Use of Heterogeneous Data for forecasting of stock market

Under the Guidance of : Dr. Ranjana Vyas

Submitted by:

Dhanush Vasa
IIT2019208

Pedada Gopal
IIT2019065

Kandagatla Meghana Santhoshi
IIB2019030

Mitta Lekhana Reddy
IIT2019204

Problem Statement

Stock Market Client Category Wise (Client, NRI, Proprietary, and Domestic Institutional Investor) Turnover and Stock Market Price of BSE using sentiment analysis.

TABLE OF CONTENTS

01
Introduction

What is Stock Market Prediction?

- Stock Market Prediction aims to forecast the movement of the stock values of a financial exchange.
- It has 4 main entities which affect the stock value:
  1. Social
  2. Psychological
  3. Political
  4. Economic

Entities Affecting The Stock Values

Social

1

Social factors have an increased attention to stock's volatility and is more significant than public sentiment

Psychological

2

Psychology factors have an enormous effect on large market swings which could trigger emotions and leading to fear-based trading

Political

3

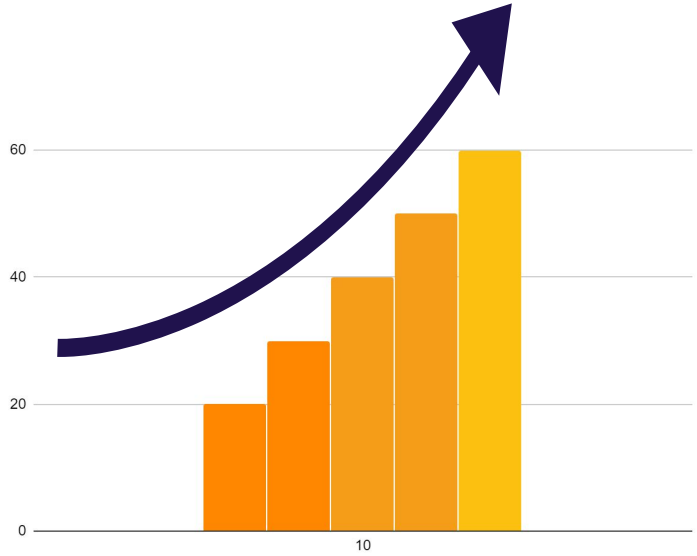
A country's government shapes how a company executes its operations and regulation, taxations and etc cause effects within the companies

Economic

4

Stock exchange firm's profits are directly proportional to the behavior of the real economy, hence stock prices are affected by the expectations of the future

Importance of Reliability In Stock Market Prediction



- When associated with reliability in the Stock Market Prediction is very serious as each and every action could have a game changing effect or a devastating result.
- Hence, as the stock market is very sensitive and with a reliable predictor will result in the market being more stable.

02
Literature Survey

Authors : Polamuri Subba Rao, K. Srinivas, A. Krishna Mohan

- [1]Stock Prediction using Machine Learning Techniques.
- Published in May 2020, Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications, Springer Link.
- **Data set** : Collected from Indian Stock Market Websites
- **Models Used** : ARIMA ,Holt-Winters, Artificial Neural Network, Hidden Markov Model,RNN
- **Results** : This paper provides a review and comparative analysis of different stock market prediction parameter techniques. These techniques are used to evaluate stock market performance and trends. The stock market forecasting system is to increase accuracy. In this study to analyze a novel approach to improve the prediction of the results of stock, it means we will c o or more methods to construct a novel approach method.

Authors : Saloni Mohan1, Sahitya Mullapudi, Sudheer Sammeta1, Parag Vijayvergia and David C. Anastasiu,

