainee, Tally Solutions PVT Ltd. Bangalore, India, Email: sahilverma0696@gmail.com
Dr. Chandramouli H, Professor, Department of Computer Science and Engineering. East Point College of Engineering and Technology, Bangalore, India. Email:hcmcool123@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

Therefore, the problem statement becomes: for a given stock, based on its history, the sentiments people are holding for it and the current business deals happening in the market. How can one select profitable listing over a period of time? An aspect of the research has gained a lot of attention in the recent past about predicting the future price of the market. The domain of Artificial Intelligence, mainly Machine learning have tried to predict the same, such as ARIMA (Auto Regressive Integrated Moving Average), CBR (Case-Based Reasoning), SVR (Support Vector Regressor) and many others.

But due to the non-linear and ever-changing trend of the market, researches are yet to have promising answers. The machine learning models provided tries to find out a specific pattern in a given dataset and follows to implement the same. Giving them a restriction on decision making on volatile terms. By which one can understand the inception of predicting the values in not just classification or regression based, but more than that, having involvement of both the domains. Optimizing the decisions of selection (classification) and flexible pattern matching to the volatility of prediction (regression). From a financial point of view, the traders, the investors and the brokers do understand the ever-changing behavior of the market.

The best thing to do is to analyze the market and figure out the best stocks to invest to, and the best timings of buying/selling the stocks. But the experts also make bad decisions choosing the right stock. The decision taken by an early investor or a rookie can be as good as an expert decision in the selection of a proper stock. There are various mathematical models provided known as technical indicators which also provides insights on making decisions. The features used for prediction of historical trends use OPEN, HIGH, LOW and CLOSE, from which further derived features are created.

In this work, a hybrid deep learning approach for stock market prediction is proposed. A deep learning model using LSTM is constructed with historical prices to predict the growth or fall of the stock price. A Neural network-based twitter sentiment analysis is proposed to decide the increase or decrease of prices. The results of LSTM and Neural prediction is combined to provide the final decision about the stock performance. The proposed hybrid deep learning approach is tested for Indian stock markets in this work, but the method can be used for any stock markets, bullions and forex.

Literature Survey

In [1] author surveyed the regression methods for prediction of stock market. The benefit is regression was done on multiple variables instead on time variant stock prices. But the accuracy was limited in these approaches. Authors in [2] created clusters of stocks based on their high or low trends using K-means clustering algorithm. The method could not cluster based on the amount of increase or decrease, so was limiting in its applications In [3] authors used machine learning algorithms to predict trends in stock market mainly Random Forest Model and Support Vector Machine (SVM). The Random Forest model is a collaborative learning model which yield high success rate in classification and regression. The SVM is a classification technique. The combination of these both can predict whether the price of stock will be high compared to today, based on historical prices. Also, the study was conducted with limited data. In [4] a sentiment analysis of stock market and Public opinion was used to find the relation between public emotions and the stock market movement. Twitter data is used for anticipating past stock prices with public opinion to predict future movements of stock market. Authors in [5] examined a predictive AI approach to analyses market news articles through various text analysis techniques such as Bag of Words, Noun Phrases, and Named Entities. Through this methodology, they researched 9,211 news stories and 10,259,042 stock price movements including the S&P 500 stocks in a time frame of five-weeks. The study revealed the discrete price movement within twenty minutes of release of news article. A customized model was built containing SVM for discrete numeric prediction and different stock quotes shown that the model containing news article words at the time of release of news article gave the closest prediction of the stock prices. Long short-term memory (LSTM) algorithm based stock market anticipation is proposed in [6]. Authors found a unique pattern amongst the trending stocks that these stocks have high volatility and show reversal of trend in short-term. Taking advantage of these findings they proposed a rules-based short-term reversal strategy. The Authors have applied deep learning models, Paragraph Vector, and LSTM for time series prediction of stock prices in [7].

