Use of Heterogeneous Data for forecasting of stock market

1st Dhanush Vasa
IIT2019208
5th Semester
IT

2nd Lekhana Reddy Mitta
IIT2019204
5th Semester
IT

3rd Meghana Santhoshi Kandagatla
IIB2019030
5th Semester
IT-BI

4th Pedada Gopal
IIT2019065
5th Semester
IT

***Abstract*—The objectives of this project work is to develop a RNN (Recurrent Neural Network) based on a LSTM (Long Short Term Memory Model) to forecast the movement of stock volatility. We have analysed various fields related to a stock market by giving our model various features as input and comparing it with the original trend of the stock market. These features includes heterogeneous data from carefully selected social media networks to analyze the public mood along with the news headlines extracted from official news papers to analyse the public mood. These public mood results are used as an influential feature in order to forecast a better outcome then traditional transaction stock price prediction.**

***Index Terms*—Heterogeneous data, Long Short Term Memory Model, Stock Prediction, Recurrent Neural Network, Social Media, News Headlines, Client Category-wise Turnover, Machine Learning**

I. Introduction

A country's GDP and many companies opulence is dictated by the stock market in a vital manner. This stipulates that if the stock market rises, the country's GDP would be high indicating an increase in the countries' economic growth. We can corroborate that the growth of the country and stock market are closely linked to behaviour of the stock market. As history as shown that the stock market is has no certainty on each and every investment made, which may lead to gaining immense fortune, impartial gain or loss, or suffer in major losses. These result have incurred due to many factors affecting the stock prices out of which historical transaction data have been a major contributor. From a lot of research we can corroborate that historical transaction data alone is insufficient to give us an accurate prediction. Hence, factors like daily financial news and social media have also been major factors in affecting the market values in them positively or negatively. We can say that all of these factors must be taken into account for a precise forecasting of the stock market. As the stakes are high for investing in the stock market a highly accurate and automated analytic system is necessary for which it can handle enormous amounts social media and financial news.

All of these challenges can be over come by the help of machine learning models and more specifically RNN (Recurrent Neural Network) based on LSTM (Long Short Term Memory Model) to forecast the movement. With the involvement of LSTM a time series prediction model which is able to build a structure of a long term memory with only short term memory. As we are aware of the model in machine learning which is LSTM the key point is the data-set that the model needs to learn. In our project we will be using LSTM as stated is a supervised machine learning model, where LSTM will take major portion of the data to create a structured model and the remaining to test. With this we will be using the social media and financial news headlines analysing there public moods and using them as influential features with the historical transaction data to forecast a better result then with plain historical data. Hence, this will help companies and investors in making decisions easier and concurrently invite more investors to the stock market which will help the countries' GDP.

II. Related Work

In this Digital era, entities that we can consider that are affecting the stock values are social, psychological, political and Economic factors. So, we have considered going through few researches.

To Refer various Machine Learning Models that can be used , Contribution Polamuri Subba Rao, K. Srinivas, A. Krishna Mohan[1] used various Machine Learning techinques like ARIMA, Holt-Winters, Artificial Neural Network, Hidden Markov Model, RNN.The stock market forecasting system is to increase accuracy.

To Refer the variants of Deep Learning Technique RNN , Contribution by Saloni Mohan1, Sahitya Mullapudi [3] used ARIMA , Facebook Prophet, RNN-p, RNN-pp, RNN-pt, RNN-mv models.

To Refer to the Social Media Related Data trained on sentimental analysis Models, Contribution by Rubi Gupta,

Min Chen [6] used Sentimental analysis with logistic regression and TF-IDF, and then analysed using percentage of positive sentiments method from the textual data collected.

III. Literature Survey

A. Stock Prediction using Machine Learning Techniques

- Citation: Polamuri Subba , Srinivas, Kudipudi Mohan, A.. (2020). A Survey on Stock Market Prediction Using Machine Learning Techniques. 10.1007/978-981-15-1420-3_101.
- Data set: Collected from Indian Stock Market Websites
- Models Used: ARIMA ,Holt-Winters, Artificial Neural Network, Hidden Markov Model,RNN
- Results: This paper comes up with an appraisal and correlative analysis of different stock market prediction parameter techniques. These techniques are widely used to access the worth of stock market production and trends. The forecasting system of stock market is to increase accuracy. In this work to analyse a novel approach to improve the prediction of the results of stock, that implies we will have to integrate two or more methods to construct novel approach method.
- Limitations: Prediction Accuracy decreased with increase in noise variation. Few models are only suitable for short term prediction.

B. Stock Price Prediction Using News Sentiment Analysis

- Citation: S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 2019, pp. 205-208, doi: 10.1109/BigDataService.2019.00035.
- Data set : Data extracted from Financial news articles, for five years .To reduce the variation high and low stock prices , applied log transformation is done.
- Models Used : ARIMA , Facebook Prophet, RNN-p, RNN-pp, RNN-pt, RNN-mv.
- Results : RNN- LSTM is the best option among three and among variants in LSTM , RNN - pp is the better one .
- Limitations: The model is not up to the mark , when the data is highly or less volatile.

C. Sentiment Analysis for Stock Price Prediction

- Citation: Rubi Gupta & Chen, Min. (2020). Sentiment Analysis for Stock Price Prediction. 213-218. 10.1109/MIPR49039.2020.00051.
- Dataset: For analysis, both stock and StockTwits data are collected with 120 days of data used for training and remaining for testing.
- Models Used: Naïve Bayes, SVM, and logistic regression and featurisation techniques used are bag of words, bigram, trigram, TF-IDF, LSA
- Results: The hybrid model of logistic regression and TF-IDF achieving high accuracy for the companies considered.
- Limitations: Accuracy can be more precise , if training data and testing data collection is more.It could we be easier to understand the movement of the stocks.

IV. Methodology

A. Data Collection

When it comes to Data Collection and Machine Learning this consists of the procedure of analyzing, calculating and precisely collecting information for the research by validated techniques. Next when we are talking about the stock market forecasting, the key step is the data collection where it is mandatory and crucial for future analysis. It is also crucial that data is not extracted from only one source as it may consist of bias and misleading information. For our project here we have used various sources of websites for transaction historical data, various financial news publishers and multiple social media platforms for there posts for data collection. Hence we decided to demonstrate this on the BSE (Bombay Stock Exchange) as it is one of the major stock markets in India.

- **Historical Transaction Data :** For this part what we did differently was we got 2 different datasets.
  - Client Categorywise Turnover
  - Normal daily stock prices of Open, High, Low, and Closed

  The way we were able to get the Client Categorywise Turnover was fairly simple as the official BSE website bseindia.com run by the government was able providing us with reliable and accurate dataset from January 2011 to October 2021. Where this dataset consisted of 4 categories:
  - Client
  - NRI (Non Residential Indian)
  - Proprietary
  - DII (Domestic Institutional Investor)

  Where each category had 3 subsections which were BUY, SALES, NET. The next dataset was the normal daily stock prices of Open, High, Low, and Closed which again we were able to extract easily from the official BSE website besindia.com run by the government again as it provided us with reliable and accurate dataset from January 2011 to October 2021. Where we combined both the datasets together to make one major historical transaction dataset.
- **Social Media :** For social media posts what we did for our project was we extracted posts from two social media networks:
  - Twitter
  - Reddit

  For Twitter what we used was the snscraper library in python with a module dedicated for Twitter. Where we were able to collect every single tweet related to BSE (Bombay Stock Exchange) from January 2011 to October

2021. For Reddit we used an API from Reddit for app development where with the help of the request library in python and the access tokens from Reddit API we collect posts related to BSE (Bombay Stock Exchange) from again January 2011 to October 2021 but due to the limitations of the API we were only able to get for 6 month from April 2021 to October 2021.

- **Financial News Headlines :** For Financial News Headlines we web-scarped. Where we were able to extract every single headline from a given website such as Business Insider, Times of India, Bloomberg and Financial Express to extract most recent and historical news of BSE.

B. Data Preprocessing

Here Data Preprocessing is essentially a major and key step when it comes to Data Mining and Machine Learning. As the effectiveness and usability of the data is predominantly important as this would directly influence the learning quality of any machine learning model. Hence, it is mandatory and essential that we pre-process the data before we input into any machine learning model. When it come to our project it was necessary that we put major focus to pre-process the social media posts and financial news headlines. What we did first was we checked if we had the duplicate social media posts or news headline on the same day. If so we removed them. Next what we did was we concatenated all of the social media posts and financial news headlines into one single section according to there dates and created a new data set for textual information with index as date. Following that what we did was we cleaned with regex commands in python with the replace command removing any thing that is not a alphabet character or space. Next what we did was we checked for any null or empty locations where there was no text present and inserted neutral instead. What this allows use to see the set of words used in the posts and headline to make it easier for sentimental analysis in the future to analysis the positive and negative public mood. Lastly when it came to the transaction data it did not have any null or empty locations which didn't require any preprocessing.

C. Sentiment Analysis

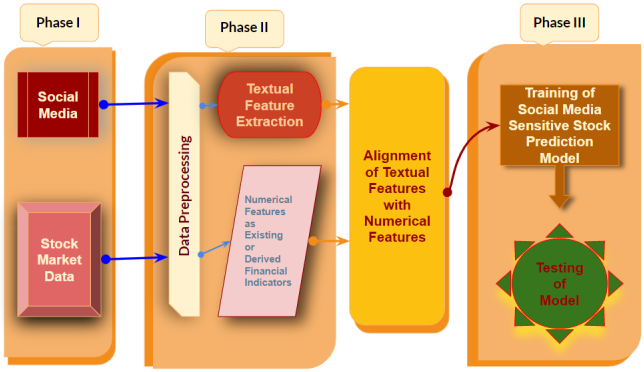
Sentimental analysis is a popular way of being able to know and monitor the public mood on a piece of news or information about a product. We have to consider Sentimental analysis because of the popularity of social media, various influential websites. Investors can also plan their future strategies using this analysis and there opinion on stocks can be affected time and time again. In our sentimental analysis we were able to categorise the data into positive, negative or neutral. In order to achieve the categorisation we used Natural Language Processing tool more specifically vaderSentiment module available on python to each piece of text's emotion/mood.

D. Testing and Training Dataset Preparation

Textual Data have been collected in various files for various dates.Several formats of dates are made into one using delimiters on excel and finally merged them. After the collection of transaction data and textual data. We have to make combined data set. Keeping transaction data on the left and textual data on the right , If we apply left join we get our final Data Set.

70% of the dataset is used for training and 30% of the dataset is used for testing in all Experiments mentioned below.

E. Long Short Term Memory(LSTM)



Long Short Term Memory is a very popular neural network technique, LSTM overcomes the limitations from RNN. RNN faces short term memory problem, to overcome that we use LSTM. In LSTM, there is a feedback mechanism which enables the LSTM to process the entire sequence of data without considering each point of data. This is different from the usual traditional feedback mechanism because of the fact that, it retains only the useful information from the data. LSTM doesn't just store the previous prediction but it remembers longer time context for the prediction. Which clearly solves the long term dependency problem facing by using RNN (Recurrent Neural Networks).

How do LSTM work ?

Prediction of the stock, at a particular state depends on following factors :

- Obviously, the current input.
- Previous prediction stored in previous hidden state.
- Long term memory of the network, stored as cell state.

LSTM has gates to filter the data. LSTM neural network has three gates namely:

- Input gate
- forget gate
- Output gate

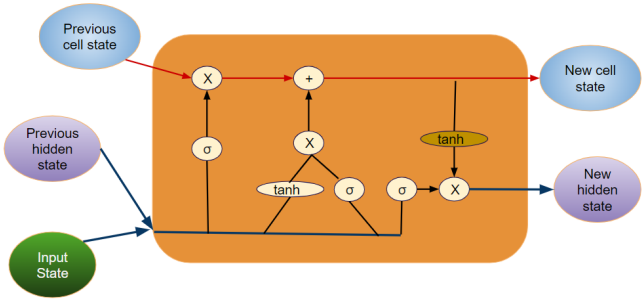
Process of working of LSTM can be explained in the following steps :

1) In the First step, With the help of previous hidden state and current input we try to figure which data is useful and which has to remembered for the further prediction.This is called Forget Gate, we basically remove irrelevant data in hidden state. Done using Sigmoid activation and point wise multiplication with cell state.
2) In the second step, to know what has to be added to the LSTM's long memory i.e; in cell state with the

help of previous hidden data and input values.Using tanh activated neural network we can combine previous hidden state data and input data to know how much should be updated. Using sigmoid function, we can know whether this input data is worth storing in the cell state. We can say process as the input gate, sigmoid activation is used as the filter. Output of the Sigmoid activation and tanh Neural network are sent to point wise multiplication to get the combined state.

3) In the last step, we need to figure out the Previous hidden state status after adding this input state. We now have the updated cell state, input data and the previous hidden to be updated.Apply sigmoid activation on the previous hidden state and input data and point wise multiplication is done with the tanh applied on the cell state to squished the data.The resultant is our new previous hidden state. Hence, this is called as output gate.

This process is repeated for the every unit of data. Here, in the stock price prediction for each date we consider these 3 steps are repeated. For example, if we consider 50 days of data as input these 3 steps are done in a loop for 50 times.



How do LSTM work ?

In our implementation we have a visible layer which consists of 1 input along with hidden layers of 150 also known as neurons with output layer which gives us single value predictions of the Stock price or Market Value. When it comes to the sigmoid activation it is by default activated in the LSTM blocks. Followed by training/fitting the network for 500 epochs but gets terminated if it is observed that the value loss is has a tolerable difference for more than 15 times. Where we kept the batch size as 1.

V. Performance Evaluation

Performance measure of LSTM is done by using Root Mean Square Error(RMSE) and Mean Absolute Error(MAE).

- To calculate the Mean Absolute Error, we need the absolute difference between actual value and forecasted value and finally get the average of the differences of the testing set.
- To calculate the Root Mean Square Error values, we find the average of square of the absolute difference of actual value and forecasted value and finally find the square root of this average.

VI. Results

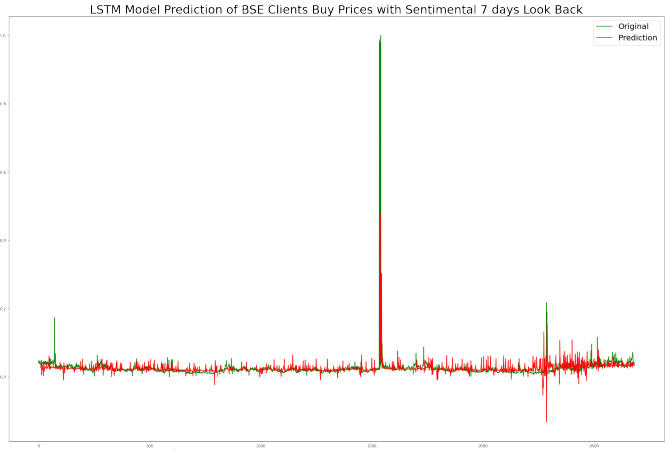
A. Client

Client, is when the trades are executed on client accounts.

When coming to the implementation part, we are able to analyse the client trading that's happening in BSE stock exchange and also forecast with good accuracy. This is achieved with various methods of testing with directly extracted transaction data and also with sentimental analysis which is an influential method from social media and financial news headlines for various look back days :

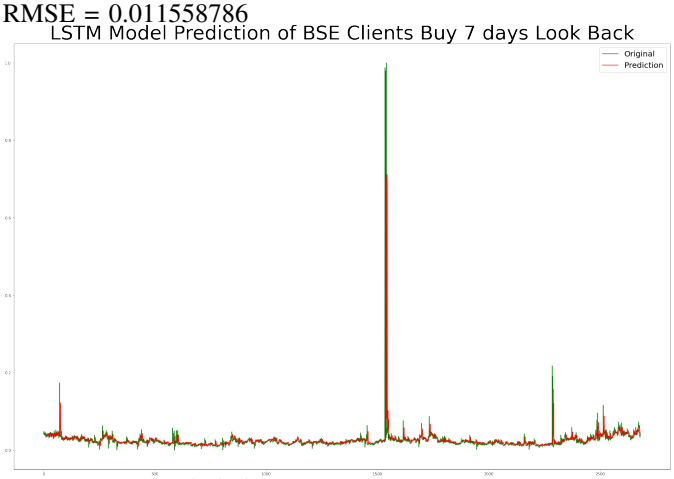
Forecasting of Client Buy Results:

1.Look Back 7 days and without sentimental analysis and with sentimental analysis:



MAE = 0.003388603

RMSE = 0.011558786

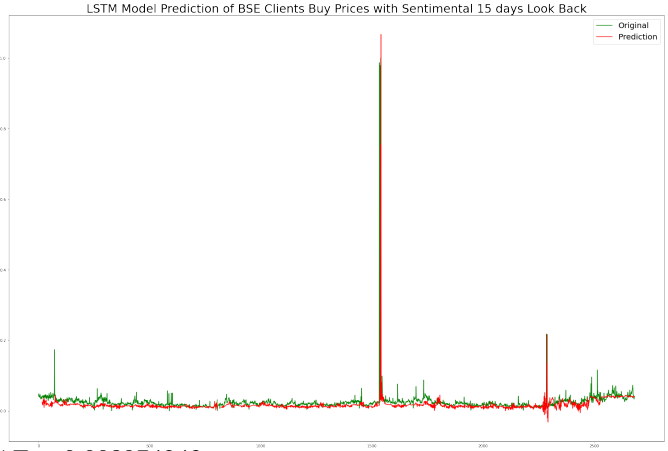


MAE = 0.0054750545

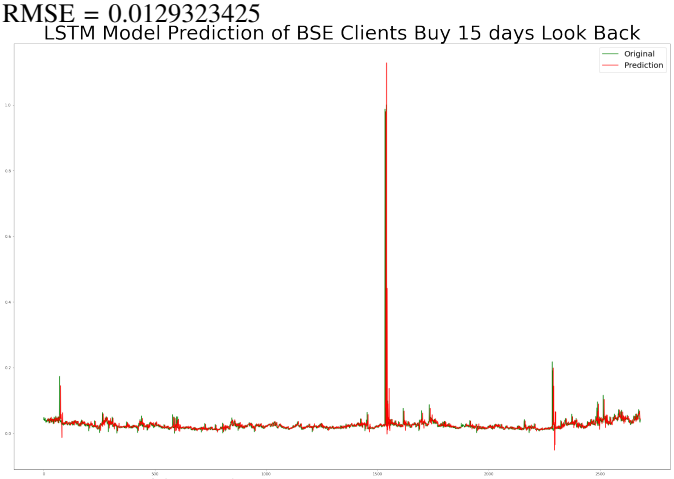
RMSE = 0.029620744

As per the graphs, Root Mean Square Error and Mean Absolute Error we can have conclusion that is approximately 39.11 percent reduction in RMSE value also 60.97 percent reduction in MAE value by introducing Sentiment analysis with look back as 7 days.

2.Look Back 15 days and without sentimental analysis and with sentimental analysis:



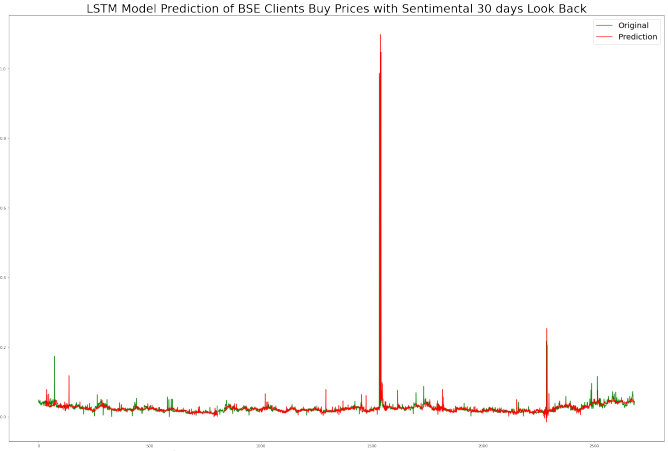
MAE = 0.008974243
RMSE = 0.0129323425



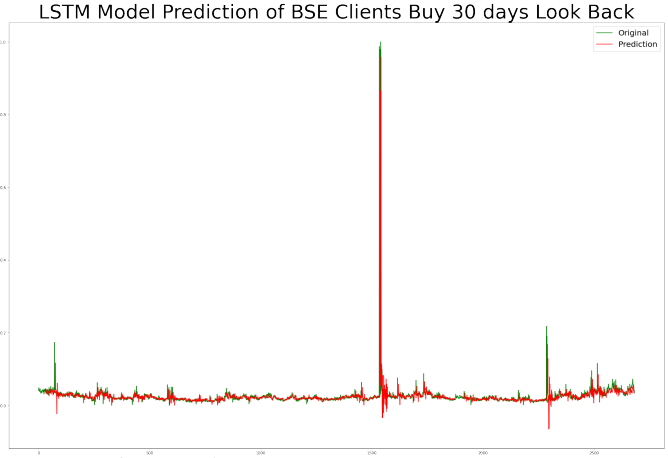
MAE = 0.0065712854
RMSE = 0.01552742

There is almost 26.77 percent reduction in MAE value, 16.71 percent reduction in RMSE value when we included sentiment analysis.