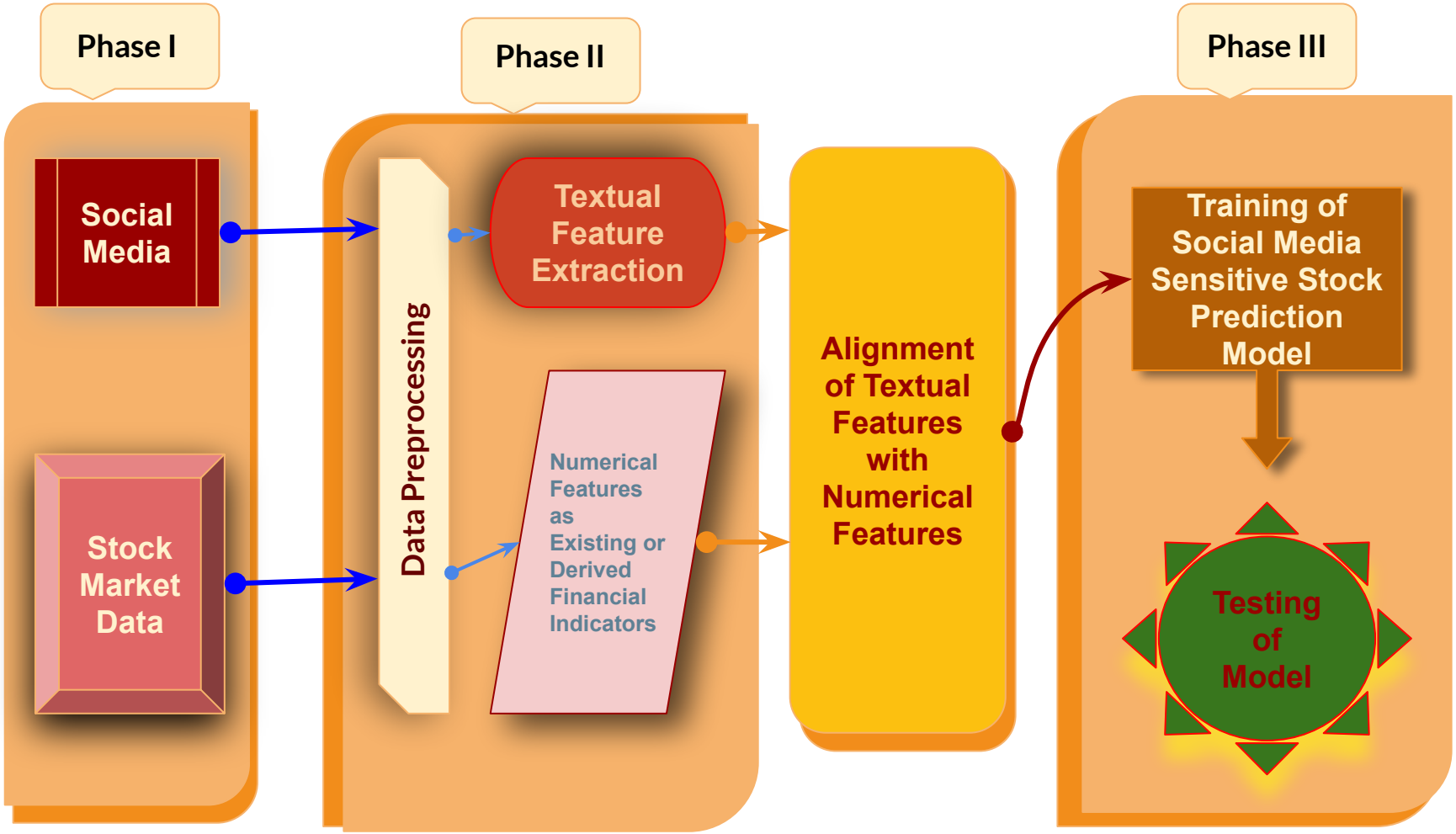
- [3]Stock Prediction using News Sentiment Analysis.
- Published in 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)
- **Data set** : Collected from Financial news articles, for S&P 500 companies for five years . Applied Log Transformation on stock prices to reduce difference between high and low stock prices.
- **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 .The models did not perform well in cases where stock prices are low or highly volatile.

Authors : Rubi Gupta, Min Chen

- [6]Performed the sentiment analysis on collected StockTwits data using three machine learning methods (Naïve Bayes, SVM, and logistic regression) and five featurization techniques (bag of words, bigram, trigram, TF-IDF, and LSA)
- **Published in:** 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR).
- Sentiment analysis is done using logistic regression and TF-IDF, and then aggregated using percentage of positive sentiments method
- **Dataset used**: For correlation analysis, both stock and StockTwits data are collected for the duration of January 1st, 2019 to September 30th, 2019, with 120 days of data used for training and remaining for testing.
- **Result :** The combination of logistic regression and TF-IDF is able to achieve reasonably high accuracy level, between 75% and 85%, for all the five companies.

03

Methodologies


Phase I
Phase II
Phase III
Social Media
Stock Market Data
Data Preprocessing
Textual Feature Extraction
Numerical Features as Existing or Derived Financial Indicators
Alignment of Textual Features with Numerical Features
Training of Social Media Sensitive Stock Prediction Model
Testing of Model

Methodology

Data collection

- For this Project we need, Two types of data. Historical transactional data and Social Media Data/News Headlines.
- Historical Data we got two different datasets , one is client wise turnover and Normal daily stock prices of open, High, Low and closed.
- Client wise turnover is divided into four categories Client, NRI, Proprietary , DII.
- We archived Historical dataset dated from January 2011 to October 2021.
- To collect Social Media data, we used Twitter and Reddit using API. We used Financial news headlines , used web scraping with parsehub with these websites which provides a API and csv of the data.

Methodology

Data Cleaning and Preprocessing

- For the historical transactional data , pre-processing of the data was not necessary.
- Coming to the textual data, the following steps are done to pre-process the data.
- If there are any duplicates in the data, we removed them first and then according to the date, all the information is concatenated.
- Using regex commands in python, we cleaned all the spaces and anything that is not an alphabet and replacing with space.
- If there is any null and empty location , we inserted “neutral” instead.
- This makes the NLTK model to classify in positive , negative and neutral emotion more simple.

