International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Stock Market Prediction Using Twitter Sentiment Analysis

Bhavya Antiya[1], Hiloni Bhimani[2], Amruta Sankhe[3]

[1,2]Department of Information Technology, Atharva College of Engineering, Maharashtra, India
[3]Professor, Department of Information Technology, Atharva College of Engineering, Maharashtra, India

---

Abstract - For a long time, economists and analysts have been interested in estimating stock market values. Stock markets are difficult to estimate due to their high volatility, which is influenced by a variety of political and economic influences, changes in politics, market sentiment, and a variety of other factors. Predicting asset markets solely on historical evidence or textual records has proved inadequate.

Machine learning algorithms have been used to design new methods for developing simulation models that can forecast stock markets and tell whether they will rise or fall. Various sentiment analysis studies were carried out using algorithms such as support vector machines, Naive Bayes regression, etc. on various stages. The precision of machine learning algorithms depends on the amount of training data offered.

In our paper, we try to maximize the prediction of stocks by collecting and reviewing data with the help of Twitter API using Random Forest Regressor.

Key Words: Stock Market Prediction, Twitter API, Machine Learning, Random Forest Regressor.

1. INTRODUCTION

One of the most dynamic, advanced means of doing business is the stock market, commonly known as the stock exchange. It's a complex model for small companies, investors and the banking sector to all generate revenue and minimize risks [2]. This paper would however attempt to use open-source datasets and current data to predict future exchange rates using a machine-learning algorithm.
In the course of years, the share market has been an important part of the growth of many companies as well as of a country's GDP [3]. Given the competitive financial market, there are still some losses that are likely to not favor the investments made on the market. In the financial markets of the global private sector, stock markets have been given the most important position in economic liberalization [1]. There have been a number of variables affecting stock markets, the most important of which are historical records.

Many approaches for forecasting stock related data were developed using different techniques and models, which used traditional prices, past revenue growth and dividends, so we know that we need data along with one of the above factors to effectively predict stocks, so that the effective market hypothesis can be built.

In this paper, the Twitter Application Programming Interface (Twitter API), which offers a streaming API, has been taken into account in the study of financial data and continually returns the data. Each data collected reflects the user's status or attitude with respect to a specific subject. This is available through a basic HTTP authentication and a twitter account [8]. After all data is collected for every tick, an interpretation is initiated of the feelings relevant to each tweet and then a mood is predicted which has a direct impact on the stock status. Sentiment analysis is basically a problem of classification in which the data content is categorized with a positive or negative opinion [1]. Various models are developed based on various learning algorithms used for the training results. The streaming data is collected through the streaming API after such a model is prepared.

2. METHODOLOGY

2.1 Tweet Extraction

The first process is to extract the tweets from twitter. This takes place after setting up the consumer key and access token. After the tweets are fetched from twitter, special characters are removed from those tweets. The tweets are then displayed with their corresponding dates in the form of a data frame.

2.2 Dataset

After the extraction of tweets, historical data of that particular company or commodity is downloaded from the Yahoo Finance website. Yahoo Finance is a website that provides live stock prices of the company or commodity and also provides downloadable csv files of the historical data.