3.Look Back 30 days without sentimental analysis and with sentimental analysis:



MAE = 0.0049145273
RMSE = 0.00869327

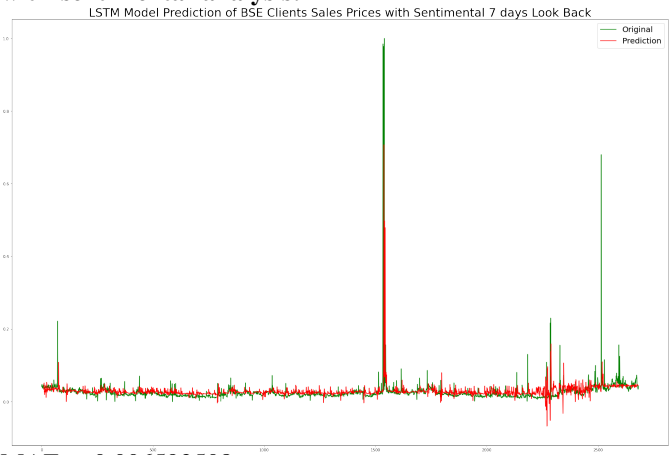


MAE = 0.0068817805
RMSE = 0.015465181

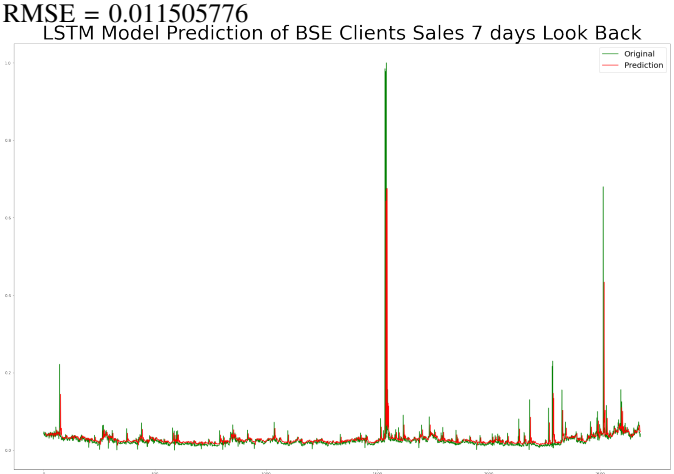
We can certainly observe 28.59 percent, 43.79 percent reduction in MAE, RMSE values when evaluated using sentiment analysis.

Forecasting of Client Sales Results:

1.Look Back 7 days and without sentimental analysis and with sentimental analysis:



MAE = 0.006593508
RMSE = 0.011505776

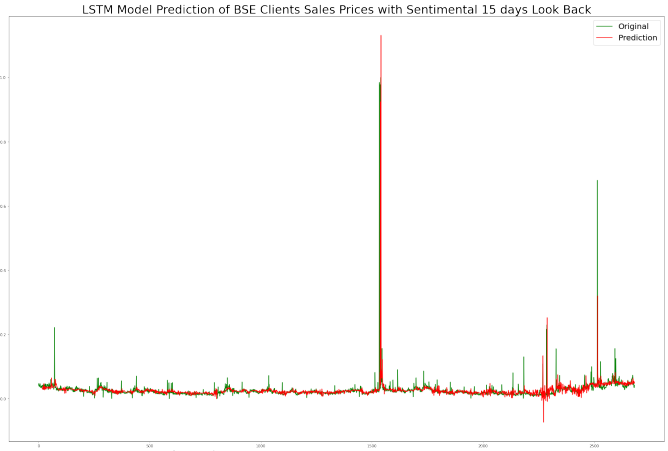


MAE = 0.009787944
RMSE = 0.03165348

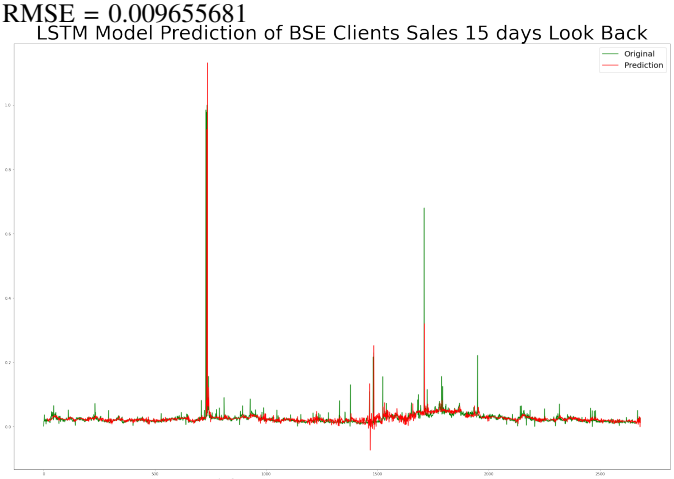
As per the graphs, Root Mean Square Error and Mean Absolute Error we can have conclusion that is approximately

32.64 percent reduction in RMSE value also 63.65 percent reduction in MAE value by introducing Sentiment analysis with look back as 7 days.

2.Look Back 15 days and without sentimental analysis and with sentimental analysis:



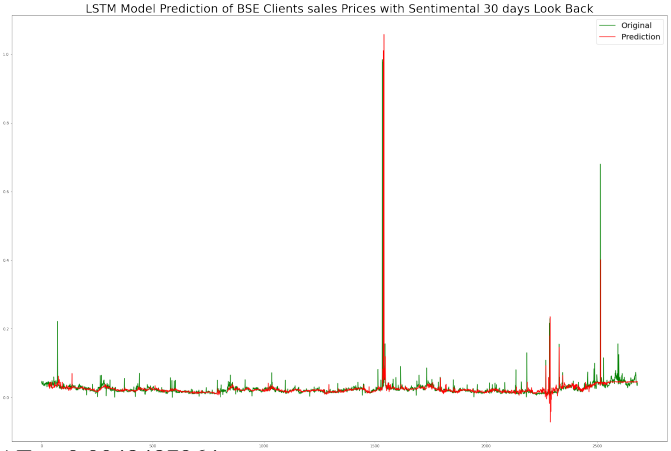
MAE = 0.0070386315
RMSE = 0.009655681



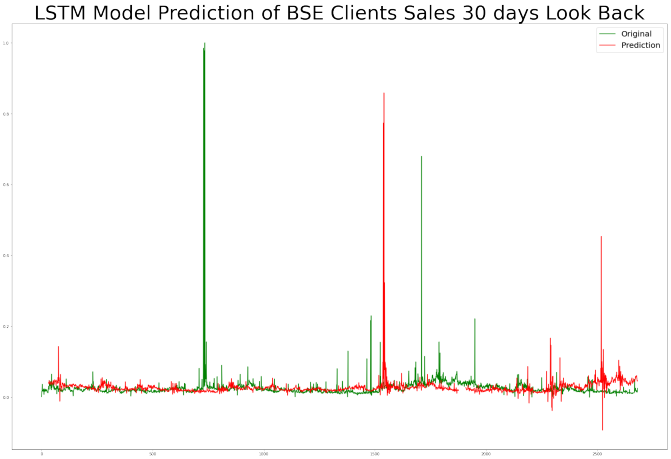
MAE = 0.010930456
RMSE = 0.035689443

There is almost 35.61 percent reduction in MAE value, 72.94 percent reduction in RMSE value when we included sentiment analysis.

3.Look Back 30 days and without sentimental analysis and with sentimental analysis:



MAE = 0.0048487964
RMSE = 0.005966673



MAE = 0.011441373
RMSE = 0.03433521

We can certainly observe 57.62 percent, 82.62 percent reduction in MAE, RMSE values when evaluated using sentiment analysis.

B. Proprietary

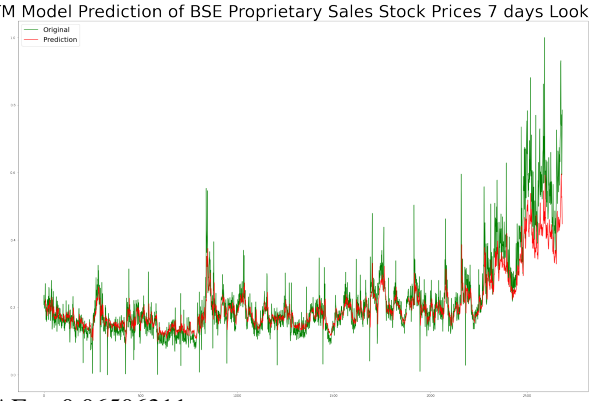
When it comes to proprietary trading what this means is that a trader trades with a firm's money instead of his/her own in order for the firm to gain profits. These may include stocks, bonds, currencies and etc that can be traded in the stock market.