FinalDATA

Date	Clients Buy	Clients Sales	Client Net	NRI Buy	NRI Sales	NRI Net	Proprietary Buy	Proprietary Sales	Proprietary Net	DII Buy	DII Sales	DII Net	Open	High	Low	Close	Text	SA
11-10-2021	3097.08	3435.48	-338.4	27	24.33	2.67	2390.26	2075.32	314.94	6585.38	6958.66	-373.28	60099.68	60476.13	59811.42	60135.78	Sensex Nifty er	0.6808
12-10-2021	2784.76	2853.05	-68.29	21.16	19.87	1.29	2073.82	2008.27	65.55	5976.21	6717.43	-741.22	60045.75	60331.74	59885.39	60284.31		
13-10-2021	3226.2	3618.35	-392.15	18.18	22.49	-4.32	2575.19	2203.22	371.97	7681.34	8113.06	-431.72	60619.91	60836.63	60452.29	60737.05	Nathan s Famou	0.7184
14-10-2021	3327.35	3558.33	-230.98	21.43	19.94	1.49	2528.78	2265.78	263	7706.79	9457.38	-1750.59	61088.82	61353.25	60978.04	61305.95	Indian shares hi	0.6808
18-10-2021	3599.31	3755.84	-156.53	22.93	32.46	-9.53	2736.57	2530.67	205.9	7720.49	9424.36	-1703.87	61817.32	61963.07	61624.65	61765.59	Indian stocks cl	0.8807
19-10-2021	4030.45	4083.35	-52.89	32.98	29.56	3.42	2975.27	2911.98	63.3	5456.9	8035.12	-2578.22	62156.48	62245.43	61594.29	61716.05	Unilever India W	0.9786
20-10-2021	3566.34	3694.76	-128.42	22.04	19.91	2.13	3148.47	2939.07	209.4	7364.58	9045.31	-1680.73	61800.07	61880.36	61109.29	61259.96	Indian stocks er	0.9393
21-10-2021	2524.82	2548.98	-24.16	20.62	14.07	6.56	2432.25	2357.81	74.45	8370.14	7941.69	428.45	61557.94	61621.2	60485.65	60923.5	Indian stocks er	0.9689
22-10-2021	2566.98	2585.93	-18.95	23.31	19.2	4.11	2299.56	2341.54	-41.98	8040.03	7010.06	1029.97	61044.54	61420.13	60551.15	60821.62	Indian shares fa	0.9316
25-10-2021	2440.77	2657.53	-216.76	26.29	19.04	7.25	2522.3	2441.92	80.38	11137.25	8747.02	2390.23	61398.75	61404.99	60449.68	60967.05	Indian shares er	0.91
26-10-2021	2058.93	2228.09	-169.16	16.41	13.75	2.66	2162.45	2120.37	42.08	8494.57	7109.16	1385.41	60997.9	61497.71	60791.29	61350.26	Indian shares ris	0.9819
27-10-2021	1977.43	2026.21	-48.78	15.11	15.09	0.02	1880.55	1888.19	-7.63	7986.46	7513.98	472.48	61499.7	61576.85	60989.39	61143.33	Wipro Appoints	0.95
28-10-2021	2209.6	2044.84	164.77	20.09	9.27	10.82	1930.18	2222.46	-292.27	8311.63	7475.03	836.6	61081	61081	59777.58	59984.7	Indian shares se	-0.7003
29-10-2021	2392.34	2446.33	-53.99	19.84	14.59	5.25	2478.55	2485.65	-7.11	9716.1	5373.59	4342.51	59857.33	60132.81	59089.37	59306.93	Indian shares ex	0.9521

Methodology

Sentimental 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.
- 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.

Methodology

Testing and Training Dataset Preparation

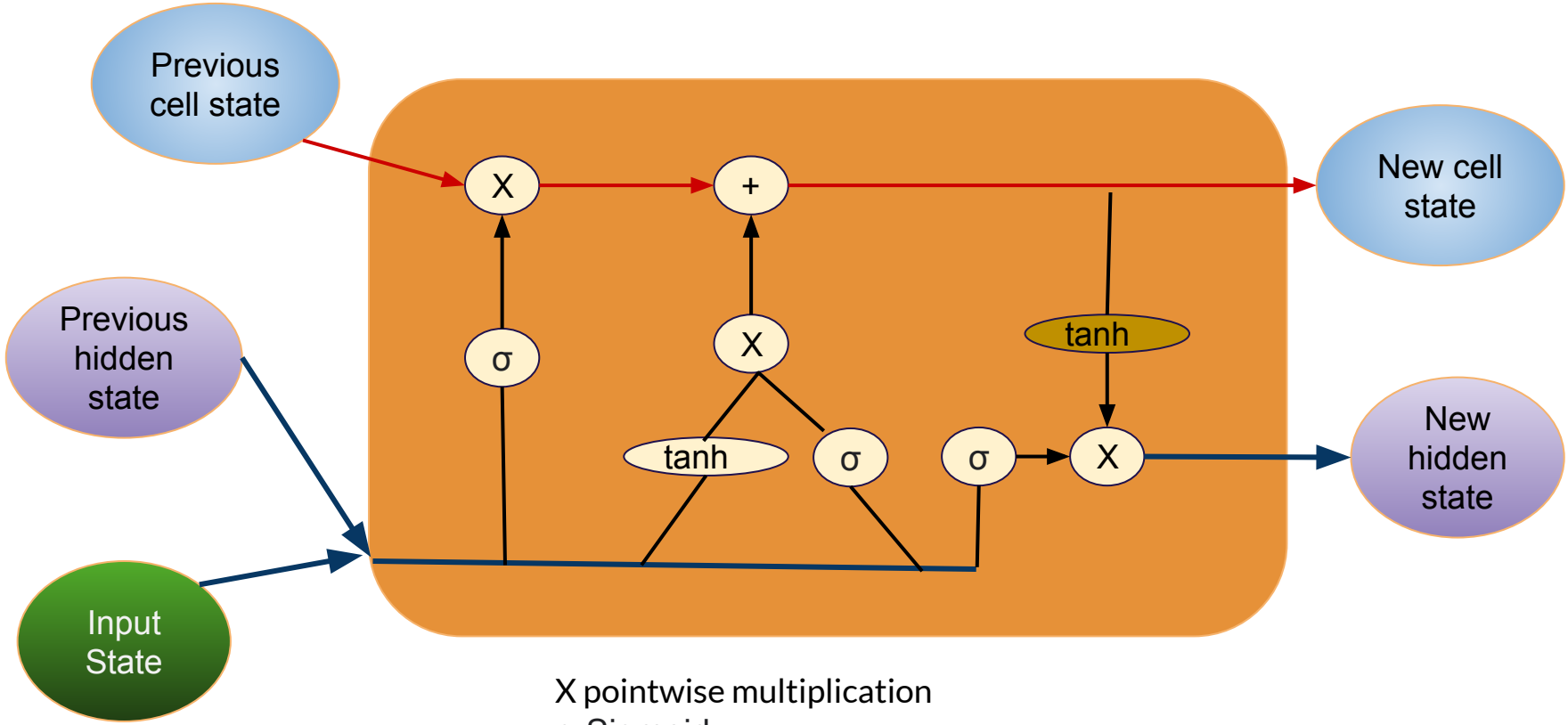
- Textual Data have been collected in various files for various dates.Several formats of dates are made into one using delimiters on excel and finally merging them.
- After the collection of transaction data and textual data. We have to make combined data set. Keeping transaction data on the left and 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.