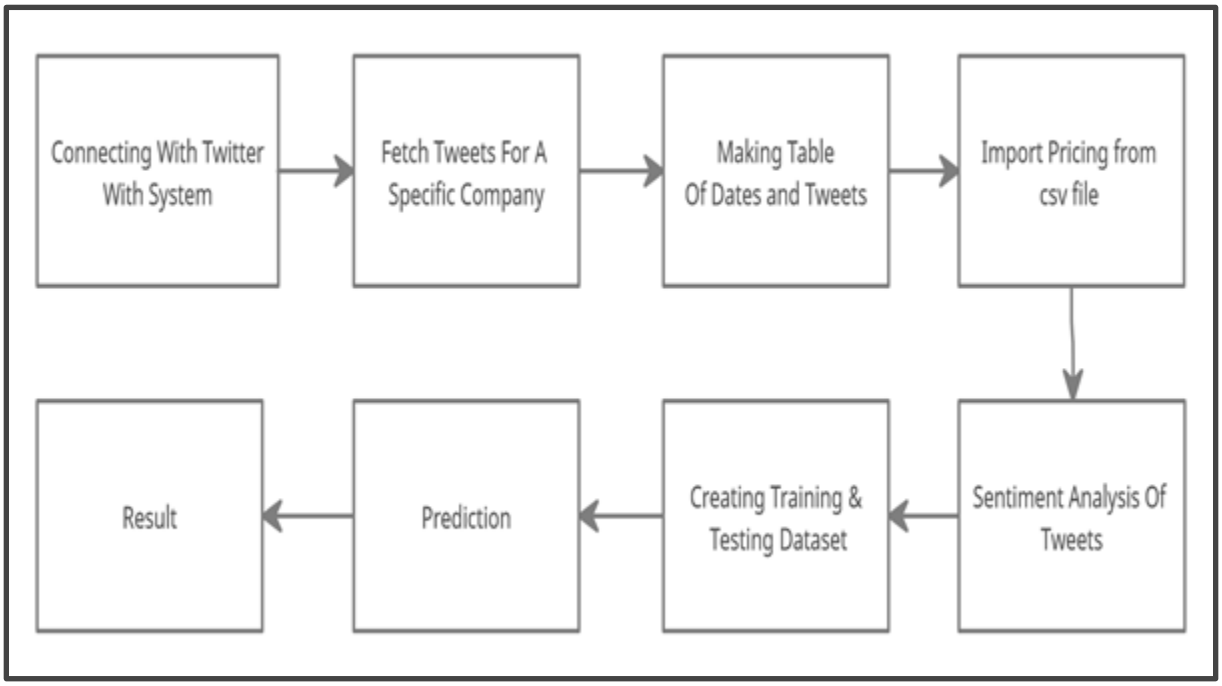

Chart - 1: Flowchart of the proposed system

2.3 Processing of Data

Price's column is then added to the data frame after the historical data is downloaded. Close Price of the company or the commodity is added to the Price column of the data frame. Some dates would not include any price due to some reasons like holidays. To fill in the values of the empty rows of the "Prices" column, mean of the available prices is determined and the empty rows are filled with this mean value.

2.4 Sentiment Analysis

Four new columns are added in the data frame. Comp, Negative, Neutral and Positive. Comp tells whether the sentence or the tweet is overall negative or positive. If the value of Comp is negative then, the sentence is negative and if the value of Comp is positive then, the sentence is positive. Vader (Valence Aware Dictionary and Sentiment Reasoner) a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media [7]. Approximately 44% positive tweets and 55% negative tweets were acquired by performing sentiment analysis using Vader Lexicon.

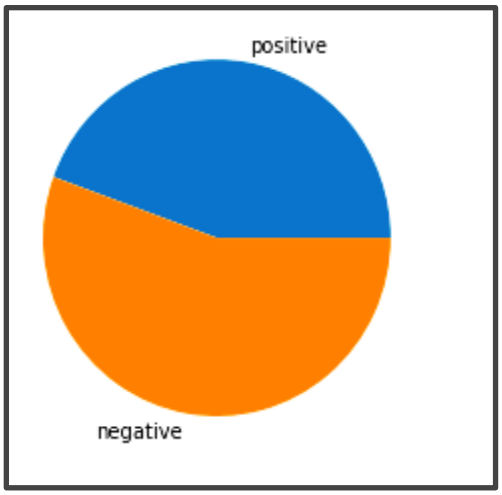

Chart - 2: Sentiment Analysis Pie Chart

3. RANDOM FOREST ALGORITHM

Random forest is a Supervised Learning algorithm which uses ensemble learning methods for classification and regression [4].

It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn't depend on one decision tree but multiple decision trees [5]. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is Aggregation.

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation,

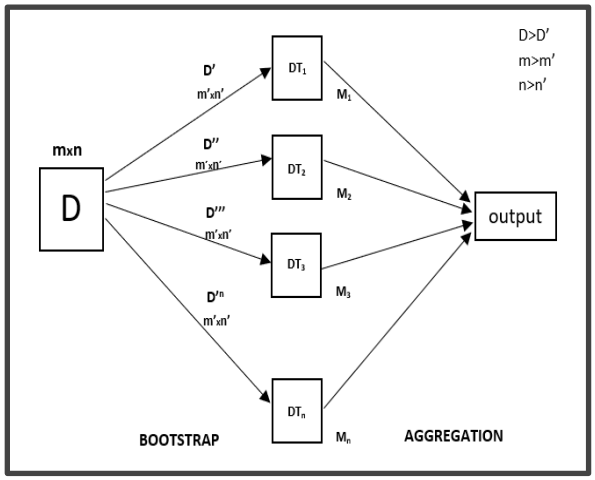

Chart - 3: Random Forest Algorithm

commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models [9]. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

+ Random forest is an ensemble of decision trees. This is to say that many trees, constructed in a certain "random" way form a Random Forest.
+ Each tree is created from a different sample of rows and at each node, a different sample of features is selected for splitting.
+ Each of the trees makes its own individual prediction. These predictions are then averaged to produce a single result.

4. RESULT AND DISCUSSIONS

Jio's current share price was obtained from Yahoo Finance page, which is treated as a basis for success measurements. The predicted stock price for the testing data is compared with the actual price.

The data collected includes 800 tweets, ranging from 14/03/2021 up to 21/03/2021, over seven days of Twitter data, and is saved in a csv format. The collection includes a set of positively and negatively graded ratings of the company. Twitter data collected over the first four days is taken as the training dataset. And for testing, the remaining days are used. Due to the limitation of data available from Twitter API, we had to jump to historical dataset. We acquired an accuracy of 51% using historical dataset.

