

Stock Market Forecasting

¹Chirag, ²Deepraj Phunyal, ³Biswabrata Mazumdar, ⁴Deepanshu Kumar, ⁵Archana B

⁵Professor, ^{1,2,3,4}School of C & IT, REVA University, Bangalore, India

⁵Assistant Professor, School of C & IT, REVA University, Bangalore, India

¹chirraaag@gmail.com, ²deepjoishi@gmail.com, ³biswabrata0007@gmail.com,

⁴deepanshusingh025@gmail.com, ⁵archanab@reva.edu.in

Article Info

Volume 83

Page Number: 4391-4398

Publication Issue:

May - June 2020

Abstract

Forecasting stock market prices has been a topic of interest lately, especially, among the researchers and the analysts. But forecasting/predicting stock prices does not come in handy as it is a pretty complex task. The stock market is a highly volatile concept as it is affected by a number of factors viz. Previous performance of the stocks, political factors, investor's sentiment, change in leadership, etc. It has been observed that historical data and previous performances of the stocks have been inefficient in forecasting the accurate nature of stock.

Existing studies focusing in or around stock market forecasting, focuses on only one aspect/parameter i.e. Historical Data/Previous Performance. There are many algorithms but, even though the accuracy rate is high for some algorithms, one cannot be completely certain that he/she can invest based on only one behaviour of the stock market. Using the sentiment analysis on the tweets collected using the Twitter API and the closing value of various stocks, one can build a handy system that can forecast the stock price movement of various companies various days prior. This paper is based on the above mentioned methodologies and generates a decent accuracy rate which can be further worked upon for more improved results.

Keywords: forecasting stock prices, historical data, sentiment analysis, twitter API

Article History

Article Received: 19 November 2019

Revised: 27 January 2020

Accepted: 24 February 2020

Publication: 12 May 2020

1. Introduction

In today's world, social media has become a platform that reflects people's thoughts and opinions to any particular event or news. Financial news and day to day impact can surely affect anything around this world. In concern with the stock market, any positive or negative sentiment of public related to a particular company can have a ripple effect on its stock prices. Thus, adding social media sentiments to your machine learning model for stock market forecasting happens to be a great motivator.

This paper discusses the methodology to forecast the stock market prices of various companies using sentiment analysis of social media data such as tweets related to that particular company. Twitter API has been used to access the tweets, monitor them and filter them to positive and/or negative. This can help an investor to make appropriate investment decisions based on the results obtained from the forecasting model. The main issue in the data

processing would be to go through all the tweets and retrieve/filter those tweets that could affect the stock behaviour of the firm taken into consideration.

In this paper, the first task is the collection of tweets. The tweets are collected and then sentiment analysis is performed upon it. In the mean time, with the help of a suitable machine learning algorithm, historical/previous data is analysed. This can easily justify a valid correlation between the stock values and the sentiments analysed. Finally, with this analysed data, the forecasting model will be trained to make stock market predictions. As the social media has grown immensely, the public reactions to any event/activity are almost expressly available on various social media platform. Catching these moods quickly and then estimating the stock prices can give a real time forecasting model similar to some of the other real time forecasting models like weather forecasting models, heart rate monitoring and predicting total sales in

a store models etc.

2. Relevance of the Project

This paper is pretty relevant as it escorts people who possess limited knowledge of finance and investments into making acquainted decisions regarding stock market investments. It totally bids a farewell to the investment experts who command sky-high wages to guide ones financial decisions by providing a straightforward solution which can be examined easily by anyone with an access to a computer or a laptop along with an internet connection. Publicizing this machine learning approach clubbed with the Sentiment Analysis, provides a very cheap or perhaps a free alternative to several stock market investment counselling organisations which are popular. The paper puts in a small effort towards assisting the inexperienced investors/capitalists and prevent them from suffering financial loss.

3. Literature Report

Machine Learning Algorithms

In this paper, many algorithms used for the prediction purpose were studied and analysed. Reference [1] used Support Vector Machines (SVM) along with Empirical Mode Decomposition (EMD) which focused on dealing with enormous datasets and yields higher prediction performance and faster convergence speed.