The decision of the retail and corporate investors depends on many factors such as Consumer Price Index (CPI), Price Earnings (PE) ratio and the news events reported. The Author used Paragraph Vector to convert news articles to distributed representations and modelled the time based impact of these events on opening price of stocks of various companies using LSTM. The news which has high impact on the market more efficient filtering mechanisms. Author proposed planar feature representation model and Deep Convolutional Neural Network (CNN) for prediction of stock prices in [8]. The historical prices of stock are represented as time series and features extracted from it are used for deep convolutional neural networks. Author proposed a deep learning method, based on CNN that anticipates the price fluctuation of stocks, using as time-series, high-frequency, and large-scale input, obtained from order book of stock exchanges in [9]. The approach can only predict short time price movements. Authors propose a hybrid model for prediction of returns in stock [10]. The proposed model is contains 2 linear models: 1) autoregressive moving average model, 2) exponential smoothing model IT ALSO CONTAIN ONE non-linear model which is recurrent neural network. A new regression model generates the training data needed for recurrent neural network. Compared to linear models the recurrent neural network capable of producing much accurate predictions. The proposed hybrid method merges the predictions obtained from all three prediction models so as to improve the prediction accuracy. Optimal weights are introduced for optimization and use of genetic algorithms better the prediction.

Use of LSTM for stock market anticipation based on technical analysis indicators is explored in [11]. The method provides only a decision of increase and decrease and does not provide the price of the stock. CNN was used to predict the stock price variation of Chinese market in [12]. Opening, High, Low, Closing price (OHLC) and volume traded of trade are used as input to CNN. Using only historical prices is not enough for accurate prediction of prices. A novel recurrent CNN for predicting trend in stock market is proposed in [13]. The proposed model can obtain the useful cues from news of financial market automatically without human intervention which is accomplished by Entity embedding layer which uses news articles about the market. The Convolutional layer is responsible for eliciting the key information that have impact on the market, it uses LSTM networks to learn about the association between the news and the market movement prediction.

This approach is breached against the errors caused due to outlier in the data. An approach of deep learning is used in [14] for market prediction, where a dense vector is created from the text extracted from the news and is trained using Neural Tensor Network, Next deep CNN is applied to know the influence of long and short term price fluctuations.

Though the approach is able to identify the events affecting the stock market prices, it cannot quantify the influence. Recurrent neural networks with character-level language model pre-training for both intraday and intera-day stock market forecasting is proposed in [15]. Financial news is used as predictors for stock market prediction. The approach cannot provide quantitative estimation of stock prices from the news forecast and especially with diverse feeds from different market news, it is very difficult to predict the trend using financial news forecast alone

II. HYBRID DEEP LEARNING STOCK PREDICTION

The proposed hybrid deep learning model for stock prediction is based on integration of two major models.

1. Sentiment analysis
2. LSTM model on historical data

The results of these model are clustered and the final output decision is made. The architecture of the model is given in Figure 1.

A. Sentiment Analysis

Sentiment Analysis is a major application in Natural Language Processing which deals with recognizing the human emotions,

through understanding the sentiments on its elemental basis. Recurrent Neural Network (RNN) is used for sentiment analysis in this work. RNN is used for sequential data. Most RNN can process sequences of variable length. Neural Network structure of RNN is used as sentiment classifier to classify the tweets into positive or negative. We have used a feed forward neuron classifier. Perceptron is the individual unit of NN,

each one responsible to make a decision on specific threshold.

Perceptron takes input of real – valued inputs, calculates the linear combination of input, and gives output of either 1 or 0, based on sigmoid function.