Methodology

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 .

How Does LSTM Work?



X pointwise multiplication
σ Sigmoid

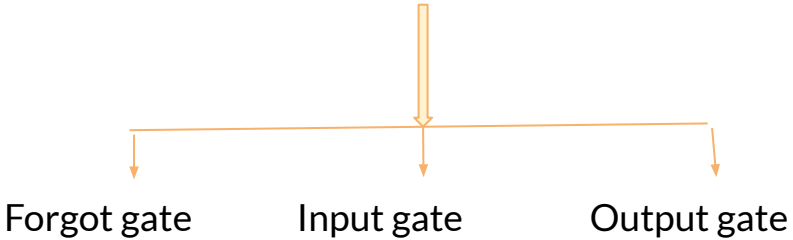
Methodology

How does LSTM work?

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

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

LSTM has three gates to filter the data.



Methodology

How does LSTM work?

STEP 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 pointwise multiplication with cell state.

STEP 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 sigmoid

Methodology

How does LSTM work?

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 pointwise multiplication to get the combined state.

STEP 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.

Methodology

How does LSTM work?

Apply sigmoid activation on the previous hidden state and input data and pointwise 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.

Methodology

Implementation of LSTM:

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 activations 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 val_loss is has a tolerable difference for more than 15 times

04
Performance and Evaluation

Performance and Evaluation

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

1. 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.
2. 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.

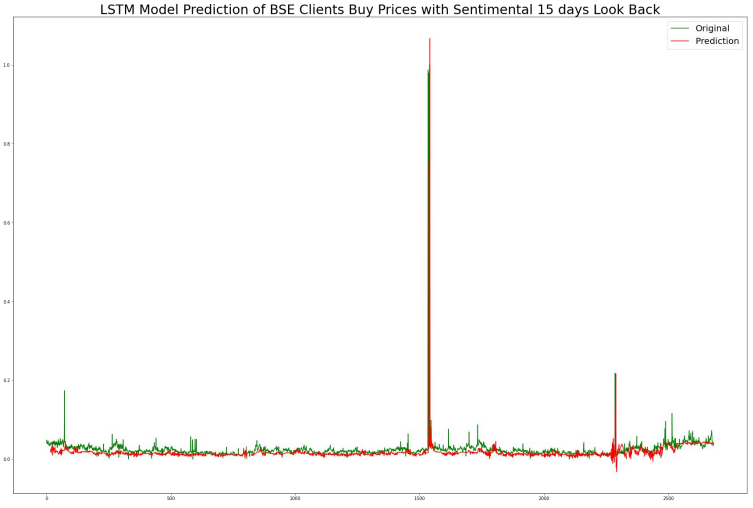
Performance and Evaluation

- The MAE and the RMSE can be used together to diagnose the variation in the errors in a set of forecasts.
- The RMSE will always be larger or equal to the MAE.
- If the RMSE=MAE, then all the errors are of the same magnitude
- The greater difference between them, the greater the variance in the individual errors in the sample.
- When RMSE>MAE ,the greater the variance in the individual errors in the samples

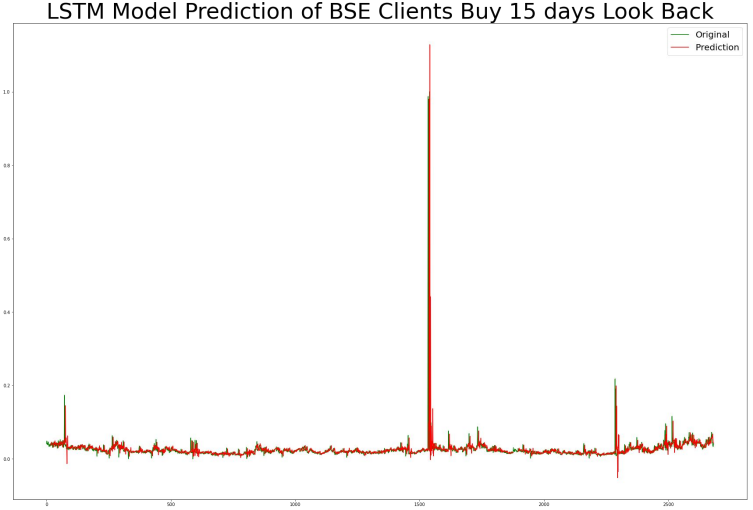
05
Results

Results Client Buy:

Forecasting of Client Buy Results - 15 days look up



With Sentimental
MAE = 0.008974243
RMSE = 0.0129323425

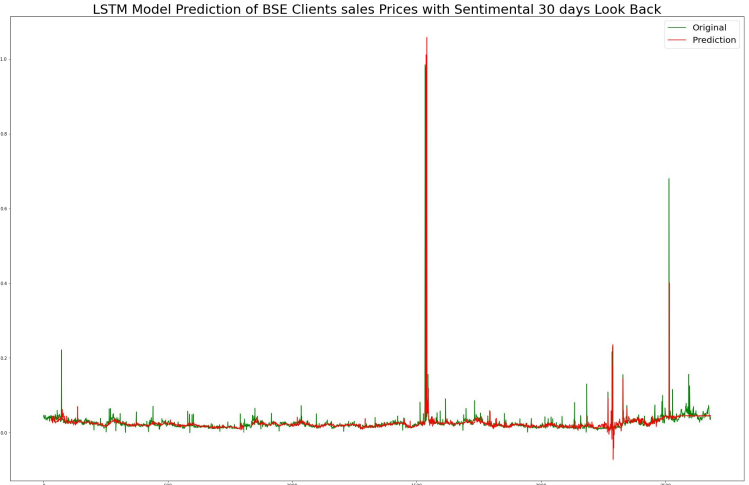


Without Sentimental
MAE = 0.0065712854
RMSE = 0.01552742

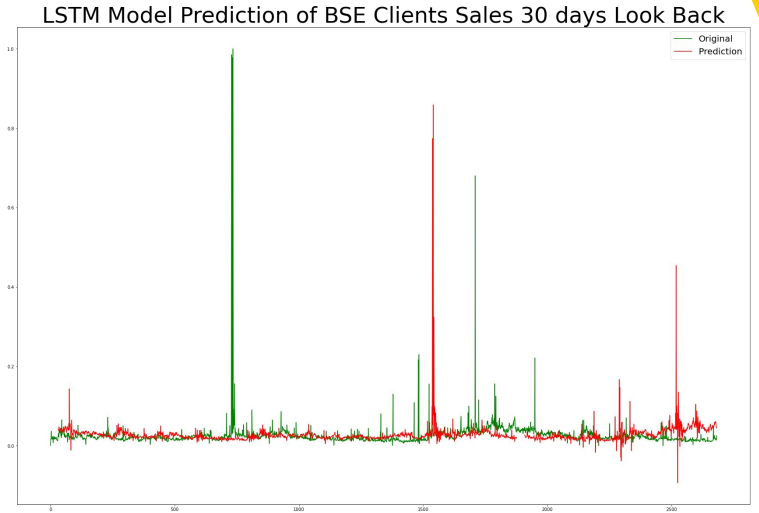
26.77 percent reduction in MAE value, 16.71 percent reduction in RMSE value when we included sentiment analysis at 15 days look back.

Results Client Sales:

Forecasting of Client Sales Results - 30 days look up



With Sentimental
MAE = 0.0048487964
RMSE = 0.005966673

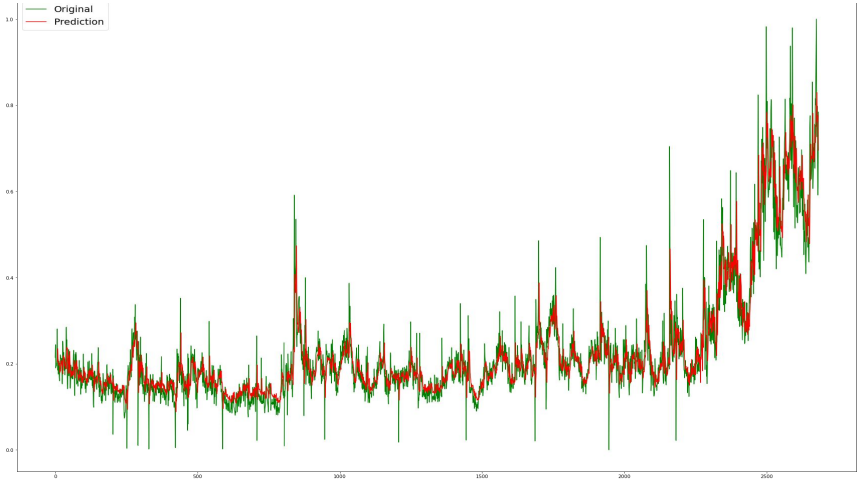


Without Sentimental
MAE = 0.011441373
RMSE = 0.03433521

57.62 percent reduction in MAE value, 57.62 percent reduction in RMSE value when we included sentiment analysis at 30 days look back..

Results Proprietary Buy:

Forecasting of Proprietary Buy Results - 30 days look up



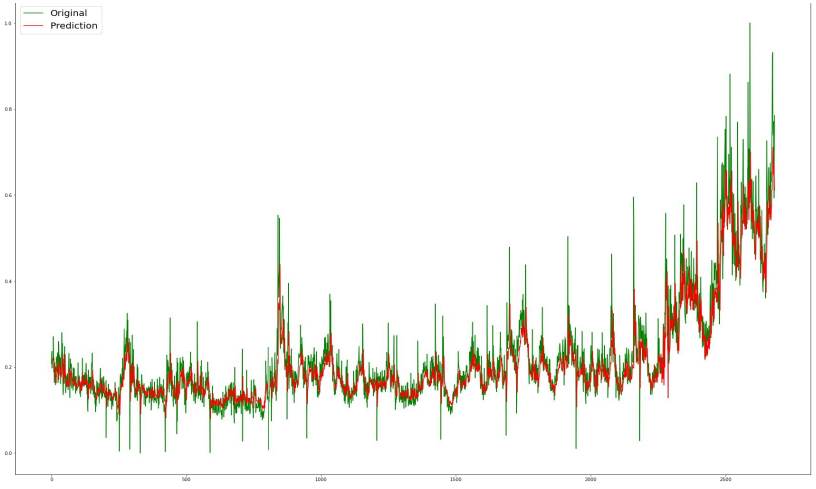
With Sentimental
MAE = 0.023039889
RMSE = 0.031927396