For implementation we were able to analysis the proprietary trading that occurred in BSE stock market and also forecast the with greater accuracy. This was achieved with various scenarios of testing with plain transaction data and also with social media and financial news headlines sentimental analysis as influence for varying days for look back:

Forecasting of Proprietary Sales Results:

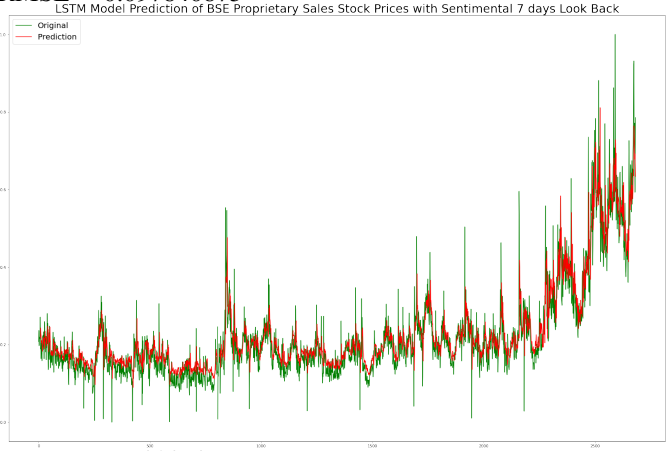
1.Look Back 7 days without sentimental analysis and with sentimental analysis:

LSTM Model Prediction of BSE Proprietary Sales Stock Prices 7 days Look Back



MAE = 0.06506311
RMSE = 0.097840875

LSTM Model Prediction of BSE Proprietary Sales Stock Prices with Sentimental 7 days Look Back

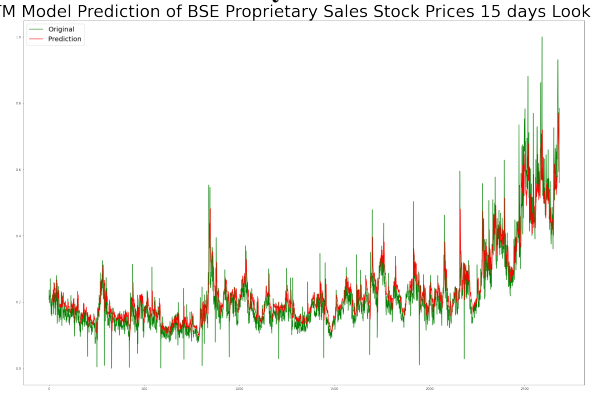


MAE = 0.02656468
RMSE = 0.034758694

On comparing both these graphs and error results we can conclude is that there is a change of 64.47 percent in RMSE error also 59.17 percent in MAE error with the introduction of sentimental analysis with having the look back as 7 days.

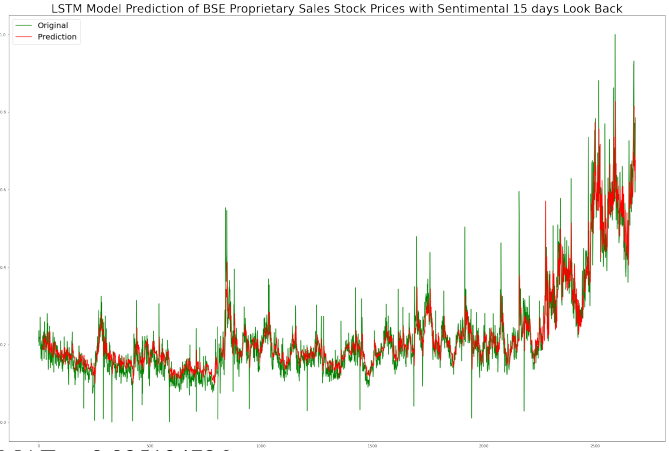
2.Look Back 15 days and without sentimental analysis and with sentimental analysis:

LSTM Model Prediction of BSE Proprietary Sales Stock Prices 15 days Look Back



MAE = 0.04740042
RMSE = 0.06720729

LSTM Model Prediction of BSE Proprietary Sales Stock Prices with Sentimental 15 days Look Back

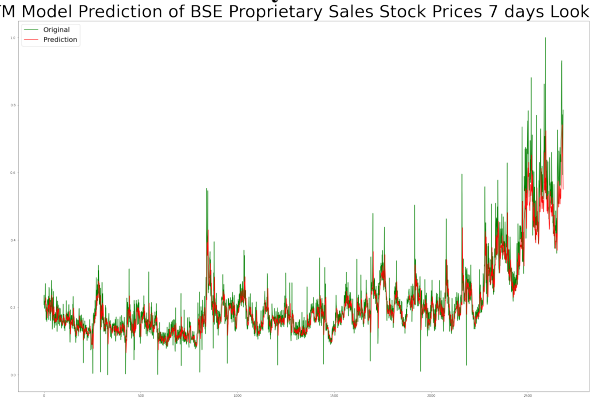


MAE = 0.025134796
RMSE = 0.033502832

On comparing both these graphs and error results we can conclude is that there is a change of 50.15 percent in RMSE error also 46.97 percent in MAE error with the introduction of sentimental analysis with having the look back as 15 days.

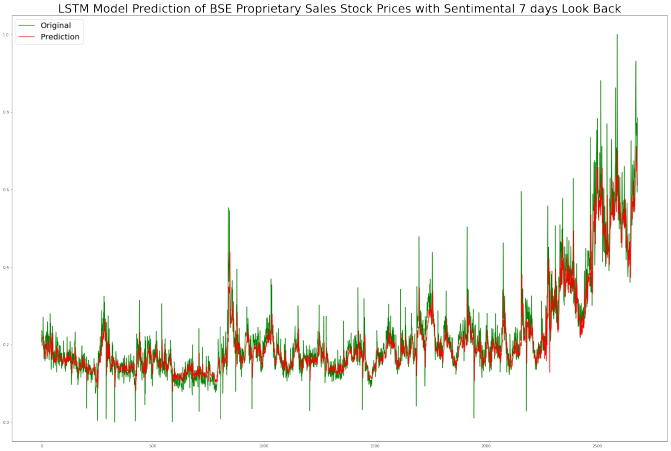
3.Look Back 30 days and without sentimental analysis and with sentimental analysis:

LSTM Model Prediction of BSE Proprietary Sales Stock Prices 7 days Look Back



MAE = 0.04622454
RMSE = 0.07039143

LSTM Model Prediction of BSE Proprietary Sales Stock Prices with Sentimental 7 days Look Back



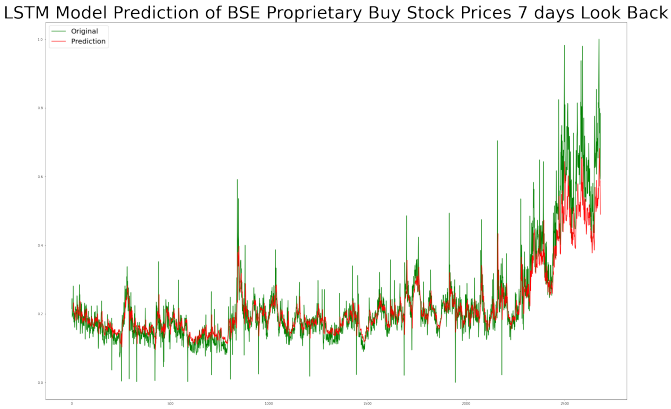
MAE = 0.02190594
RMSE = 0.030971821

On comparing both these graphs and error results we can

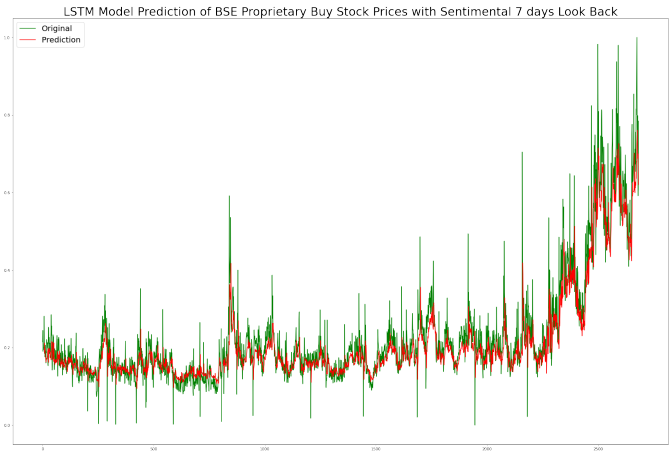
conclude is that there is a change of 56 percent in RMSE error also 52.61 percent in MAE error with the introduction of sentimental analysis with having the look back as 30 days.

Forecasting of Proprietary Buys Results:

1.Look Back 7 days and without sentimental analysis and with sentimental analysis:



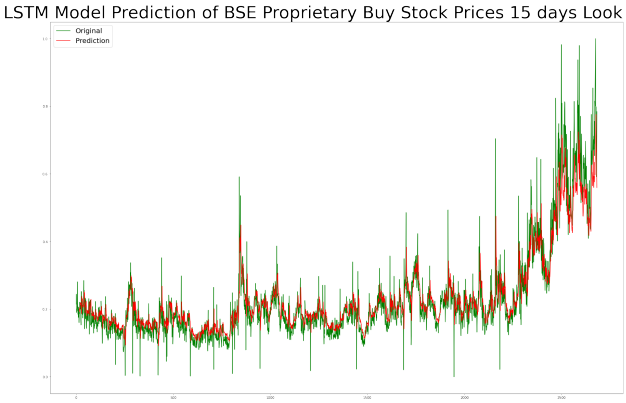
MAE = 0.0633481
RMSE = 0.09530285



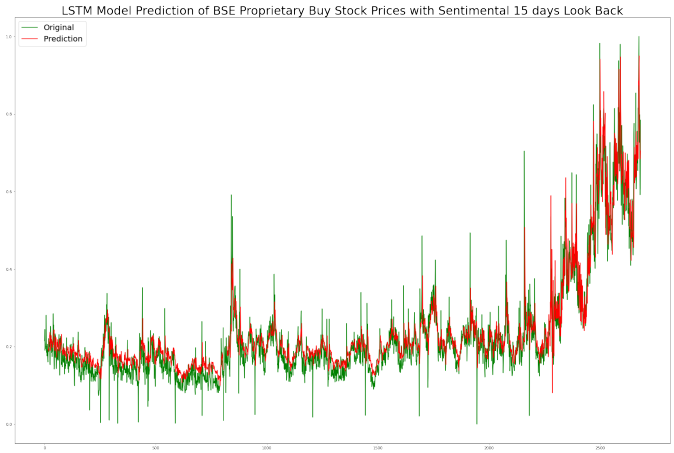
MAE = 0.02272685
RMSE = 0.032000337

On comparing both these graphs and error results we can conclude is that there is a change of 64.47 percent in RMSE error also 66.42 percent in MAE error with the introduction of sentimental analysis with having the look back as 7 days.