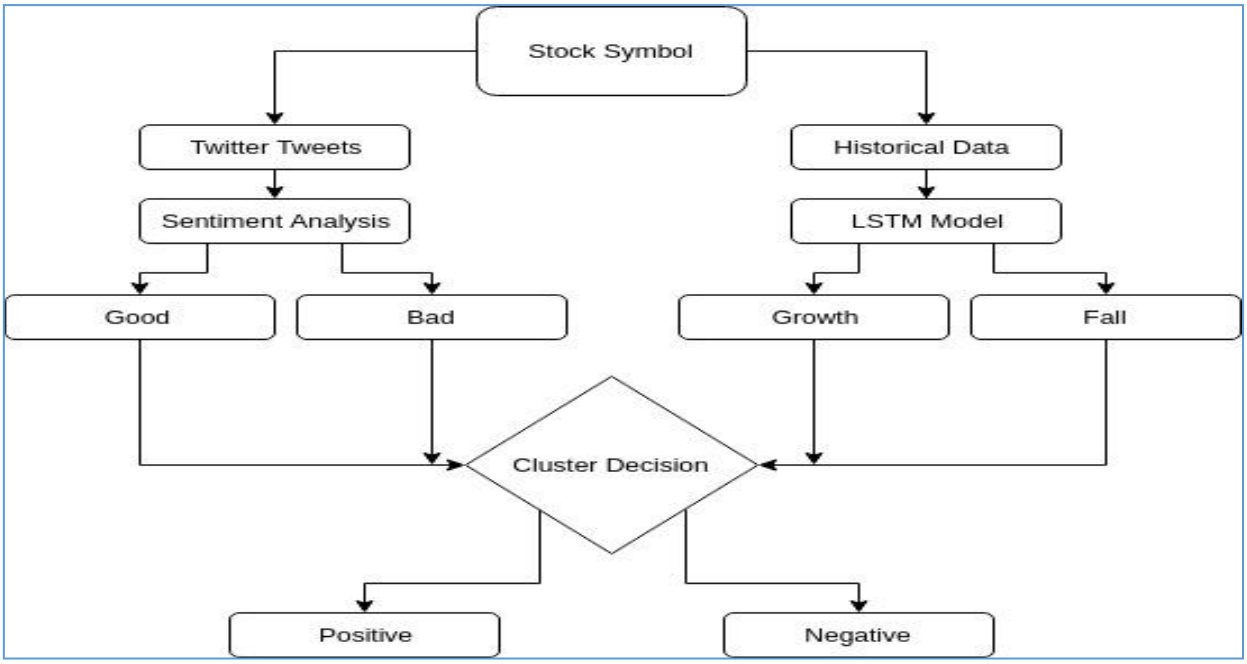

Figure 1 Sentiment Analysis & LSTM Model for historical data

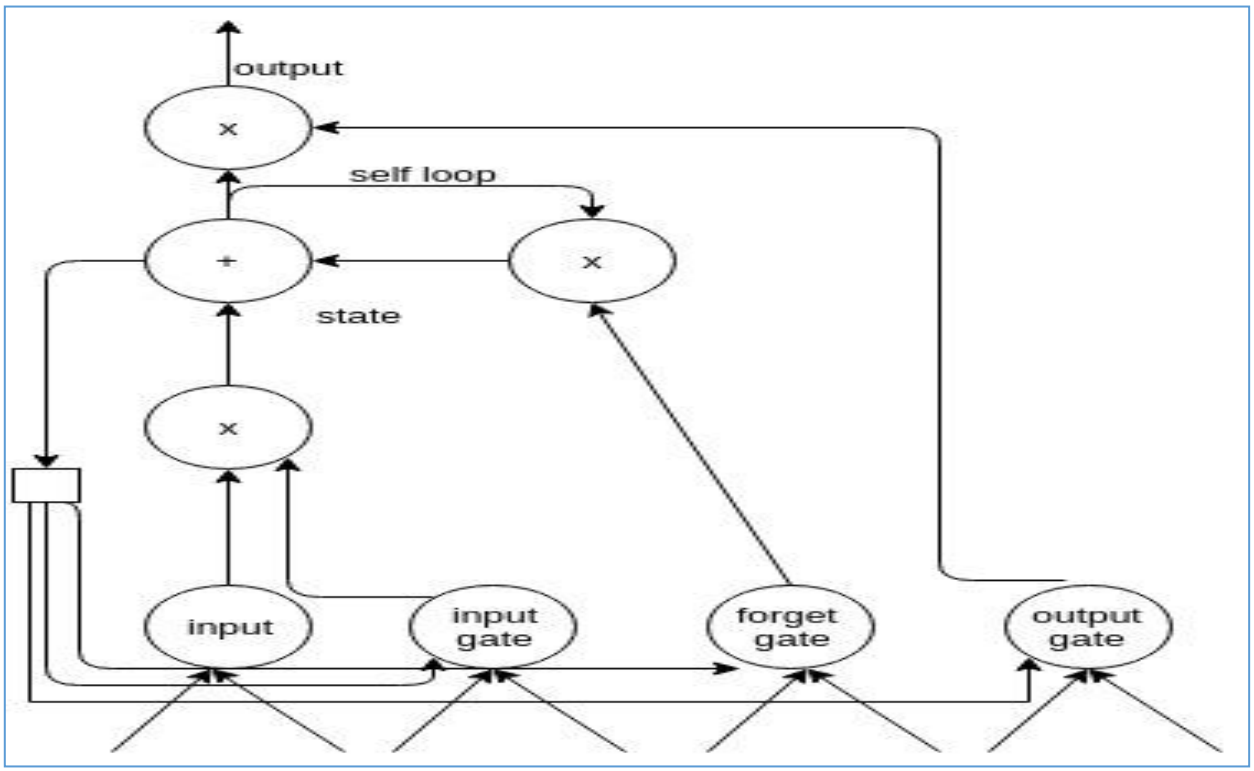

Figure 2 Hybrid Deep Learning Stock Prediction

= 1 if $w_0\, x_0 + w_1\, x_1 + \cdots w_{n-1}\, x_{n-1} < 0$

$O(\vec{x})$=Sgn$(\vec{w}, \vec{x})$ x= input vector w= weights

Sigmoid = sgn(x) = $\frac{1}{1+e^{-x}}$

The tweets obtained from the twitter API are made to pass through preliminary filters to remove the noise as

1. Remove the emoticons
2. Removal of hashtags
3. Other language texts

Once the data is parsed and cleaned. It's is converted to Bag-of-Words and vectors are obtained which are fed to a trained supervised Neural Network, classifying the tweet in a positive or negative class. Considering a specific threshold given to a neural network, which uses the Bag-of-Words, to store specific number of most occurred words from pre-processed data without any regards of grammar. Each word in BOG is converted to a vector.

Size of Vector = size of BOG* number of word occurrence

The resultant vector of a complete tweet is taken by the commutative sum of all the vectors present in the specific instance.This vector is used as training feature in classified neural network of target value as Positive or Negative (-1,1). Once the model has been trained on the basis of the above data, it can be used to classify the sentiments tweeted about a stock in a chosen time period. The design of the proposed RNN for sentiment analysis has following layers.The flow of sentiment analysis-based stock growth prediction is given below

Input layer	To obtain training sample and to transfer feature vector to network. This layer calculates the distance between training vector and input vector upon receiving the input.
Pattern layer	This layer computes the Euclidean distance for each test case from center point of neuron, the computed distance is then applied to Gaussian function using sigma value
Summation layer	This layer adds up the inputs obtained from previous layer, which matches to a category of the selected training pattern
Output layer	This layer selects the highest value from the input it got from previous layer and defines the class of test case.