Reference [2] has taken into consideration Random Forest using Least Square (LS) Boost according to which one single regression model is not sufficient and the main focus has been kept on reducing the error estimates and a decent efficiency and accuracy has been displayed by their model.

Reference [3] consists of works on non-traditional methods which comprises of Hidden Markov Mode (HMM) along with Neural Networks and SVM. Better accuracy than some of the traditional techniques has been achieved but lacks in performance with increase in the data.

Reference [4] is a survey on the stock market prediction approaches using machine learning and the main focus has been kept on Regression algorithm and its various types.

Reference [7] takes into account the deep learning approach of the Artificial Neural Network (ANN) and the focus has been kept on improving the accuracy of the model.

Reference [8] talks about unsupervised learning along with another form of ANN called Jordan Recursive Neural Network (JRNN). High dimensionality of data is the sub problem focused in [8]. Success has been achieved in creating Indicators without the human intervention and have also shown improvements in the Mean Absolute Percentage Error (MAPE).

Reference [9] dives into Long Short-Term Memory (LSTM), optimized by Mini-Batch Gradient

Descent (MBGD) which takes into account the extreme maxima and minima for future price prediction and has achieved smaller error rate.

Sentiment Analysis

According to [5] any positive or negative sentiment of public related to a company can have a ripple effect on the stock prices and therefore has made use of sentiment analysis along with the closing and have shown that handling large datasets with sentiments.

Reference [6] has shown the work upon the social media mining along with machine learning algorithms. [5] states that stock market follows a random pattern work and therefore value forecasting cannot rely on a single factor. Also, the focus has been restricted to short term investors.

Reference [10] tends to create a dictionary based on news sentiment model but domain specific sentiments were difficult to be scaled. Also, a number of assumptions were made by the researchers in this model.

Reference [11] has tried to improve the accuracy of stock price prediction by gathering a large amount of time series data and furthermore fake news filtering has been done by opting for only financial news articles. Unlike many other references like [5] and [6], this reference has used Financial News Articles over Twitter Sentiment Analysis. Facebook Prophet has been used as the platform to make predictions with good accuracy.

4. Methodology

The methodology for the above mentioned model has been summed up in the following modules:

Data Fetching

In regards with the collection of tweets, the Twitter API allows you to access the features of Twitter without having to go through the website interface. Twitter is an information network and communication mechanism that produces more than 200 million tweets a day. The Twitter platform offers access to that corpus of data, via APIs. The Twitter API also provides an API key and an API Secret key for authentication purpose.

For the stock history data, one can go again for financial APIs provide by Yahoo or Google or any other domain, and can fetch the data of certain firms online. But in this paper, the focus has been kept at forecasting the stock prices of only one firm, so the dataset containing 5 years of data history was downloaded from kaggle, an online domain the provides free datasets. The firm taken into consideration is TATAGLOBAL. The data worked upon ranges from the 8th of August 2013 to the 8th of August 2018. Although, there were some dates with missing data, so for the time being, they have been omitted.

The challenging aspect to work and research in this field is making use of the available data to get accurate decisions/results. In this paper, a lot of data is being dealt

with and also a lot of data is generated. So, if these data were to be organised physically, it would become almost paradoxical to arrive at a result. Thus, making use of machine learning models(here, Linear Regression) to process the data as and when it is being generated.

Data Processing

This sub-section aims at providing the details about the pre-processing steps followed in the paper. Since large data sets can have various kinds of noises in it, it was important to look through the dataset. It was found that the dataset was filled with null values. Now, there were two options, either to find those missing values or to just ignore the tuples. The latter method was considered over the former one. So, the dataset was then directly fed into the machine learning model. Also, the data set contains many attributes, but for the model, only the attribute 'Close' was focused upon which gives the closing value of each day.

For going through the retrieved tweets, Search API named REST API has been used which helps us request specific query of the recent tweets. So, after authenticating using the keys, the tweets were accessible using the python library Tweepy. Now, the tweets contained so many irrelevant texts like links, special characters, tags and many other symbols which did not lead to any sentiment. Therefore, it was necessary to remove these irrelevant texts so as to avoid any errors. And, to do so, a function named cleanTweet had been used which completely get rid of these irrelevant texts.