Without Sentimental
MAE = 0.052677557
RMSE = 0.07954283

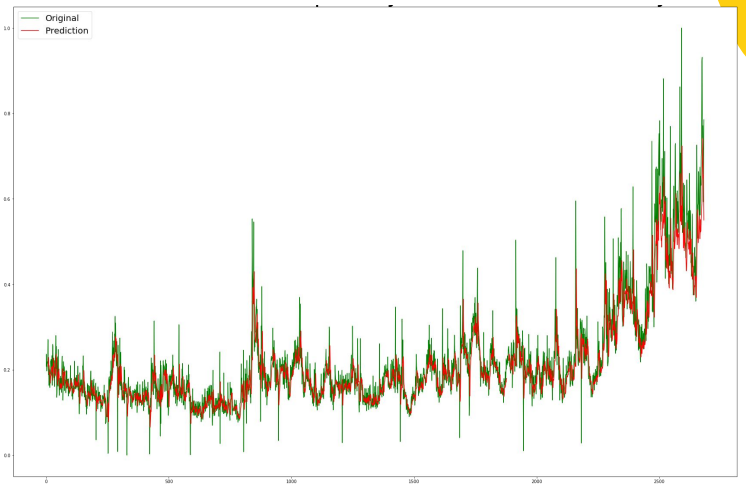
We can certainly observe 56.26 percent, 59.86 percent reduction in MAE, RMSE values when evaluated using sentiment analysis at 30 days look up.

Results Proprietary Sales:

Forecasting of Proprietary Sales Results - 30 days look up



With Sentimental
MAE = 0.02190594
RMSE = 0.030971821

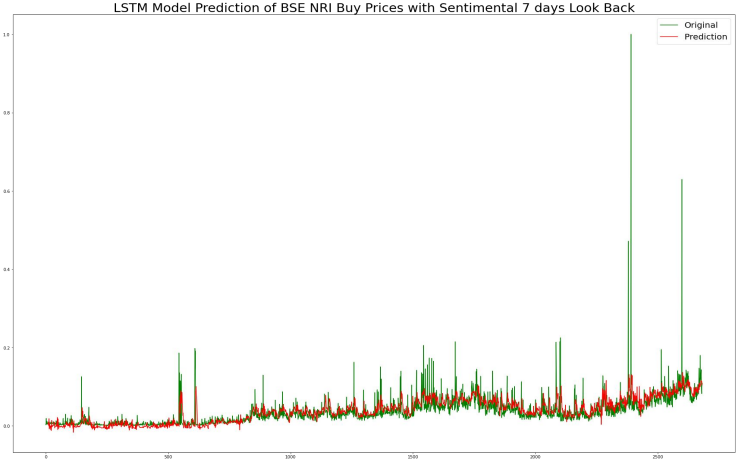


Without Sentimental
MAE = 0.04622454
RMSE = 0.07039143

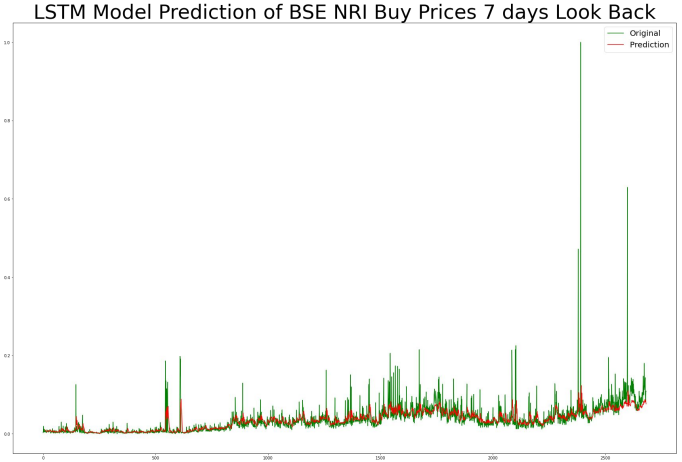
We can certainly observe 52.61 percent, 56 percent reduction in MAE, RMSE values when evaluated using sentiment analysis at 30 days look up.

Results NRI Buy:

i)Forecasting of NRI Buy Results - 7 days Lookup



With Sentimental
MAE = 0.02656468
RMSE = 0.034758694

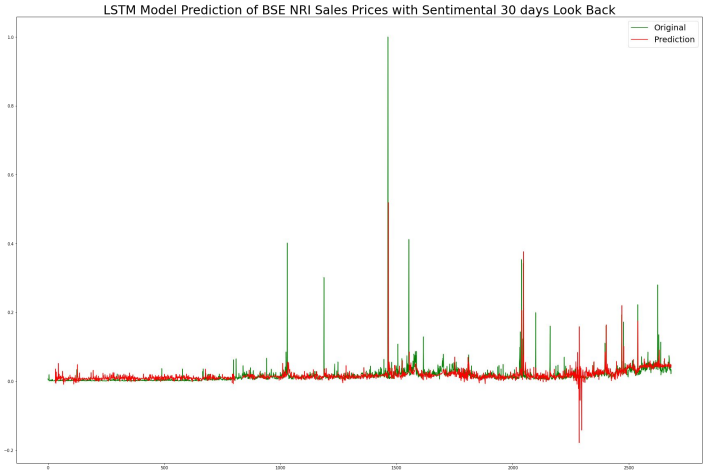


Without Sentimental
MAE = 0.06506311
RMSE = 0.097840875

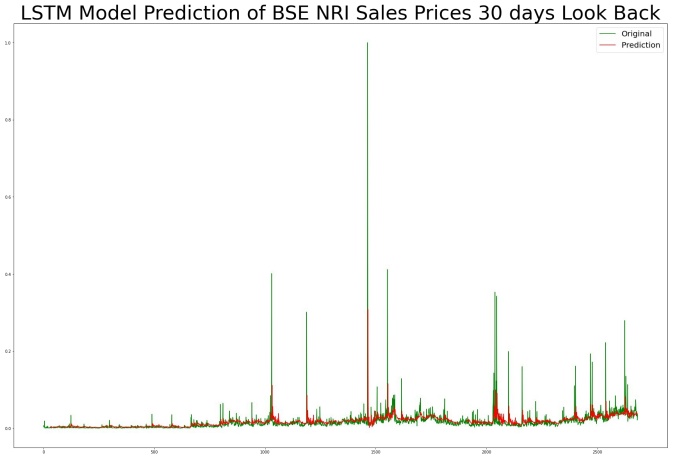
64.47 percent reduction in RMSE value also 59.17 percent reduction in MAE value by introducing Sentiment analysis with look back as 7 days.