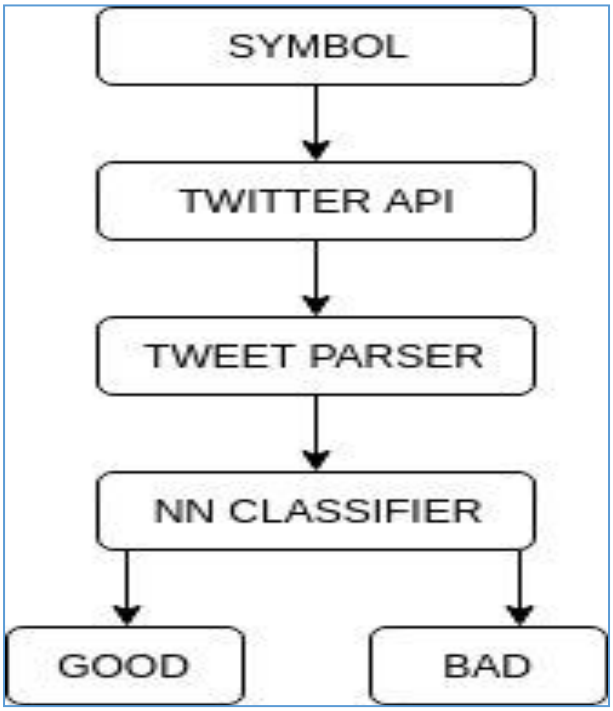

Figure 3 Sentiment analysis

B. LSTM model for historical data

Recurrent Neural Network (RNN) is a class of Artificial Neural Networks (ANN), which uses their recursive internal state to obtain better decisions. The RNN are one of the best classes of NN used to understand the pattern on a time series data, because the ability to map entire historical pattern of x to y. Gated RNNs are the most effective classification models used in practical applications. RNN include the LSTM and networks based on Gated-RNNs.

The LSTMs uses the self-loops to yield paths where the gradient can flow for long durations making it the core of initial models introduced, and taking decisions with self-gated loops (handled by another hidden layer), changes the time scale dynamically, helping to obtain very powerful results. The LSTM had been found exceptionally effective in many applications, like speech recognition and time series data.The LSTM block model is been illustrated in Figure 2.

Target gate $f_i^{(t)}$

$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right)$

$x^{(t)}$: is current input vector

$h^{(t)}$: is current hidden layer, containing of all LSTM

b^f = bias

U^f = input weight

W^f = Weights of recurrent target gate

$$S_i^{(t)} = f_i^{(t)} S_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}\right)$$

U & W respectively denote the biases, input weights, and recurrent weights into the LSTM cell. The external input gate unit $g_i^{(t)}$ is computed likewise to the forget gate (with a sigmoid unit to obtain a gating value between (0 and 1), but with its own parameters:

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}\right)$$

The output $h_i^{(t)}$ of the LSTM cell can also be closed using the output gate $q_i^{(t)}$ which also uses a sigmoid unit for gating:

$$h_i^{(t)} = \tanh\left(S_i^{(t)}\right) q_i^{(t)}$$

$$q_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}\right)$$

As we can see the cells are connected recurrently to each other, replacing the hidden units of ordinary recurrent networks. As the input feature is computed with regular artificial neuron unit.

The architecture is having 3 memory cells, input gate, forget gate and output gate.

The input gate decides provisionally which values from the input gate to be updated in the memory state. The forget gate decides provisionally what information to throw away from the block. The output Gate decides provisionally what is the output based on input and memory of the block.

At a given time t, the memory cells in the network contains information from previous state (t-1). When the LSTM receives an input x, it's given the vector from previous state (t-1). The grouping is based on the magnitude of the input gate and the forget gate. Hence the memory cells is updated to the latest value. Finally, the output value is computed by the LSTM cell by passing current cell value through a non-linearity. How much of these computed output must be actually passed as the final output is determined by output gate.

From the parsed historical data, we compute the OHLC average, which is used to train the LSTM network. The data size distribution for train-test size is taken as 75-25% respectively.

The LSTM model consists of 2 LSTM layer of 32,16 nodes, followed by a dense layer. The final activation function used is linear. Total number of epochs taken are 200 for each stock.

C. Clustering

The results of both the above models are clustered and then based on the result of maximum clustered decision, the final result of high or fall of the stock is made and given as the result.

III. RESULTS

Dataset has been acquired from NSE (National Stock Exchange), which is the leading stock.

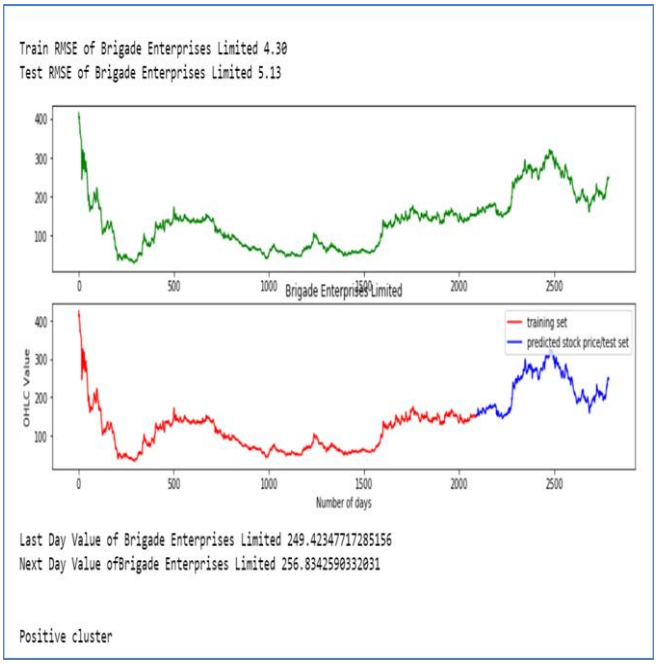

Figure 4 Results for Brigade Enterprises Limited

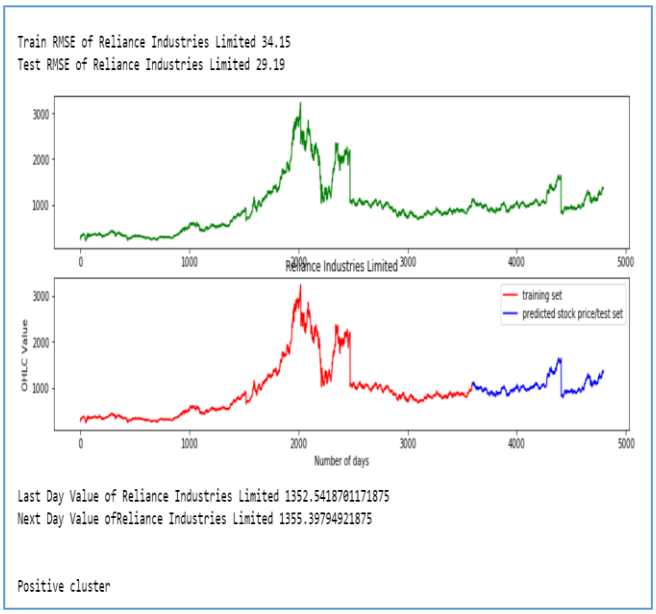

Figure 5 Results for Reliance Industries Limited

exchange of India. The Data is on day-to-day interval over 19 years of data, starting from Jan 2000. For the companies not listed from 2000, their data is taken from the day they got listed. We have taken the selected number of stocks for experimental study, on a random basis.

Stock market data is having multiple attributes, as trades, volume, turnover and others. Which are not useful for the prediction in this model, therefore we drop them.

The major attributes used for predictions and visualization by traders are OPEN, HIGH, LOW, CLOSE or OHLC graph. Therefore, the same is used for the model. Open-high-low-close chart (or OHLC Charts) are used as a trading tool to visualize and analyze the price changes over time for securities, currencies, stocks, bonds, commodities, etc.

OHLC Charts are useful for interpreting the day-to-day sentiment of the market and forecasting any future price changes through the patterns produced. New feature is used as VOLATILITY which is built upon the attributes of CLOSE- OPEN. The attribute shows the volatility of the stock, indicating the chances of growth and fall. The visual representation of the volatility and OHLC graph shows the important change and growth in a stock. The results for stocks in National stock exchange of India using the proposed solution is given in figure 4. For the LSTM Historical Data Module, we obtained an RMSE value from 5-50 units. And 10-4 loss values in epoch results.

As we can see the results of stock predicted values one can observed values for a given period of time. By visualization also we can observe the prediction of stock is in same pattern to observed values, even in exceptional instances also. The clustering of the stock gives us the idea of increasing and losing stock over the period of time, by putting them in cluster not by giving a specific value to growth or fall.

IV. CONCLUSION

The proposed hybrid deep learning-based stock prediction gave a new modular approach to find the pattern of prediction in stock markets. As shown in survey, there are several methods on predicting signals through machine learning algorithms and numerical methods. Sentiment analysis and LSTM were implemented side by side, through which we were able to cluster the stocks in growth or loss. This work can be extended by adding the predictive text analysis on the basis of information present about the stock.

REFERENCES

1. DINESH BHURIYA, UPENDRA SINGH, "Survey of Stock Market Prediction Using Machine Learning Approach", ICECA 2017.
2. Mansoor Momeni, Maryam Mohsen, "CLUSTERING STOCK MARKET COMPANIES VIA K- MEANS ALGORITHM", Kuwait Chapter of Arabian Journal,2015
3. Sahaj Singh Maini, Govinda.K, "Stock Market Prediction using Data Mining Techniques ", IEEE, 2018.
4. Rajat Ahuja, Harshil Rastogi, Arpita Choudhuri, "Stock Market Forecast Using Sentiment Analysis", Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference.
5. Robert P. Schumaker and Hsinchun Chen, "Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System",ACM Transactions on Information Systems (TOIS), 2009
6. Fischer, T., & Krauss, C. Deep learning with long short-term memory networks for financial market predictions.2018 European Journal of Operational Research, 270,654–669
7. R. Akita, K. Uehara, et al., Deep learning for stock prediction using numerical and textual information. IEEE ICIS, Vol. 15, 2016
8. J. F. Chen, W. L. Chen, C. P. Huang, S. H. Huang and A. P. Chen, “Financial Time-Series Data Analysis Using Deep Convolutional Neural Networks,” 2016 7th International Conference on Cloud Computing and Big Data (CCBD), Macau, 2016, pp. 87-92
9. A. Tsantekidis, N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj and A. Iosifidis, “Forecasting Stock Prices from the Limit Order Book Using Convolutional Neural Networks,” 2017 IEEE 19th Conference on Business Informatics (CBI), Thessaloniki, 2017, pp. 7-12.
10. Rather, Akhter Mohiuddin, A. Agarwal, and V. N. Sastry. “Recurrent neural network and a hybrid model for prediction of stock returns.” Expert Systems with Applications 42.6(2015):3234-3241.
11. Nelson, David M. Q., A. C. M. Pereira, and R. A. D. Oliveira. “Stock market's price movement prediction with LSTM neural networks.” International Joint Conference on Neural Networks IEEE, 2017:1419-1426.
12. Sheng Chen "Stock Prediction Using Convolutional Neural Network",IOP Conf. Series: Materials Science and Engineering 435 (2018)
13. Bo Xu, Dongyu Zhang, Shaowu Zhang, Hengchao Li,"Stock Market Trend Prediction Using Recurrent Convolutional Neural Networks",© Springer 2018
14. Ding, X., Zhang, Y., Liu, T., et al.: Deep learning for event-driven stock prediction. In: Ijcai, pp. 2327–2333 2015
15. Leonardo dos Santos Pinheiro, Mark Dras "Stock Market Prediction with Deep Learning: A Character-based Neural Language Model for Event-based Trading" In Proceedings of Australasian Language Technology Association Workshop 2017

AUTHORS PROFILE

Guruprasad S, B.E, M.Tech in CSE, currently working as Assistant Professor in Dept.of CSE, BMSIT&M Bangalore**.** Perusing research in stock market prediction.

Sahil Verma is a Software Development Engineer at Tally Solutions, Bangalore. He is also a researcher and developer in Machine learning. He has completed his Bachelor of Engineering in computer science at BMS Institute of Technology and Management. Presently working on Time series data and financial market analysis

Dr Chandramouli H received his Ph.D in the year of 2014 and currently working as a Professor in the Department of Computer Science and Engineering at East Point College of Engineering and Technology, Bengaluru. He has 22 years of rich experience in the Academics. He has **published more than 25 research articles in National and** International Journals. He holds CSI membership and an active member in CSI events. His research area includes Wireless sensor network, Resource allocation in Networking, Big Data Analytics.