The adjoining figure gives a brief overview of the Sentiment Analysis from the tweets collected from Twitter:

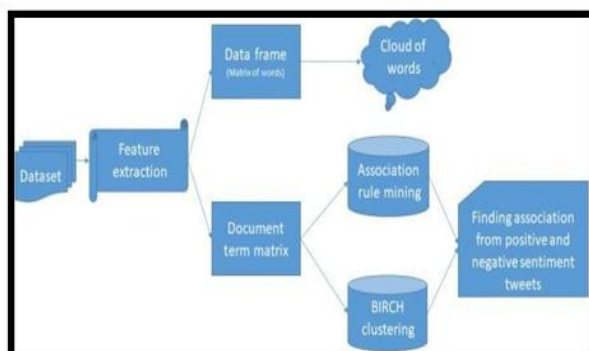


Figure 1: Overview of Sentiment Analysis

Evaluation Method

Speaking about the linear regression algorithm used, it is basically used to estimate relationships between variables. To be more specific, this algorithm tells us how the typical value of the dependent variable changes when any one of the independent variables is varied. For the forecasting module, there is a scalar dependent variable, termed as y and a dependent variable, denoted by X . Since, there is only one independent variable in the module, it comes under Simple Linear Regression. Linear

Regression has been used to fit a predictive model to an observed dataset of y and X values. After training and testing this model, if any additional value of X is given without its corresponding value of y , then this trained and tested model can easily forecast the value of y . So, in this model, X is a array consisting of the attribute 'Close' extracted from the dataset, and y is an array which will hold the forecasted values.

For the Sentiment Analysis, the count for the number of tweets to be analysed has been summed up to 1000, and only those tweets have been taken into consideration containing the keyword "TATA". It then analyses these 1000 tweets and divides them into seven categories viz. Positive, Weakly Positive, Strongly Positive, Neutral, Negative, Weakly Negative and Strongly Negative. Based on this analysis, the Sentiment Analysis is then plotted in the form of a pie chart and assigned a Sentiment Score. Based on the final Sentiment Score(ranges from -1 to 1), the final forecasting is made that is based on both the Linear Regression algorithm and Sentiment Analysis.

The Methodology section can be summed up with the following data flow diagram:

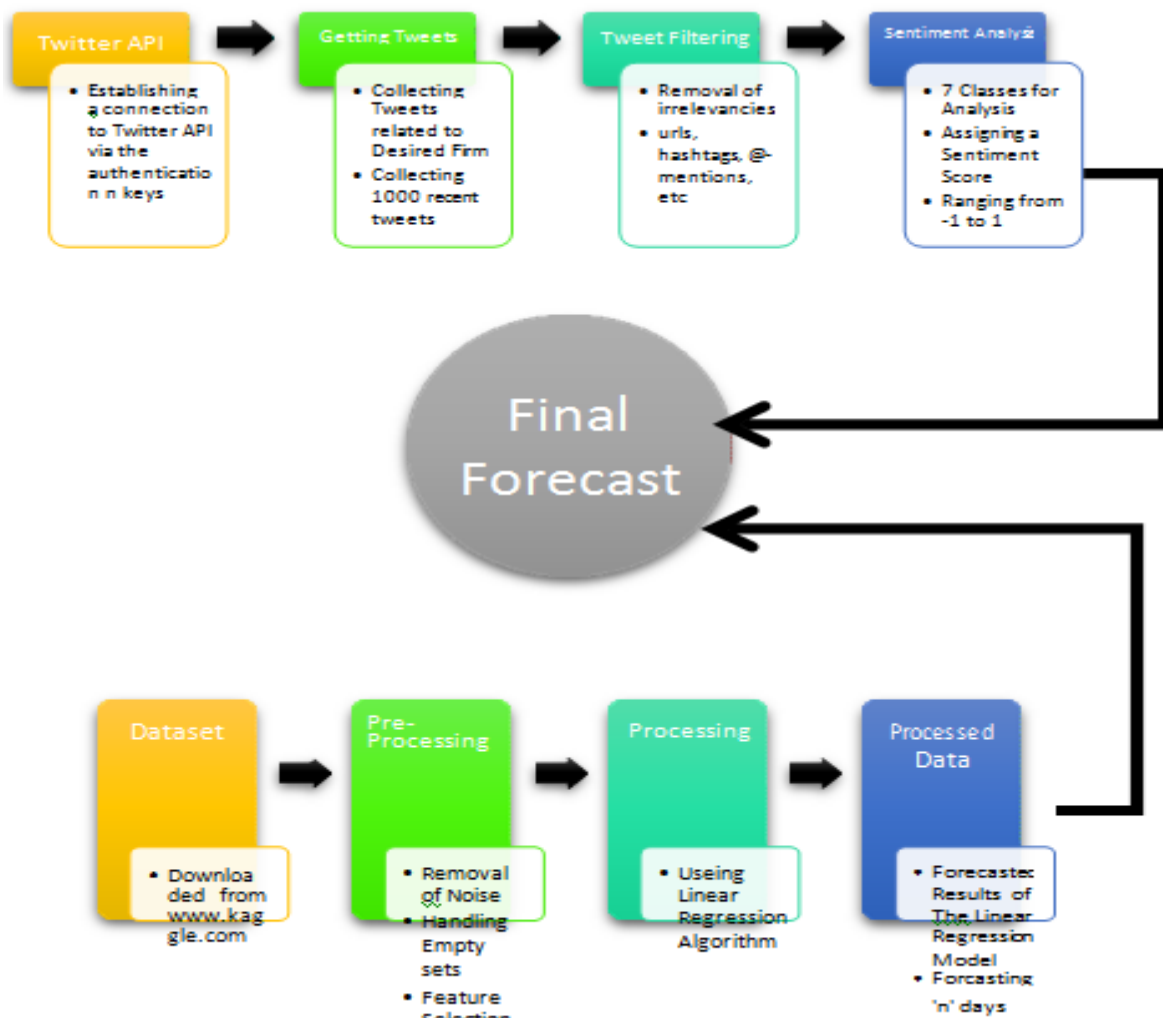


Figure 2: Data Flow Diagram

5. Experimental Results

'n' days value of X is given without its corresponding value of y, then this trained and tested model can easily forecast the value of y. So, in this model, X is a array consisting of the attribute 'Close' extracted from the dataset, and y is an array which will hold the forecasted values.

For the Sentiment Analysis, the count for the number of tweets to be analysed has been summed up to 1000, and only those tweets have been taken into consideration containing the keyword "TATA". It then analyses these 1000 tweets and divides them into seven categories viz. Positive, Weakly Positive, Strongly Positive, Neutral, Negative, Weakly Negative and Strongly Negative. Based on this analysis, the Sentiment Analysis is then plotted in the form of a pie chart and assigned a Sentiment Score. Based on the final Sentiment Score(ranges from -1 to 1), the final forecasting is made that is based on both the Linear Regression algorithm and Sentiment Analysis.

The Methodology section can be summed up with the following data flow diagram:

In this paper, the datasets were divided into two parts. The first part comprises of the 80% of the dataset which was used for the training the model, and the second part comprises of the remaining 20% of the dataset, used for the testing purpose. The model has been trained and then used to show the behaviour of the company's stock prices for the next 30 days. First, the forecasting is done without the use of the Sentiment Analysis. The accuracy achieved on the 5 years of data of TATAGLOBAL using the Linear regression is quiet remarkable. The accuracy rarely dropped below 85%. After this, Sentiment Analysis was performed on 1000 tweets related to the the TATAGLOBAL and framed into seven categories. It was found that the sentiment for the firm was Weakly Positive. And then using this Sentiment Score, the final forecasting result was obtained.