2.Look Back 15 days and without sentimental analysis and with sentimental analysis:



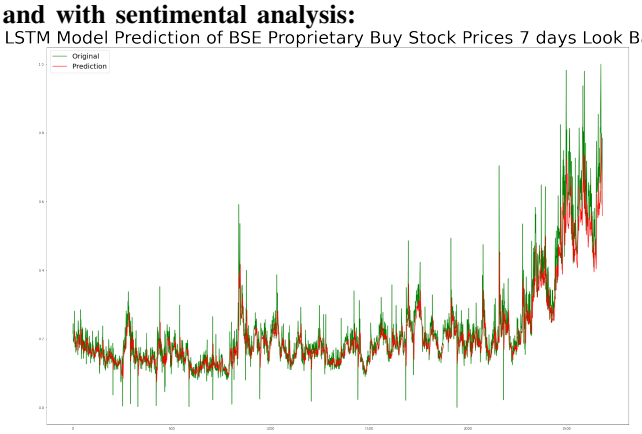
MAE = 0.050764315
RMSE = 0.07603101



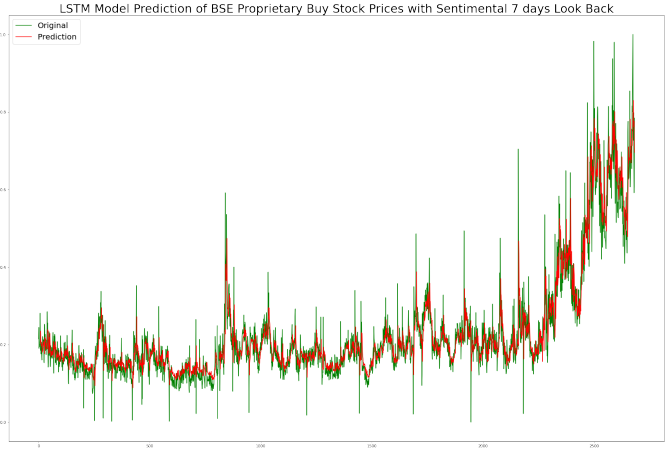
MAE = 0.029201943
RMSE = 0.037482142

On comparing both these graphs and error results we can conclude is that there is a change of 50.70 percent in RMSE error also 42.48 percent in MAE error with the introduction of sentimental analysis with having the look back as 15 days.

3.Look Back 30 days and without sentimental analysis and with sentimental analysis:



MAE = 0.052677557
RMSE = 0.07954283



MAE = 0.023039889
RMSE = 0.031927396

On comparing both these graphs and error results we can conclude is that there is a change of 59.86 percent in RMSE error also 56.26 percent in MAE error with the introduction of sentimental analysis with having the look back as 30 days.

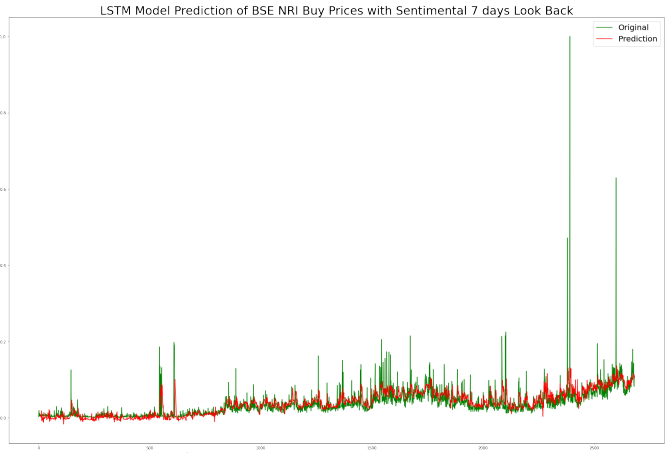
C. NRI - Non Residential Indian

NRI, is an Indian citizen residing outside India for a sum up days of more than 183 is considered to be NRI. NRIs are eligible to vote, and most importantly the earnings made in India are valued for taxation.

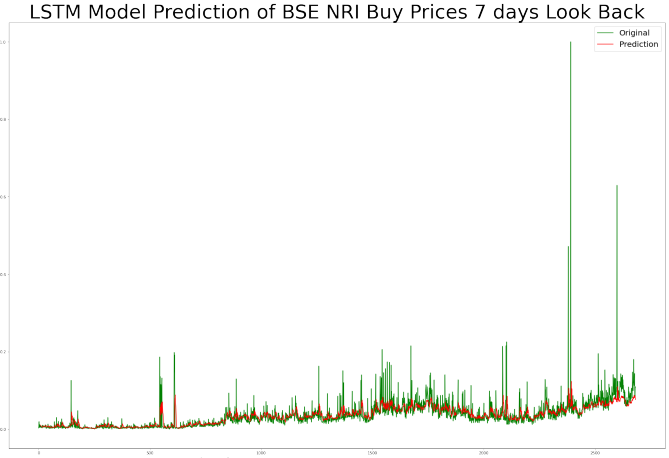
When coming to the implementation part, we are able to analyse the NRI trading that's happening in BSE stock exchange and also forecast with good accuracy. This is achieved with various methods of testing with directly extracted transaction data and also with sentimental analysis which is an influential method from social media and financial news headlines for various look back days :

Forecasting of NRI Buy Results:

1.Look Back 7 days and without sentimental analysis and with sentimental analysis:



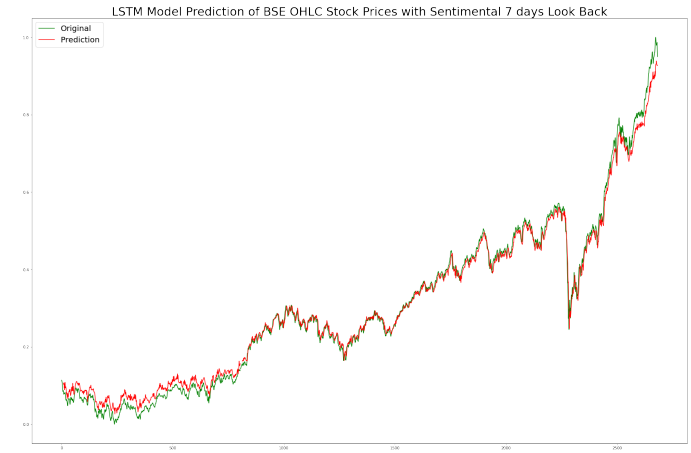
MAE = 0.008837361
RMSE = 0.017581824



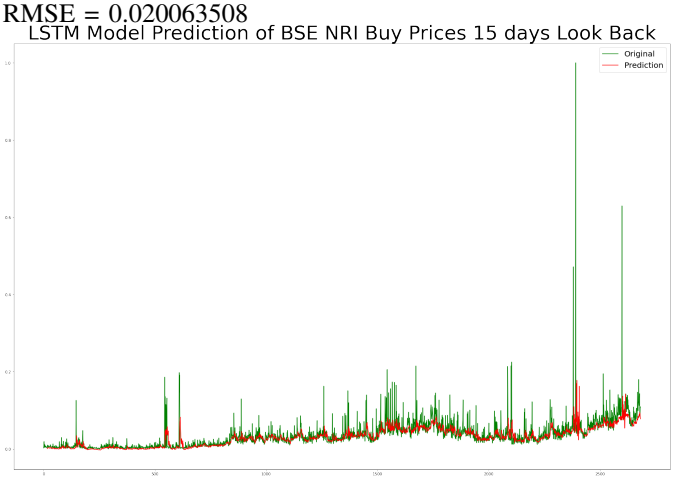
MAE = 0.009830546
RMSE = 0.047404557

As per the graphs, Root Mean Square Error and Mean Absolute Error we can have conclusion that is approximately 10.10 percent reduction in MAE value also 62.91 percent reduction in RMSE value by introducing Sentiment analysis with look back as 7 days.

2.Look Back 15 days and without sentimental analysis and with sentimental analysis:



MAE = 0.014909723
RMSE = 0.020063508

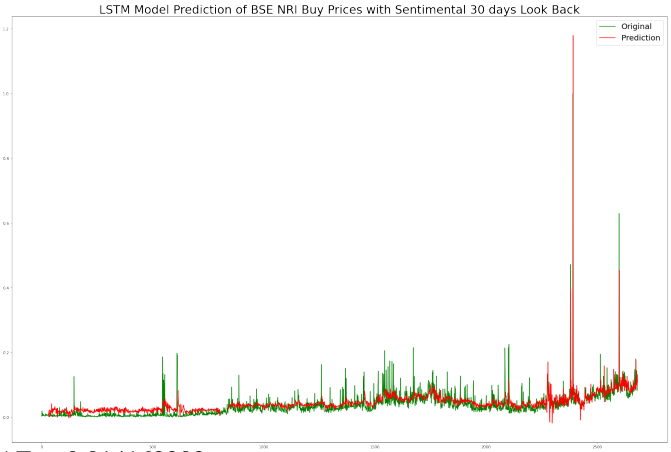


MAE = 0.017097829

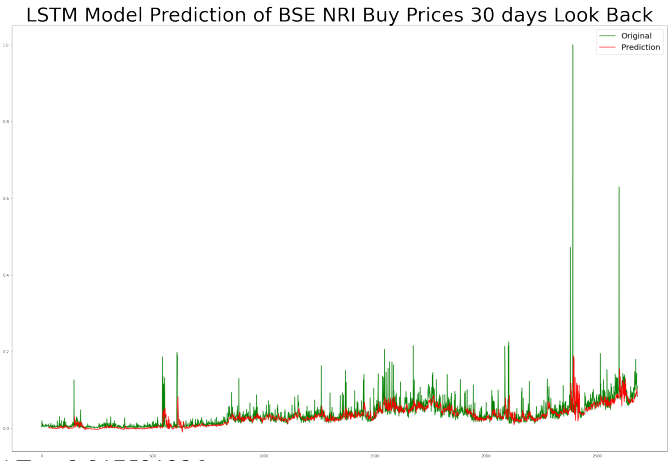
RMSE = 0.048306443

There is almost 12.80 percent reduction in MAE value, 58.46 percent reduction in RMSE value when we included sentiment analysis.