Results NRI Sales:

i)Forecasting of NRI Sales Results - 30 days Lookup



With Sentimental
MAE = 0.010306023
RMSE = 0.008943314

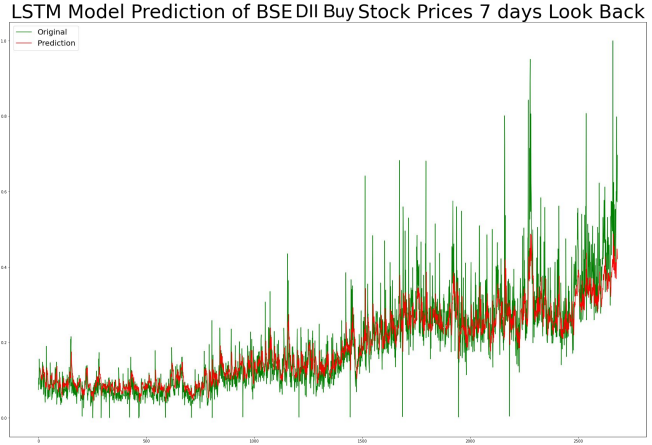


Without Sentimental
MAE = 0.011090927
RMSE = 0.028352953

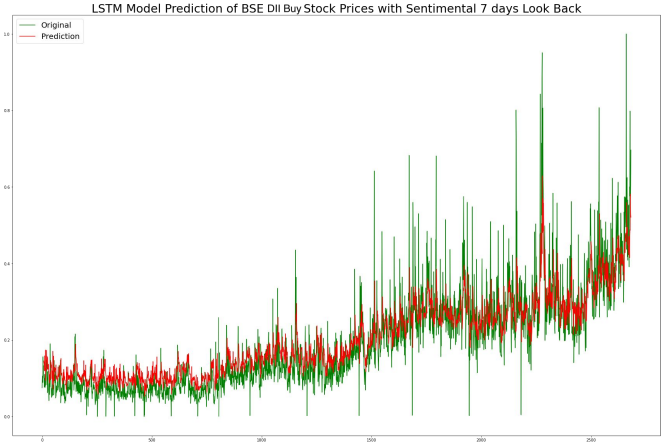
We can certainly observe 46.97 percent, 68.45 percent reduction in MAE, RMSE values when evaluated using sentiment analysis.

Results DII Buy:

Forecasting of DII Buy Results - 7 days Lookup



With Sentimental
MAE = 0.0647
RMSE = 0.0956

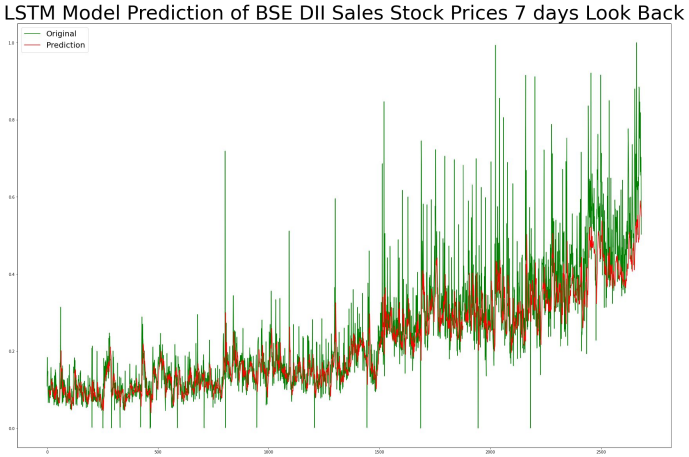


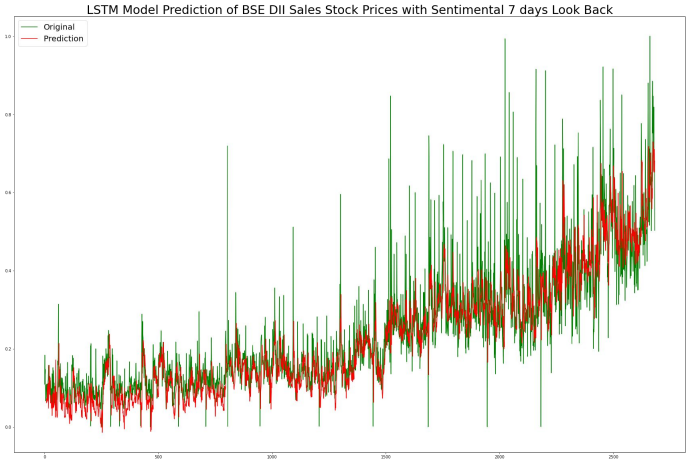
Without Sentimental
MAE = 0.0331
RMSE = 0.0391

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.

Results DII Sales:

Forecasting of DII Sales Results - 7 days Lookup





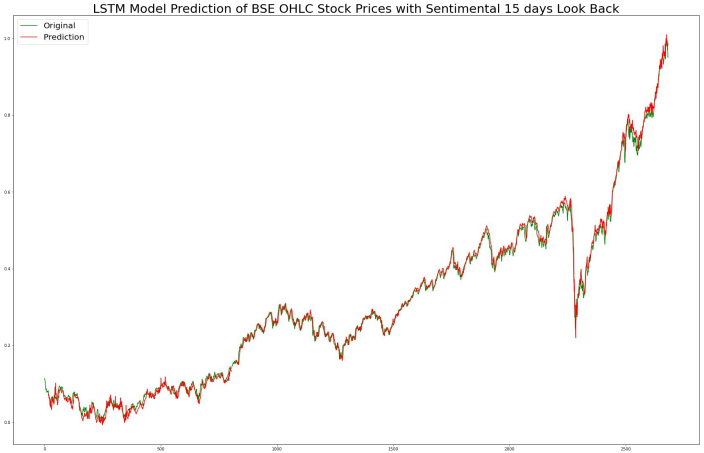
Without Sentimental
MAE = 0.08789
RMSE = 0.12604

With Sentimental
MAE = 0.03658
RMSE = 0.04727

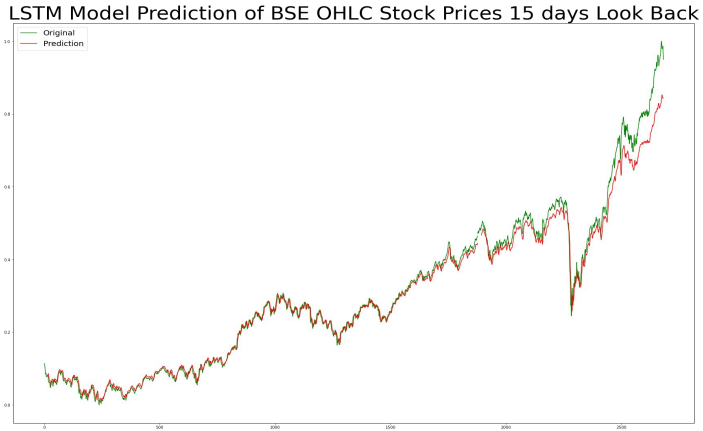
Here is a change of 80 percent in RMSE error also 50 percent in MAE error with the introduction of sentimental analysis with having the look back as 7 days

Results OHLC:

Forecasting of OHLC Results - 15 days Lookup



With Sentimental
MAE = 0.007780917
RMSE = 0.010715311



Without Sentimental
MAE = 0.03644404
RMSE = 0.04876546

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.

06
Conclusion

Conclusion

1. For Client's Buy, the better performing scenario is the one with look back days as 15 and with sentimental analysis and with a higher accuracy.
   Overall Average change of 31.49% for MAE and 33.20 % change for RMSE

2. For Client's Sales, 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.
   Overall Average change of 52.29% for MAE and 62.73% change for RMSE

3. For Proprietary Buy, 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.
   Overall Average change of 56.05% for MAE and 58.34 % change for RMSE

4. For Proprietary Sales, 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.
   Overall Average change of 52.92% for MAE and 56.87 % change for RMSE

5. For NRI Buy, 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.
Overall Average change of 14.13 % for MAE and 59.45 % change for RMSE

6. For NRI Sales, 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.
Overall Average change of 37.90 % for MAE and 68.90% change for RMSE

7. For DII Buy, we can say that with the look back as 7 days with sentimental analysis has given the best accuracy with the least error.
Overall Average change of 40.61 % for MAE and 29.67 % change for RMSE

8. For DII Sales, we can say that with the look back as 7 days with sentimental analysis has given the best accuracy with the least error.
Overall Average change of 55.61 % for MAE and 33.20 % change for RMSE

9. For OHLC, after observing the three scenarios that having look back days as 15 along with sentiment analysis has acquired smaller values of MAE.
Overall Average change of 53.76 % for MAE and 55.98 % change for RMSE

Overall improvement: 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**.

This model creation is cretual as this is not only focused on the stock price but more on how clients turnover. Which is important for businesses and companies to analyze there growth and look forward for future investments.

Our model offers the capability to let the businesses to pre plan for any thing in the future

Limitations : Financial and Digital news publishing in India is minute in order to collect each and every news and post for execution of sentiment analysis, which may compromise the results to what is expected.

Future Proposal

1. We can automate the entire process of transactional data extraction and social media/news headlines from online, followed by preprocessing the data and modeling..
2. Selecting more specific stock markets with major social media presence. To make use of our model to a better extent.
3. Creating a User Interface for ease of access to all users interested.

07 References

References

1. A Survey on Stock Market Prediction Using Machine Learning Techniques Polamuri Subba Rao K. SrinivasA. Krishna Mohan
2. Stock Market Prediction Using Machine Learning , Ishita parmar,Navanshu agarwal,Lokesh Chauhan
3. Stock Price Prediction Using News Sentiment Analysis Saloni Mohan1, Sahitya Mullapudi1, Sudheer Sammeta1, Parag Vijayvergia1 and David C. Anastasiu1,*
4. R. Gupta and M. Chen, "Sentiment Analysis for Stock Price Prediction," 2016 IEEE Conference on MIPR
5. Transformers for Time Series Forecasting , Natasha Killenbrunn
6. Sentiment Analysis for Stock Price Prediction Rubi Gupta, Min Chen
7. Usmani S, Shamsi JA. News sensitive stock market prediction: literature review and suggestions. PeerJ. Computer Science.2015 ;7:e490. DOI: 10.7717/peerj-cs.490. PMID: 34013029; PMCID: PMC8114814.

THANK YOU :)