This model is ahead of the models spoken about in

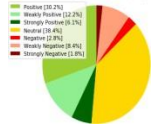
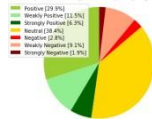
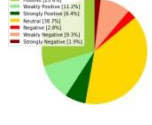
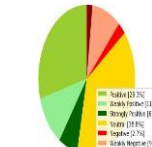
the Machine Learning Algorithms module of the Literature Survey section of this paper. The models in [1] and [2] showed very good accuracy rates, in fact better than that of the model which this paper focuses on. But, since they([1] and [2]) focus on only one aspect i.e. historical data, it can be termed as a downgraded model when compared to this model because this model takes into account, two aspects that play a major role on the stock market

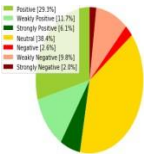
i.e. the historical data along with the sentiment

analysis. The models in [7] and [8] also give remarkable results but once again lack in providing an accurate real time forecasting as only a single parameter cannot decide the precise course of a forecasting model.

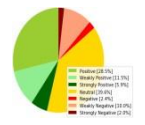
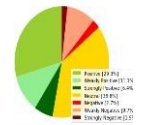
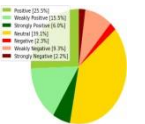
Below mentioned are two tables which evidently depict the outputs achieved by the model discussed in the paper. The two tables correspond to two different executions of the model. Both the runs/executions have forecastings in 5 slots viz. 1 day, 5 days, 10 days, 15 days and 30 days.

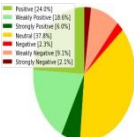
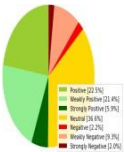
First Run:

No. of Days	Linear Regression(LR) Accuracy (Percentage)	Forecasted Output(Only LR)(Day wise)	Sentiment Analysis(Pie Chart)	Final Forecasted Output (Clubbed Result)
1	99.55	[155.31]		[155.63]
5	97.08	[158.78,159.37, 159.46,154.95, 155.19]		[159.09,159.68, 159.78,155.24, 155.49]
10	95.74	[160.85,158.67, 163.12,160.99, 157.18,158.53, 159.11,159.21, 154.76,155.00]		[161.14,158.94, 163.42,161.28, 157.44,158.80, 159.38,159.48, 155.00,155.24]
15	92.65	[161.06,161.87, 162.44,163.95, 157.60,160.54, 158.41,162.77, 160.68,156.94, 158.27,158.83, 158.93,154.57, 154.80]		[161.46,162.29, 162.87,164.43, 157.90,160.92, 158.73,163.22, 161.07,157.22, 158.58,159.17, 159.27,154.78, 155.03]
		[145.19,144.46, 147.33,143.91, 144.23,153.59,		[145.10,144.38, 147.24,143.83, 144.15,153.47,

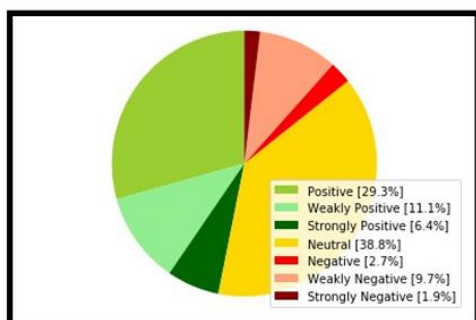
30	86.39	155.41,158.65, 158.88,161.80, 165.50,167.23, 165.59,162.21,		155.29,158.52, 158.75,161.66, 165.35,167.07, 165.44,162.07,
		163.13,160.75,		162.98,160.61,
		161.53,162.08,		161.39,161.93,
		163.54,157.42,		163.39,157.29,
		160.25,158.20,		160.11,158.07,
		162.40,160.39,		162.25,160.25,
		156.78,158.06,		156.66,157.93,
		158.61,158.70,		158.48,158.57,
		154.50,154.73]		154.38,154.61]

Second Run:

No. of Days	Linear Regression(LR) Accuracy (Percentage)	Forecasted Output(Only LR)(Day wise)	Sentiment Analysis(Pie Chart)	Final Forecasted Output (Clubbed Result)
1	99.28	[155.29]		[155.40]
5	97.80	[158.74,159.32, 159.42,154.91, 155.15]		[158.76,159.35, 159.45,154.91, 155.16]
10	94.91	[161.01,158.83, 163.28,161.15, 157.33,158.68, 159.27,159.36, 154.91,155.15]		[161.18,158.99, 163.46,161.32, 157.48,158.84, 159.43,159.52, 155.05,155.29]