3.Look Back 30 days and without sentimental analysis and with sentimental analysis:

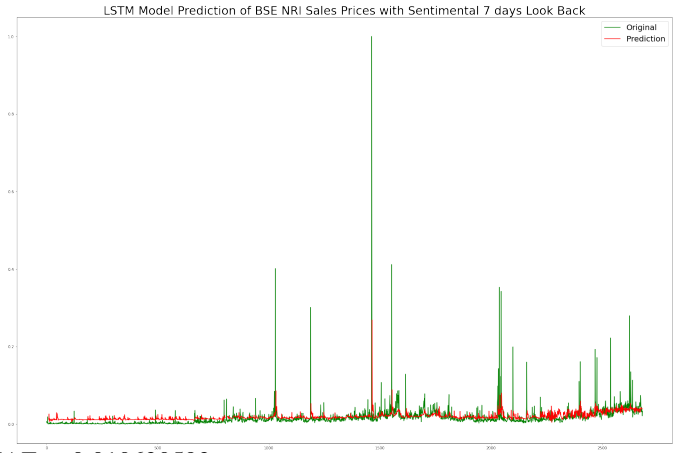


MAE = 0.014162203
RMSE = 0.021084223

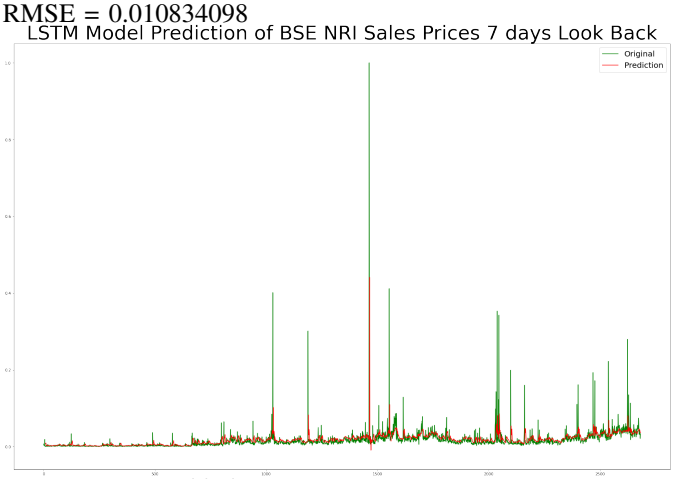


MAE = 0.017591286
RMSE = 0.049028702
We can certainly observe 19.49 percent, 56.99 percent reduction in MAE, RMSE values when evaluated using sentiment analysis.

Forecasting of NRI Sales Results:
1.Look Back 7 days and without sentimental analysis and with sentimental analysis:

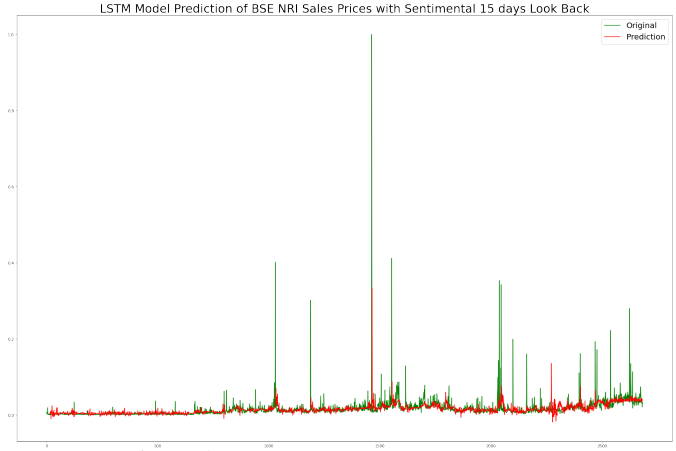


MAE = 0.010630593
RMSE = 0.010834098



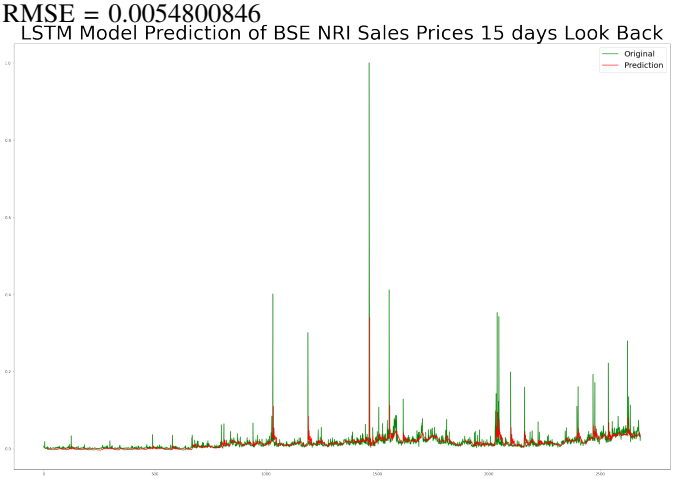
MAE = 0.010996696
RMSE = 0.028052555
As per the graphs, Root Mean Square Error and Mean Absolute Error we can have conclusion that is approximately 3.33 percent reduction in MAE value also 57.65 percent reduction in RMSE value by introducing Sentiment analysis with look back as 7 days.

2.Look Back 15 days and without sentimental analysis and with sentimental analysis:



MAE = 0.0036971958

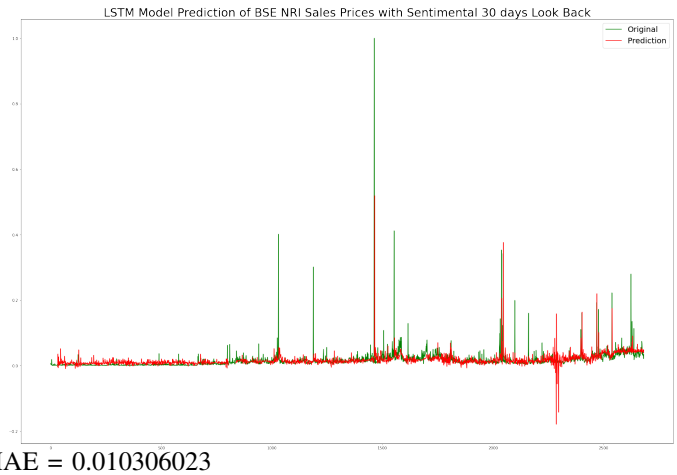
RMSE = 0.0054800846



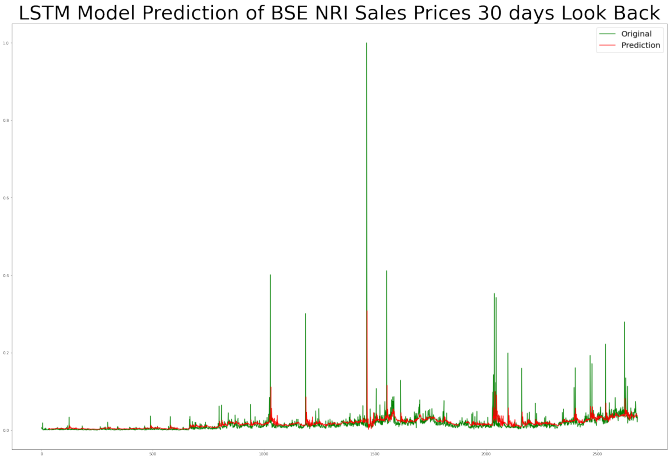
MAE = 0.010098477
RMSE = 0.0282736

There is almost 63.39 percent reduction in MAE value, 80.62 percent reduction in RMSE value when we included sentiment analysis.

3.Look Back 30 days and without sentimental analysis and with sentimental analysis:



MAE = 0.010306023
RMSE = 0.008943314



MAE = 0.011090927
RMSE = 0.028352953

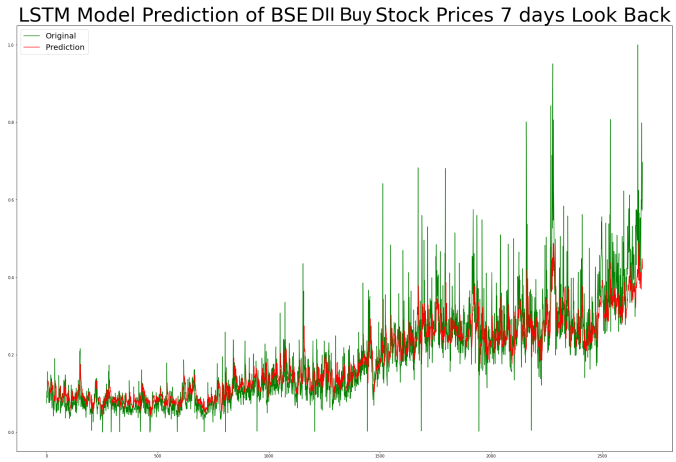
We can certainly observe 46.97 percent, 68.45 percent reduction in MAE, RMSE values when evaluated using sentiment analysis.

D. Domestic Institutional Investors

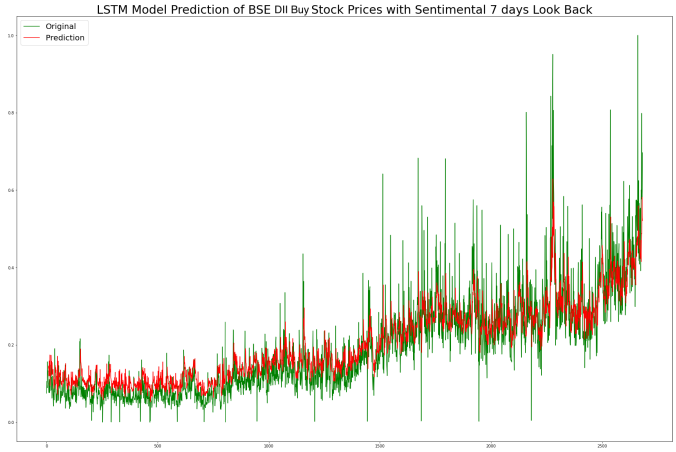
DII refers to Indian institutional investors who are investing in finanacial market in INDIA for eg:stock market.

Forecasting of Domestic Institutional Investors Buy Results:

1.Look Back 7 days and without sentimental analysis and with sentimental analysis:



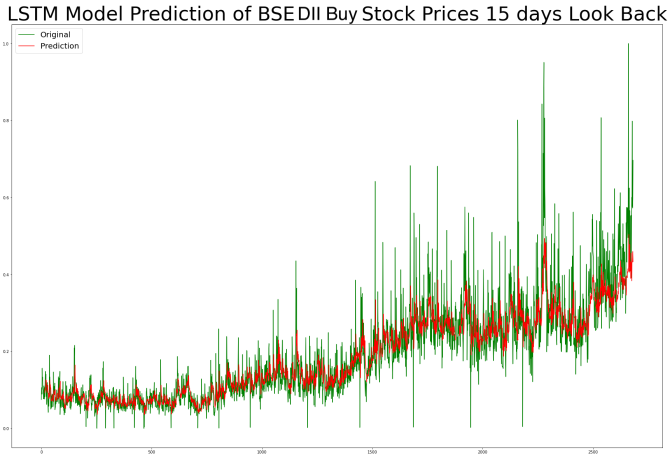
MAE = 0.0647
RMSE = 0.0956



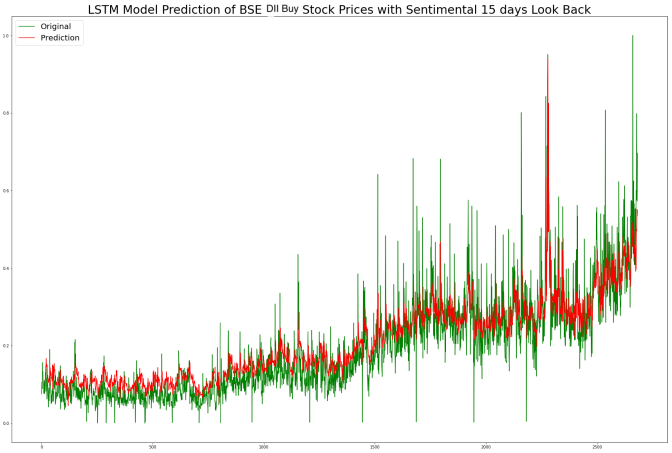
MAE = 0.0331
RMSE = 0.0391

On comparing both these graphs and error results we can conclude is that there is a change of 56 percent in RMSE error also 31 percent in MAE error with the introduction of sentimental analysis with having the look back as 7 days.

1.Look Back 15 days and without sentimental analysis and with sentimental analysis:



MAE = 0.06356
RMSE = 0.09314

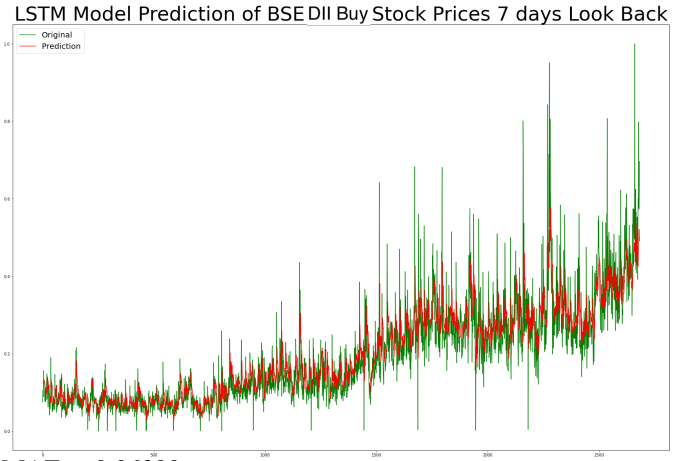


MAE = 0.034181
RMSE = 0.04000

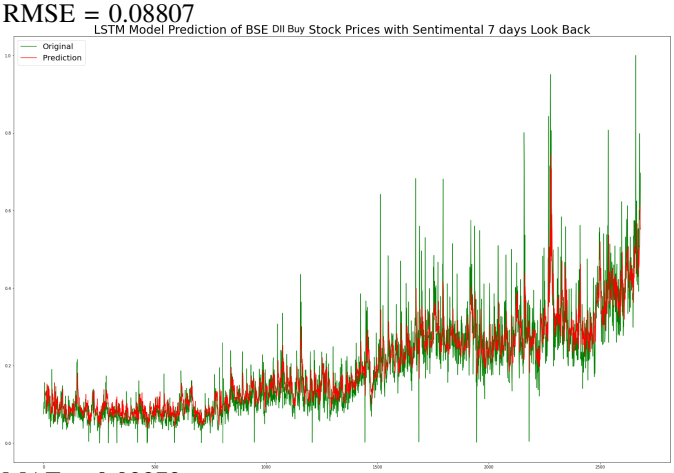
On comparing both these graphs and error results we can conclude is that there is a change of 53 percent in RMSE error also 29 percent in MAE error with the introduction of sentimental analysis with having the look back as 15 days.

Forecasting of Domestic Institutional Investors Buy Results:

3.Look Back 30 days and without sentimental analysis and with sentimental analysis:



MAE = 0.06300
RMSE = 0.08807

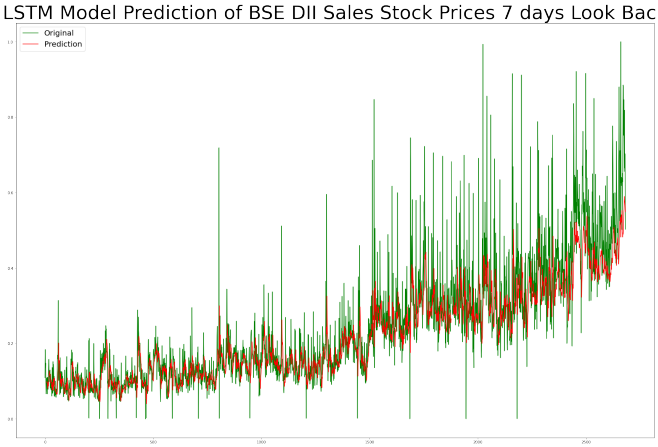


MAE = 0.02272
RMSE = 0.02917

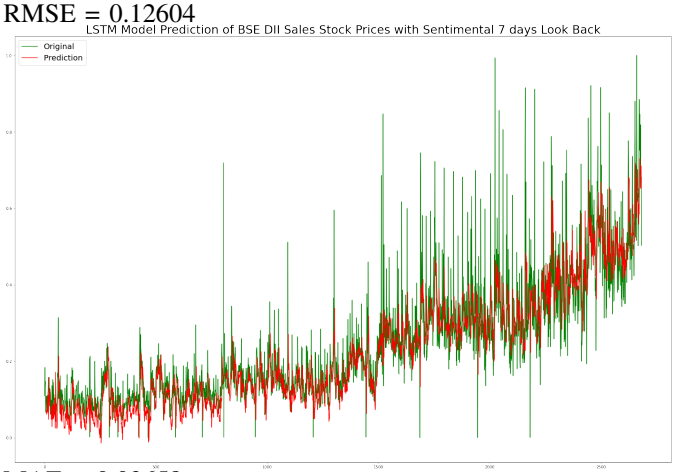
On comparing both these graphs and error results we can conclude is that there is a change of 59 percent in RMSE error also 40 percent in MAE error with the introduction of sentimental analysis with having the look back as 30 days.

Forecasting of Domestic Institutional Investors sales Results:

1.Look Back 7 days and without sentimental analysis and with sentimental analysis:



MAE = 0.08789
RMSE = 0.12604

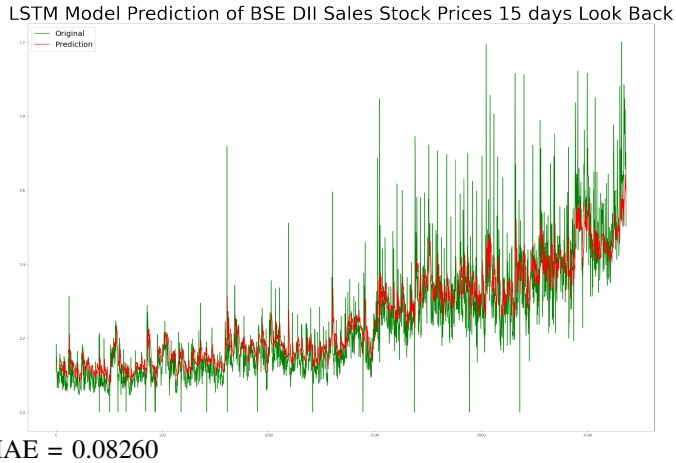


MAE = 0.03658
RMSE = 0.04727

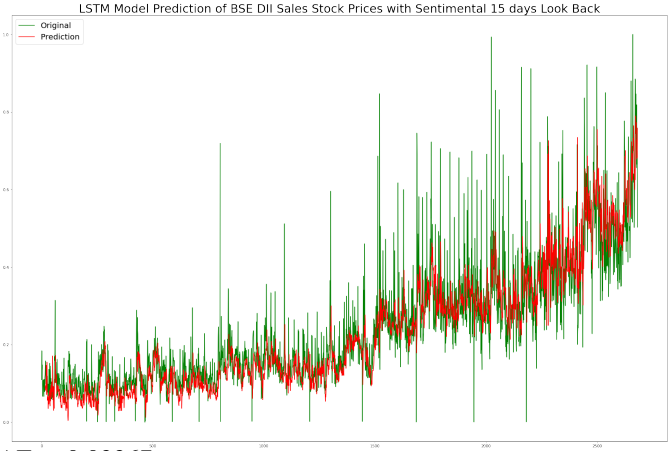
On comparing both these graphs and error results we can conclude is that there is a change of 80 percent in RMSE error also 50 percent in MAE error with the introduction of sentimental analysis with having the look back as 7 days.

Forecasting of Domestic Institutional Investors sales Results:

2.Look Back 15 days and without sentimental analysis and with sentimental analysis:



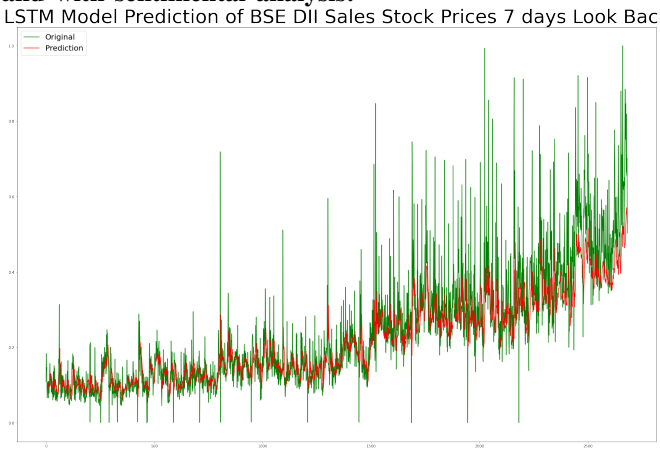
MAE = 0.08260
RMSE = 0.11634



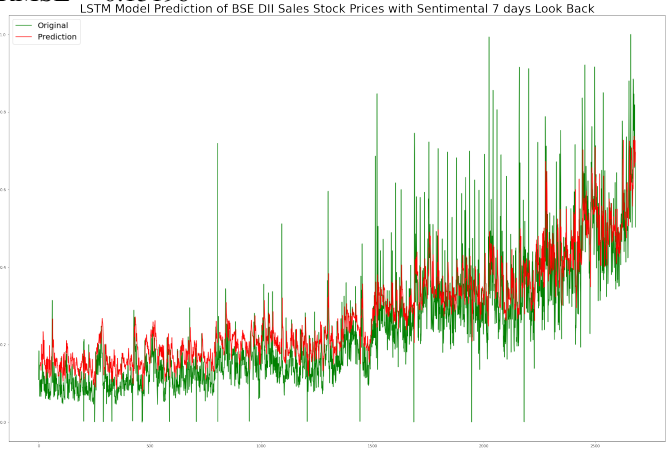
MAE = 0.03367
RMSE = 0.0444

On comparing both these graphs and error results we can conclude is that there is a change of 70 percent in RMSE error also 50 percent in MAE error with the introduction of sentimental analysis with having the look back as 15 days.

3.Look Back 30 days and without sentimental analysis and with sentimental analysis:



MAE = 0.09283
RMSE = 0.13196



MAE = 0.05538
RMSE = 0.06255

] On comparing both these graphs and error results we can conclude is that there is a change of 70 percent in RMSE error also 40 percent in MAE error with the introduction of sentimental analysis with having the look back as 30 days.

E. Forecasting the BSE Stock Market Prices

In this section what we have done is we calculated a variable call OHLC which stores average of the 4 Open, High, Low and Close values. Which we used to forecast the BSE stock market.

Forecasting of OHLC BSE Stock Market Prices:

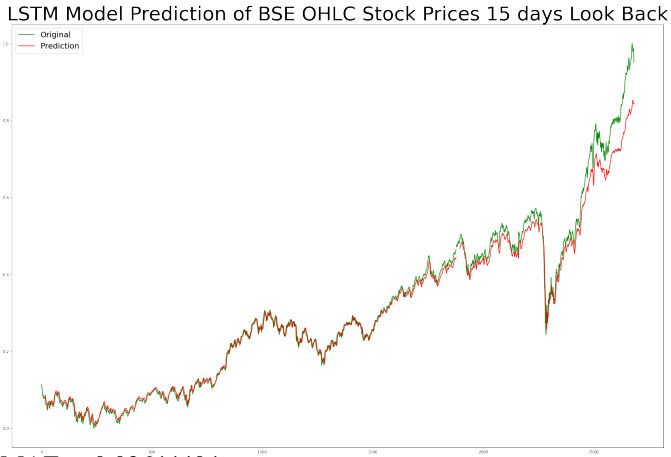
1.Look Back 7 days and without sentimental analysis and with sentimental analysis:

MAE = 0.012060656
RMSE = 0.015673973

MAE = 0.008570193
RMSE = 0.010936281

On comparing both these graphs and error results we can conclude is that there is a change of 30.23 percent in RMSE error also 28.94 percent in MAE error with the introduction of sentimental analysis with having the look back as 7 days.

2.Look Back 15 days and without sentimental analysis and with sentimental analysis:



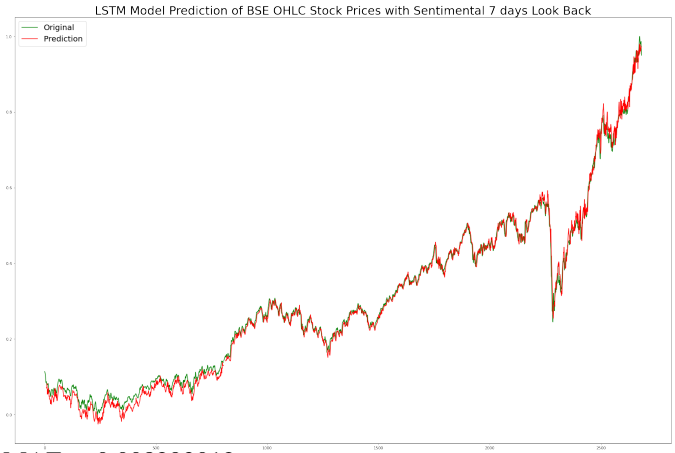
MAE = 0.03644404
RMSE = 0.048765462

MAE = 0.007780917
RMSE = 0.010715311

On comparing both these graphs and error results we can conclude is that there is a change of 78.02 percent in RMSE error also 78.65 percent in MAE error with the introduction of sentimental analysis with having the look back as 15 days.

3.Look Back 30 days and without sentimental analysis and with sentimental analysis:

MAE = 0.017996717
RMSE = 0.025944564



MAE = 0.008293012
RMSE = 0.011648685
On comparing both these graphs and error results we can conclude is that there is a change of 55.10 percent in RMSE error also 53.92 percent in MAE error with the introduction of sentimental analysis with having the look back as 30 days.

VII. Conclusion

In conclusion to our project work on Stock price prediction using heterogeneous data it can be said that the deviations observed in every category by taking different look back days:

A. Client

From Clients Buy it can be observed that there has been a substantial reduction of error for both MAE and RMSE for varying look back days with the of use sentiment analysis. That being said with sentimental and without sentimental has shown an average reduced error of 31.49 percent for MAE and 33.20 percent for RMSE for varying number of look back days. Overall what can be said about Client Buy data predictions is that the moderate number of look back days along with having the sentimental analysis has given the prediction a higher accuracy. Which was the case for look back days as 15 and with sentimental analysis gave the least MAE = 0.0065712854 and RMSE = 0.01552742.

On the contrary to Clients Buy, Clients Sales has seen reduced error changes in the values of MAE and RMSE for various look back days when sentiment analysis at a more substantial reduction then Client Buy. The implementation of the influential feature made the LSTM model to perform better in comparison prediction to plain transaction data. Where the predictions with sentimental analysis on comparison with, without sentimental analysis has shown a average reduced error of 52.29 percent for MAE and 62.73 percent for RMSE. Under consideration of all the outcomes here is that the better performing scenario is the one with look back days as 15 and with sentimental analysis and with a higher accuracy in comparison to the rest and best improvement with MAE = 0.010930456 and RMSE = 0.035689443.

B. Proprietary

In Proprietary Buy there is considerable amount of change in MAE and RMSE values while using sentiment analysis and without. That being said there has been an average improvement of 56.05 percent in MAE and 58.34 percent in RMSE using sentimental analysis then without. After observing the all the three scenarios is that it can be observed that having look back as 30 days along with sentimental analysis has yielded smaller values for MAE and RMSE along with the greatest change of RMSE and MAE values reduced. With MAE = 0.023039889 and RMSE = 0.031927396.

In comparison to Proprietary Buy, Proprietary Sales has also done equally well with reducing MAE and RMSE values when using sentiment analysis for different look back days. Where on average the MAE value has reduced by 52.92 percent and RMSE value has reduced by 56.87 percent with the use of sentiment analysis to with sentiment analysis. Followed by comparing the all the results we can say that with having look back days as 30 and with sentimental analysis has provided with the most accurate results and with the least MAE = 0.02190594 and RMSE = 0.030971821 values.

C. NRI

When it comes to NRI Buy there has been a change in the MAE and RMSE values being reduced with the sentimental analysis as a extra feature and various look back days. Where on average the MAE error was reduced by 14.13 percent and RMSE error is reduced by 59.45 with the use of sentimental analysis for different look back days. Altogether, the LSTM model performed best when it came to the look back days being 7 and with sentimental analysis have proven to have yielded the best results with MAE = 0.008837361 and RMSE = 0.017581824 to have been the most accurate.

On the contrary with NRI Sale we have again seen the MAE and RMSE have performed better with sentimental analysis on varying look back dates. On average the MAE value has decremented by 37.90 percent and the RMSE value decremented by 68.90 percent with sentimental analysis. From consideration of all the result we can say that with having the look back as 30 days with sentimental analysis has given the best accuracy with the least error for MAE = 0.010306023 and RMSE = 0.008943314 as the best result.

D. DII

In DII buy there is a decrements observed in the changes of MAE and RMSE values with sentiment analysis for different look back days. Perhaps, decrements is mostly observed in RMSE values but not much in MAE values. On an average it is observed a 40.61 percent reduction in MAE value and 29.67 percent reduction with RMSE. We can say that with the look back as 7 days with sentimental analysis has given

the best accuracy with the least error for MAE = 0.0331 and RMSE = 0.0391.

In DII sales there is a decrements observed in the changes of MAE and RMSE values with sentiment analysis for different look back days. Perhaps, decrements is mostly observed in RMSE values but not much in MAE values. On an average it is observed a 55.61 percent reduction in MAE value and 69.67 percent reduction with RMSE. We can say that with the look back as 7 days with sentimental analysis has given the best accuracy with the least error for MAE = 0.033671 and RMSE = 0.04444.

E. OHLC

Being the most popular feature of predicting stock market, there is a substantial increase in change in MAE and RMSE values while using sentiment analysis and without. It is observed that there has been an average improvement of 53.76 percent in MAE value and 55.98 percent in RMSE value using sentiment analysis than without. After observing the three scenarios that having look back days as 15 along with sentiment analysis has acquired smaller values of MAE and RMSE along with highest change of RMSE and MAE reduction. Here, MAE = 0.007780917 and RMSE = 0.0107151311.

In the end what can be said is that there has been a substantial difference to the stock market being predicted with plain old transaction data to heterogeneous data. Accordingly our model with LSTM was able to reduce the error over all on the scenarios by on average 43.86 percent for MAE and 50.93 percent for RMSE. In all we can say and conclude is that the accuracy of prediction of stock mark is better handled with heterogeneous data with its respective sentimental analysis over the traditional method on analysis of only transnational data. After all the success, when it came to the limitations what was observed was that for social and online news headlines were very limited for the India market. Hence, what can be said is that with a greater wide range of more social and news headlines we can be assured that we will be achieving greater results with SA.

VIII. Abbreviations and Acronyms

LSTM : Long Short Term Memory
OHLC : Open-High-Low-Close
RNN : Recurrent Neural Network
DII : Domestic Institutional Investors

IX. References

[1]Polamuri Subba , Srinivas, Kudipudi Mohan, A.. (2020). A Survey on Stock Market Prediction Using Machine Learning Techniques. 10.1007/978-981-15-1420-3_101.

[2] S. Mohan, S. Mullapudi, S. Sammeta, P. Vijayvergia and D. C. Anastasiu, "Stock Price Prediction Using News Sentiment Analysis," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 2019, pp. 205-208, doi: 10.1109/BigDataService.2019.00035.

[3]Rubi Gupta & Chen, Min. (2020). Sentiment Analysis for Stock Price Prediction. 213-218. 10.1109/MIPR49039.2020.00051.

[4]Transformers for Time Series Forecasting , Natasha Killenbrunn

[5]Usmani S, Shamsi JA. News sensitive stock market prediction: literature review and suggestions. PeerJ. Computer Science. 2021 ;7:e490. DOI: 10.7717/peerj-cs.490. PMID: 34013029; PMCID: PMC8114814.