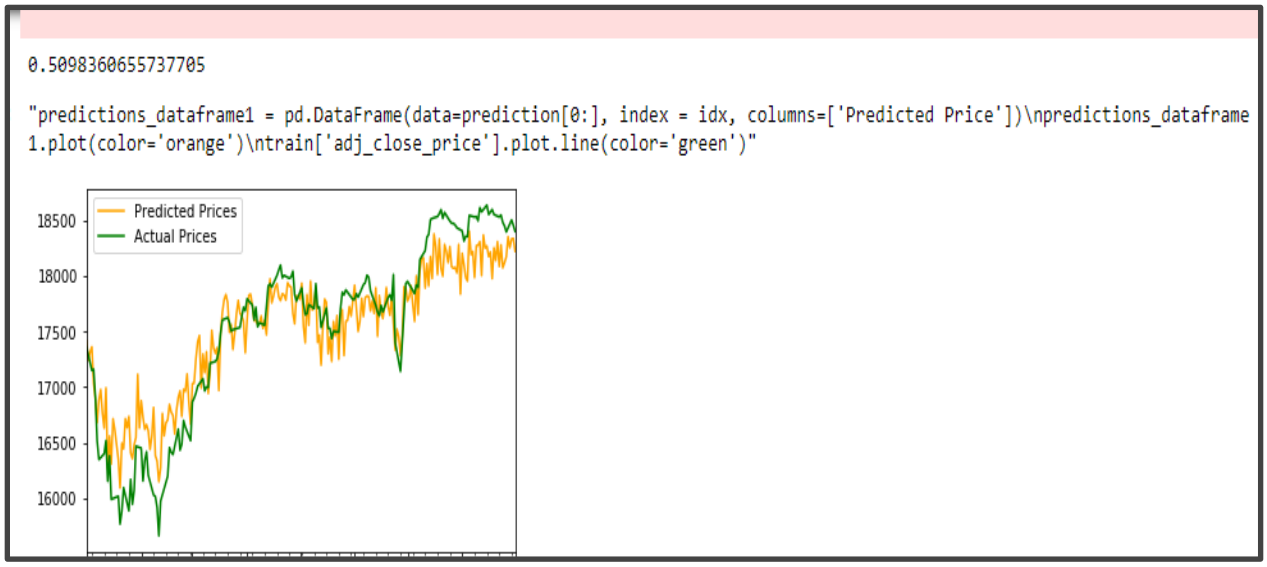

Chart - 4: Final Result

5. CONCLUSION

The Random Forest Regressor is used for predicting future data prices. Dataset is built on the basis of live results. With the assistance of Yahoo Finance, the live data and historic data is retrieved. Also, with support of twitter APIs, required for sentiment analysis is obtained. This algorithm provided an accuracy of 51%. The model was not effective in situations of low or high volatile stock values. There are already various approaches to design stock models, which we leave as work to be done in the future. Some of them involve developing a business model by grouping firms based on their business, taking account of adverse impact on a company's stock price due to news about other similar businesses, and examining more general industry and global news that could suggest general stabilization of the market.

REFERENCES

[1] A. Graves, Supervised Sequence Labelling with Recurrent Neural Networks, ser. Studies in Computational Intelligence. Springer, 2012, vol. 385.

[2] Y. Singh and A. S. Chauhan, "Neural Networks in Data Mining," Journal of Theoretical and Applied Information Technology, pp. 37-42, 2009.

[3] Predicting Stock Price using RNN by Lilian Weng, Unpublished.

[4] A. V. Devadas and T. A. A. Ligori, "Forecasting of stock prices using multi layer perceptron," Int J Comput Algorithm, vol. 2, pp. 440–449, 2013

[5]A.J.P. Samarawickrama and T.G.I. Fernando, A Recurrent Neural Network Approach in Predicting Daily Stock Prices. 978-1-5386-1676-5/17/$31.00 ©2017 IEEE

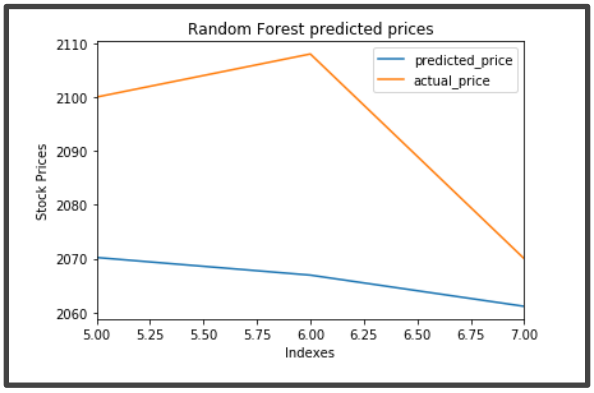

Chart - 5 : Random Forest Prediction Graph

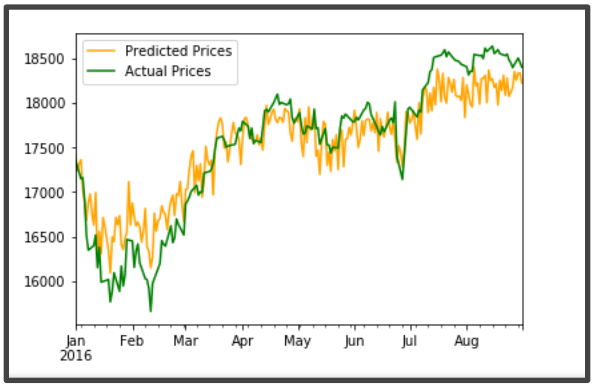

Chart - 6 : Final Graph

[6] A. Pannetrat, R. Molva. "Authenticating real time packet streams and multicasts", Proceedings ISCC 2002 Seventh International Symposium on Computers and Communications, 2002.

[7] Sreelekshmy Selvin, R Vinayakumar, E. A Gopalakrishnan, Vijay Krishna Menon, K. P. Soman. "Stock price prediction using LSTM, RNN and CNNsliding window model", 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2012

[8] Q. LI, L. JIANG,P. LI, H. CHEN : Tensor-Based Learning for Predicting Stock Movements. AAAI Conference on Artificial Intelligence, North
America, feb. 2015.

[9]Aparna Nayak, M. M. ManoharaPai, Radhika M. Pai : Prediction Models for Indian Stock Market, Procedia Computer Science, Volume 89, 2016.