15	92.89	[161.46,162.27, 162.84,164.37, 157.97,160.93, 158.78,163.17, 161.07,157.30, 158.64,159.21, 159.31,154.92,		[161.27,162.07, 162.64,164.16, 157.81,160.75, 158.67,162.97, 160.89,157.15, 158.47,159.04, 159.14,154.78,
		155.2]		155.02]
		[145.30,144.56,		[144.85,144.12,
		147.46,144.00,		146.99,143.57,
		144.33,153.78,		143.89,153.24,
		155.62,158.90,		155.06,158.30,
		159.13,162.08,		158.53,161.45,
		165.82,167.57,		165.14,166.87,
30	84.19	165.91,162.50,		165.23,161.86,
		163.42,161.02,		162.77,160.40,
		161.80,162.36,		161.17,161.72,
		163.83,157.65,		163.18,157.07,
		160.51,158.44,		159.90,157.84
		162.68,160.65,		162.04,160.03,
		157.01,158.30,		156.43,157.71,
		158.85,158.95,		158.26,158.37,
		154.70,154.93]		154.15,154.38]

The pie chart may not be visible in the tables above, so the following pie graph gives a brief of all the attributes and colours associated with it.



6. Challenges

- ★ Finding and getting rid of the unwanted items in the datasets.
- ★ The linear regression model is sensitive to outliers.
- ★ Obtaining the tweets between a specified period of time.
- ★ Filtering of false tweets.
- ★ Authentication is required every time to access real time Twitter Data which consumes time.
- ★ Difficult for the model to recognise things like sarcasm and irony, negations, jokes and exaggerations.

7. Conclusion

The aim of this paper is to help the brokers and the stock investors for investing money in the stock market. It was a pretty good outcome by using the Linear Regression Algorithm and the Sentiment Analysis together and a feasible model has been developed for forecasting the behaviour of stock market. Where, on one hand, the Linear Regression algorithm is used to make the base of the forecasting model while, on the other hand, using tweets for the Sentiment Analysis gives a good overview of the public mood. To sum up, the paper shows that good results have been achieved with this model and there is a huge scope on further researching on the same to enhance the obtained results.

References

- [1] Honghai Yu and Haifei Liu on “Improved Stock Market Prediction by Combining Support Vector Machine and Emperical Mode Decomposition”, 2012 Fifth International Symposium on Computational Intelligence and Design.
- [2] Nonita Sharma and Akanksha Juneja on “Combining of Random Forest Estimates using LSboost for Stock Market Index Prediction” 2012 Second International Conference for Convergence in Technology.
- [3] Poonam Somani, Shreyas Talele and Suraj Sawant on “Stock Market Prediction using Hidden Markov Model” in 2014 IEEE Conference.
- [4] Ashish Sharma, Dinesh Bhuriya and Upendra Singh on “Survey of Stock Market Prediction using Machine Learning Approach” 2017.
- [5] Tejas Mankar, Tushar Hotchandani and Mahesh Madhwani on “Stock Market Prediction based on Social Media Sentiments using Machine Learning” 2018.
- [6] Yaojun Wang and Yaoqing Wang on “Using Social Media Mining Trchnology to assist in Price Prediction of Stock Market”.
- [7] Aditya Menon, Shivali Singh and Hardik Parekh on “A Review of Stock Market Prediction using Neural Networks” in 2019.
- [8] Nima Gozalpour and Mohammad Teshnehlab on “Forecasting Stock Market Price using Deep Neural Network” in 2019.
- [9] Dou Wei on “Prediction of Stock Prices based on LSTM Neural Network” in 2019.
- [10] Dev Shah, Haruna Isah and Farhana Zulkernine on “Predicting the Effects of News Sentiments on the Stock Market” in 2018.
- [11] Saloni Mohan, Sahitya Mullapudi, Sudhir Sammeta, Parag Vijayvargia and David C. Anastasiu on “Stock Price Prediction using News Sentiment Analysis” in 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications.