International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

AI and ML-based - News Classifier for Share Market

Smita Deshmukh[1], Suyash Saxena[2], Pooja Devrukhkar[3], Sagar Sanil[4]

[1]Sr. Asst. Professor, Department of Information Technology, Terna Engineering College, Nerul, New Mumbai, India
[2,3,4] Department of Information Technology, Terna Engineering College, Nerul, New Mumbai, India

---

Abstract - *Share market is an investor's paradise, and hence it is necessary for an investor to have an idea about the market trends, share prices, its ups and downs due to the news around. The principle target of this project is to characterize the news as positive, negative, or neutral, and study its impact on the movement of the stock index. Sentimental analysis will be used to achieve this task. Based on this segregation we will predict the ups and downs that may happen to the BSE SENSEX index. The financial headlines will be collected in real-time from sites like Money control, FT, etc. and then will be processed and our system will predict, by studying the sentiment of news[1], whether headlines have an effect on the movement of the index, if it will increase, decrease or stay the same. The model which predicts the movement of the index more accurately is studied[2].*

***Index Terms*—stock market, prediction, machine learning, LSTM, news, Text mining.**

1. INTRODUCTION

The process of determining the future value of a company's stock is termed as Stock Market prediction. It also helps to determine other financial instruments being traded. Significant profit can be achieved, and losses can be minimized by successful stock predictions. Since the prices of stock are related or dependent on each other, the process of conventional batch processing cannot be used with much efficiency for analysis of the stock market. Share market is an investor's paradise, and hence it is necessary for investors to have an idea about the market trends, share prices and ups and downs related to the market. Profits can be maximized in the stock market with minimal risk by predicting the market value. This process of share market trend analysis involves a lot of data, but nobody can predict this trend accurately even with 100 different sources. We will be taking into account the historical data of BSE Sensex, and scrape the financial news headlines from Money Control website and do Sentimental Analysis on them and classify them into 3 groups of positive, neutral and negative news. Based on this classification we will find the correlation between the ups and downs that may happen to the Sensex index with respect to the sentiment of the news. The analysis of stock depends on three broad factors which mostly overlap with each other. They are fundamental analysis which studies intrinsic factors of stock, technical analysis where Sensex prices and volume is considered and technological methods. The real-time news impact must be known to the investor for the actions to be taken to maximize the profit and minimize the loss that might happen. News must be perused by the machine to comprehend its sentiment and subsequently sentimental analysis, by utilizing the NLP toolkit in python [3] is done. The model which predicts the movement of the stock movement more accurately using historical data is built, and thus may be applied in real-time[4].

2. LITERATURE SURVEY

2.1 Stock Price Prediction Using Long Short Term Memory

Raghav Nandakumar, Uttamraj K R, Vishal R, Y V Lokeswari [5] have suggested the use of stochastic learning paradigm since each data point in the model was trained. Since within a day the prices vary very drastically, the two factors which are very important in predicting the trends are Speed and Accuracy. To forecast the closing price of the stock at the end of each day using LSTM, the authors proposed the methodology of learning which will be online. A comparison of LSTM is done with ANN and LSTM proves to be better than ANN. They specify the benefits of LSTM over RNN. Root Mean Squared Error is used to evaluate the exactness of the model. Since LSTMs can keep a track of the temporal dependencies that are present between the prices of stocks for a greater period of time, hence they are considered better than others. Correlation of sentimental analysis output and LSTM is not obtained in this research.

2.2 Predicting Stock Prices Using LSTM

Murtaza Roondiwala, Harshal Patel, Shraddha Varma [6] have emphasized on the need of efficient systems for prediction of the stock market. It suggests that the system consisted of several stages which also helped them improve the accuracy of the model. The various stages were 1) Pre-processing of the data. 2) Extracting the features. 3) Training the neural network. The authors here used the same procedure for analyzing the efficiency of the model that was built using the Root Mean Square Error. The value of RMSE is used to minimize the difference between error and the output value that is obtained. RMSE proved to be better than Mean absolute error. The paper suggests that LSTM had accuracy more than any other variant. The effect of volume

feature on stock is not studied and is thus been studied by us.

2.3 Stock Market Prediction based on Deep Long Short-Term Memory Neural Network

Xiongwen Pang, Yanqiang Zhou, Pan Wang, Weiwei Lin and Victor Chang [7] demonstrated the notion of a 'stock vector', very much like the notion of 'word vector' found in deep learning. The input was a high-dimensional historical data of multiple stocks and not of a solitary stock or index. The "word vector", utilizes a sequence of numbers to represent the words. The authors stated that when historical data of multiple stocks are taken into account, the input dimension increases exponentially, had they used this information for predicting the stock market, it would have been a waste due to useless and repetitive information. Hence, the 'stock vector' concept was introduced by the authors. The stock vector's dimension was reduced and was hence represented in a low dimension space. Finally, the forecasting was done using stock vector. An Embedded Layer LSTM neural network (LSTM) was then used for prediction. The predictions had a drawback of less accurate sentiment analysis.

2.4 A Comparative Study of LSTM and DNN for Stock Market Forecasting

Dev Shah, Wesley Campbell and Farhana H. Zulkernine [8] did a comparison between two artificial neural network models. One of them was the recurrent neural network (RNN), the Long Short-Term Memory (LSTM), and another one was the deep neural network (DNN) for predicting the movements in the Sensex index on a daily and weekly basis. They created a model and tested it with one more stock to check its generalization. After model compilation comparison between the models was made. According to the authors both the created models were able to make acceptable predictions, both daily and weekly. A generalizable test was also done to check for overfitting of the model, and it was found that the LSTM was better able to adapt and adjust to the changes found in the original data. Here, price data was the only parameter used for the prediction of the stock which can be optimized by using volume, high, low price data.

2.5 Stock Price Prediction Using LSTM on Indian Share Market

Achyut Ghosh, Soumik Bose, Giridhar Maji, Narayan C. Debnath, Soumya Sen [9] had the objective to predict the future price and calculate the future growth of the company in the different time span. The training data set is fed into a NN model and the model is initiated using random weights and biases. The proposed LSTM model consists of a total of 5 layers, 1 sequential input layer, 3 LSTM layers and a final dense layer with activation. The output layer is a dense layer having a linear activation function. The target values are contrasted with the RNN generated output and the error difference is calculated. The Back-propagation algorithm is utilized to lessen the difference in error by fine-tuning the parameters of the neural network. The prediction would have been more accurate if the model were trained on a larger data set. So based on the results it was considered to apply the LSTM based model to predict the share price on long time historical data. The data set studied here was small for accurate learning and prediction by the model.

2.6 Stock Movement Prediction Using Machine Learning on News Articles

In this research paper the authors B R Ritesh, Chethan R and Harsh S Jani [10] have tried to find a method by which they can predict a stock's future change in price on a daily basis. To do this, the authors have used news articles. The author concentrates on two parts: 1) Prediction of a stock's price; and 2) A system through which a high return on investment can be achieved using the predicted stock price. The linear SVM [11][12] classifier by far yielded the best results among the other classifiers. From the results, news articles and the and the development of the stock regularly, showed a correlation between each other. They also suggested that machine learning[13] can be applied to understand the decision-making ability of the trading simulation process. Sentimental analysis was not implemented, instead raw textual data was used for learning of classifiers thus decreasing the accuracy of the data.

3. PROPOSED SYSTEM

- Steps followed for building a model of stock market prediction using news headlines-

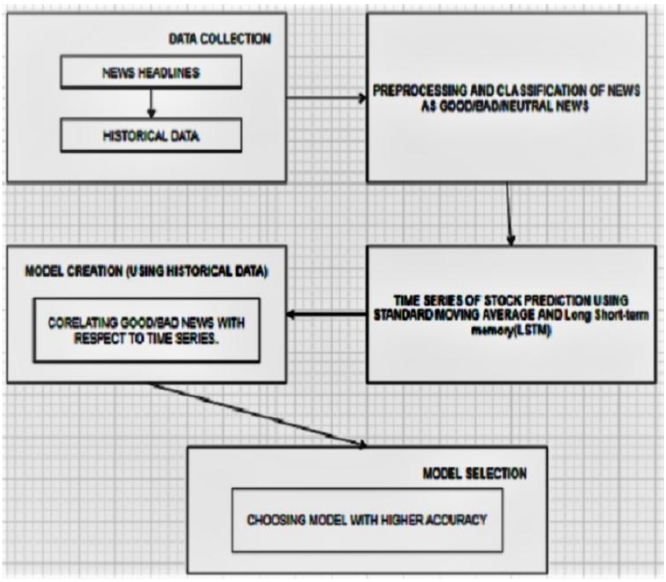

Fig-1: System Framework

1. **Data Collection:** News headlines are collected from a website which provides all business and other headlines at a single point. The news headlines are collected and stored in our file in csv format along with the date of the source headline. BeautifulSoup package is imported which scrapes the headlines from the source website.
2. **Preprocessing data:** The method of converting raw data into a machine-understandable and easily analyzable form. It involves various steps such as Data Cleaning, Data Integration, Data Transformation, Data reduction, Data Discretization. Irregularities, duplication in the data are removed and data is streamlined for further analysis.
3. **Polarity score:** The polarity of the collected news headlines is obtained via sentiment analysis and then used to build models that give maximum accuracy of the stock price movement. The data here is already collected for training and hence only past data is used. Live data collection can be done for real-time prediction.
4. **Building Model:** The model with maximum accuracy is selected after training on the historical news which was collected. The study of literature survey suggests the use of LSTM, RNN, Naive Bayes.

- **Working:**

1. The news headlines are collected from the website Moneycontrol.com from the year 2016 to 2020. The website had 25 headlines per page, and around 15000 pages worth of headlines was scraped and collected. The Sensex data was collected from the BSE website. BSE gives stock market index of 30 companies which are representative of the industrial sector of the Indian economy. The data stores 'Date', 'Open', 'High', 'Low', 'Close', 'Adjusted Close' and 'Volume' of the stocks traded in the market. The data from 1997 to 2020 is collected. The collected data was then preprocessed. Natural language processing was then applied to the data, to make it computer understandable. One of the many applications of NLP is sentiment analysis. Opinion mining or sentiment analysis is a step-based technique of using NLP algorithms to analyze textual data. It was done on the news headline dataset to obtain meaning in machine-understandable form and thus the polarity of the headlines was obtained.
2. In our project we have compared VADER and TextBlob methods for sentimental analysis. TextBlob was used to perform NLP tasks like noun phrase extraction, opinion mining, part-of-speech tagging and more. TextBlob is designed to handle both structured and unstructured forms of data. The polarity of the sentence states the positivity and negativity of the sentence, ranging from – 1 to 1. On the scale 1 indicates positive statement and -1 indicates a negative statement. Where anything greater than 0 is positive and below 0 is negative. The subjectivity ranges from 0 to 1 which implies the fact to opinion about that statement. 0 indicates fact and 1 indicates the opinion percentile of the sentence.
3. VADER which stands for Valence Aware Dictionary and sEntiment Reasoner is a sentiment analysis tool, which is a lexicon and rule based. As stated and proved by C.J. Hutto, Eric Gilbert in their paper [14] 'VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text' by comparing VADER sentiment analysis tool to several other sentiment analysis lexicons, details of which can be found in their paper, they found that VADER performs highly well compared to others. Hence by comparing the two, Textblob and Vader, we concluded that VADER provides more accurate results.
4. Long short-term memory (LSTM) [15] is an advanced recurrent neural network (RNN) model. LSTM is composed of a cell which further consists of an input gate, a forget gate, and an output gate. The value is remembered by the cell for an arbitrary time interval. The regulation of the flow of information in and outside of the cell is managed by the three gates. The drawback of RNN is the problem of Vanishing Gradient i.e. to forget the long-time data. This drawback is removed by LSTM. LSTM's are specially designed to remove the long-term dependency problem of RNN. 1. Forget gate: This gate decides and removes the data which is no longer required for the LSTM and thus the cell state is been replaced with more recent information. 2. Input gate: The input cell learns to add or store the information in the cell. 3. Output gate: The output data depends on the input data and the cell state. In this project, we have created a two-layered LSTM network, with the output of the first layer being fed into the second layer. Both the LSTM layers are of 50 neurons each with two additional Dense layers, with the final dense layer working as the output layer. The historical Sensex data from the year 2016 to 2020 was used for prediction in the LSTM model. The Closing price from this data was used for the prediction purpose. The comparisons of LSTM and Standard Moving Average (SMA) are based on RMSE. Thus, Root mean squared error is used for evaluation of data.

4. RESULTS

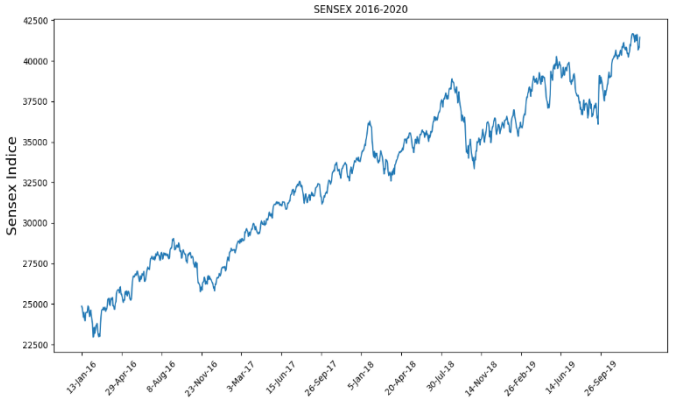

Fig-2: Sensex plot for the year 2016 to 2020

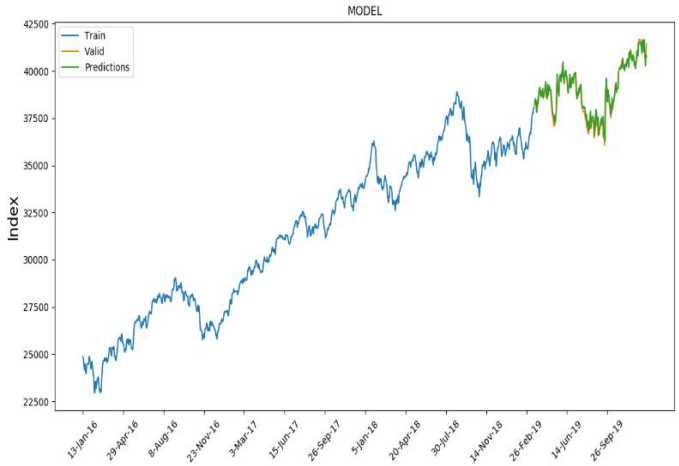

Fig-3: LSTM prediction after training on the Sensex data with a loss value of 3.7348e-04

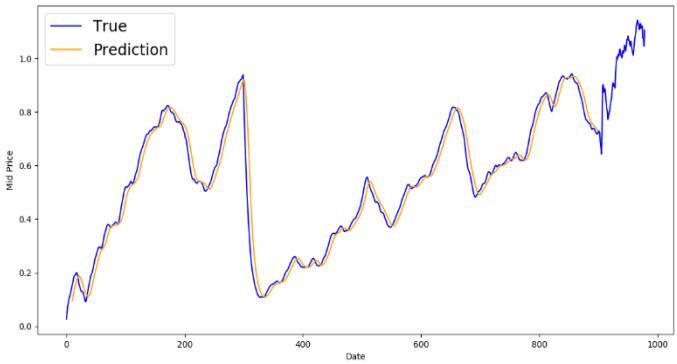

Fig-4: Standard Moving Average Plot for Sensex data

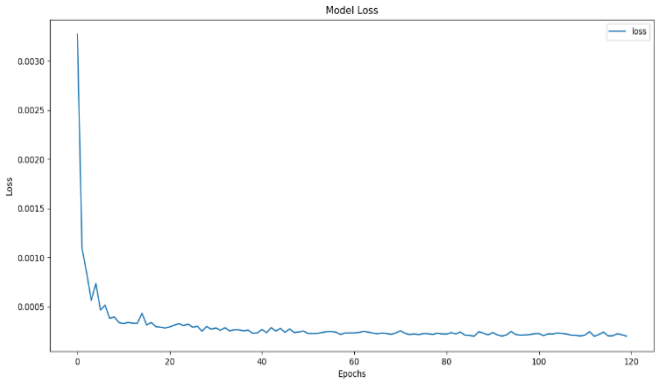

Fig-5: LSTM Model Loss

```
See the caveats in the documentation:
  valid['Predictions']=prediction
PREDICTED PRICE:  [[41627.168]]

Process finished with exit code 0
```

Fig-6: Predicted index value for Sensex on 10th January 2020 (Actual index value was 41,599.72)

```
{'neg': 0.255,
 'neu': 0.532,
 'pos': 0.213,
 'compound': -0.1613,
 'headline': "SAD not quitting NDA to save Harsimrat's chair: Amarinder Singh"},
{'neg': 0.0,
 'neu': 0.753,
 'pos': 0.247,
 'compound': 0.3182,
 'headline': 'OPPO to increase its marketshare in Tamil Nadu'},
```

Fig-7: Result of Sentiment Analysis done using VADER

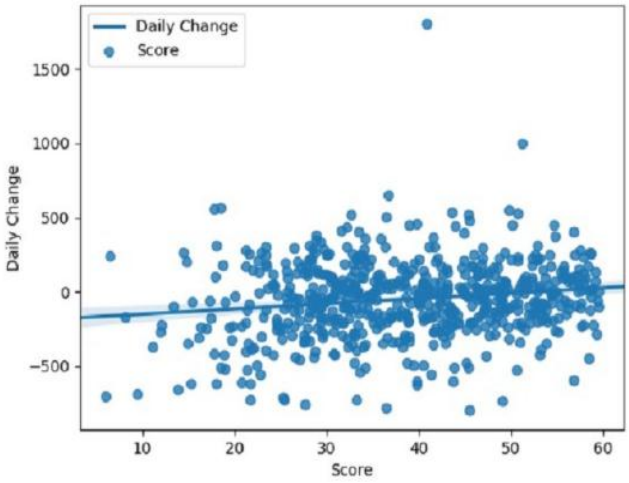

Fig-8: Regression plot showing a loose correlation of 16% between news sentiment and the daily change in index value

5. CONCLUSION

In this project, we found that the news sentiments alone are not the only driving factor behind the change in Sensex indices. We found that while Standard Moving Average technique for prediction can give good results for the index of the next day, it falls short when the predicted value is far into the future. LSTM on the other hand gave us a better predictive value for future values with a much lower error rate. Hence, LSTM is the better choice for stock prediction.

6. IMPROVEMENTS

In this study, we have only considered the closing price for the prediction, thus a combination of various other features can also be used for prediction and increasing the accuracy. The use of more sophisticated sentiment analysis methods can also be done by the researchers. A user-created dictionary of positive and negative words can also be created and used, thus building a more robust model. Supervised learning of sentiment analysis can also be done by the researchers [16]. An LSTM model with far more features than the ones being used in this research may be able to predict indices with a much greater accuracy.

ACKNOWLEDGEMENT

Our sincere appreciation goes to the Engineering Product Innovation Center (E.P.I.C) and staff members of the Department of Information Technology, Terna Engineering College, Nerul.

REFERENCES

[1] Marti A. Hearst, Untangling text data mining, Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics.

[2] Kari Lee and Ryan Timmons Predicting the Stock Market with News Articles. CS224N Final Project.

[3] Yashwant Singh Patel, Supriyo Mandal- Stock Market Prediction using Daily News Articles

[4] Kalyani Joshi, Bharathi H.N, Jyothi Rao- Stock trend prediction using news trend analysis. International Journal of Computer Science and Information Technology. Vol. 8. June 2016.

[5] Uttamraj K R2, Vishal R3, Y V Lokeswari4- Stock Price Prediction Using Long Short Term Memory Raghav Nandakumar1, International Research Journal of Engineering and Technology (IRJET).Volume: 05 Issue: 03. Mar-2018.

[6] Murtaza Roondiwala, Harshal Patel, Shraddha Varma- Predicting Stock Prices Using LSTM. International Journal of Science and Research (IJSR). Vol. 6. April 2017.

[7] Xiongwen Pang1 , Yanqiang Zhou1 , Pan Wang2 , Weiwei Lin3 and Victor Chang4- Stock Market Prediction based on Deep Long Short Term Memory Neural Network.Proceedings of the 3rd International Conference on Complexity, Future Information Systems and Risk -Volume 1: COMPLEXIS, 102-108, 2018.

[8] Dev Shah, Wesley Campbell,Farhana H. Zulkernine- A Comparative Study of LSTM and DNN for Stock Market Forecasting. Volume: 1, Pages: 4148-4155.Year: 2018.

[9] Achyut Ghosh , Soumik Bose , Giridhar Maji2 , Narayan C -Stock Price Prediction Using LSTM on Indian Share Market. Proceedings of 32nd International Conference on Computer Applications in Industry and Engineering.vol 63, pages 101–110.September 2019.

[10] B R Ritesh, Chethan R and Harsh S Jani- Stock Movement Prediction Using Machine Learning on News Articles, International Journal on Computer Science and Engineering (IJCSE), Vol. 9 No.8 ,Aug 2017.

[11] Rong-Kuan Shen, Cheng-Ying Yang, Victor R. L. Shen, Wei-Chen Li, Tzer-Shyong Chen - A stock market prediction system based on high level fuzzy petri-nets.International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 26(05), 771-808, October 2018

[12] Ching-Te Wang and Yung-Yu Lin, "The prediction system for data analysis of stock market by using Genetic Algorithm," 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Zhangjiajie, 2015, pp. 1721-1725.

[13] Atkins, Adam Niranjan, Mahesan Gerding, Enrico. (2018). Financial News Predicts Stock Market Volatility Better Than Close Price. The Journal of Finance and Data Science. 4. 10.1016/j.jfds.2018.02.002.

[14] Hutto, C.J. Gilbert, Eric. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

[15] Hasim Sak, Andrew W. Senior, Franc¸oise Beaufays - Long short-term memory recurrent neural network architectures for large scale acoustic modeling, INTERSPEECH (2014), pp. 338-342.

[16] T. Kimoto, K. Asakawa, M. Yoda and M. Takeoka, "Stock market prediction system with modular neural networks," 1990 IJCNN International Joint Conference on Neural Networks, San Diego, CA, USA, 1990, pp 1-6 vol.1.