International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217475

Stock Market Prediction Using Twitter Sentiment Analysis

Padmanayana, Varsha, Bhavya K
Department of Computer Science, Srinivas Institute of Technology, Mangalore, Karnataka, India

Article Info
Volume 7, Issue 4
Page Number: 265-270

Publication Issue :
July-August-2021

Article History
Accepted : 15 July 2021
Published : 22 July 2021

ABSTRACT

Stock market prediction is an important topic in financial engineering especially since new techniques and approaches on this matter are gaining value constantly. In this project, we investigate the impact of sentiment expressed through Twitter tweets on stock price prediction. Twitter is the social media platform which provides a free platform for each individual to express their thoughts publicly. Specifically, we fetch the live twitter tweets of the particular company using the API. All the stop words, special characters are extracted from the dataset. The filtered data is used for sentiment analysis using Naïve bayes classifier. Thus, the tweets are classified into positive, negative and neutral tweets. To predict the stock price, the stock dataset is fetched from yahoo finance API. The stock data along with the tweets data are given as input to the machine learning model to obtain the result. XGBoost classifier is used as a model to predict the stock market price. The obtained prediction value is compared with the actual stock market value. The effectiveness of the proposed project on stock price prediction is demonstrated through experiments on several companies like Apple, Amazon, Microsoft using live twitter data and daily stock data. The goal of the project is to use historical stock data in conjunction with sentiment analysis of news headlines and Twitter posts, to predict the future price of a stock of interest. The headlines were obtained by scraping the website, FinViz, while tweets were taken using Tweepy. Both were analyzed using the Vader Sentiment Analyzer.

Keywords : Sentiment Analysis, Stock market prediction, Machine Learning, Twitter

I. INTRODUCTION

Stock market prediction is an important topic in financial engineering especially since new techniques and approaches on this matter are gaining value constantly. Predicting the stock market price is the main challenge for many of the researchers today as it has the complexity for predicting the accurate value which can match the actual stock market price. Stock market prediction is the process of evaluating the future value of the stock of particular company, thus giving an idea of gain or loss to the investors to invest on that particular company stock. Social media plays an important role in predicting the stock price

namely Twitter. Twitter is the social media platform where around millions of tweets are sent daily. Newspaper headlines also provide information related to stock market which can also use for the prediction purpose. Using the twitter data, prediction process can be performed. Various tweets related to different companies are obtained in the Twitter API. There may be many tweets which is not used for prediction purpose. Live twitter data can be extracted from the twitter API and analysed using the classifier. Stock data can be fetched using Yahoo finance API to analyses the value. Various machine learning algorithms are used to train the model to predict the stock price. XGBoost and Naïve bayes are the important classifiers that are used as the training model to provide the accurate value after the prediction.

II. METHODS AND MATERIAL

2.1. Literature Review

Over the past two decades many important changes have taken place in the environment of stock markets. The development of powerful communication and

leading facilities has enlarged the scope of selection for investors as well as for users. Sentiment Analysis is an information extraction task that aims to obtain writer's feelings expressed in positive, negative or neutral comments.

Agarwal and Apoorv in [3] examine the various machine learning techniques on providing a positive or negative sentiment on a tweet. The author uses different techniques are Naïve Bayes, support vector machine etc. Naïve Bayes classifier used to analyze sentiment in the tweet data set and the support vector machine techniques would be used for predicting market movement.

Fazel Zarandi M.H, Rezaee B, Turksen I.B and Neshat E [6] used a type 2 fuzzy rule based expert system is developed for stock price analysis. The proposed type 2 fuzzy model applies the technical and fundamental indexes as the input variables. The type 1 method was used for inferences and to increasing the robustness of the system, flexibility and error minimization.

2.2 Algorithms

In this project we used two main algorithms, Naïve Bayes classifier and XGBoost. Naïve Bayes classifier is used for sentiment analysis. This algorithm is structured to provide either of the three classes: positive, negative and neutral from the news headlines and twitter tweets. Naïve Bayes classifier is one of the simple and most effective classification algorithms which helps in building the fast machine learning models that can make quick productions. It is a probabilistic classifier, which means it predicts on the basis of probability of an object. In sentiment analysis we figure out, if a text express negative or positive feeling. Written reviews are great datasets for doing sentiment analysis because they often write a score that can be used to train an algorithm. Naïve Bayes classification algorithm tends to be a baseline solution for sentiment analysis task. The basic idea of Naïve Bayes technique is to find the probabilities of classes assigned to texts by rising the joint probabilities of words and classes. To avoid underflow, log probabilities can be issued.

XGBoost algorithm is used for stock price prediction. After the sentiment analysis process, we combine it to most recent and in trend algorithm to process with stock data to predict stock price. XGBoost is a most powerful machine learning algorithm today. XGBoost stands for gradient boosted trees and that means it's a big machine learning algorithm with lots of parts remember boosting is an ensemble method. Every tree within or boosting seen here is going to boost the attributes that led to misclassification of previous tree. In boosting, different model get train one after another, so first model gets trained, then the second

model, then the third model and then many models combine to give you a better result. XGBoost is routinely wins Kaggle competitions. It is very easy to use and it is very effective computationally. XGBoost can automatically handle the missing values. Regularized boosting or prevents overfitting, parallel processing, tree pruning some of the features of XGBoost algorithm.

III. RESULTS AND DISCUSSION

1. Proposed System

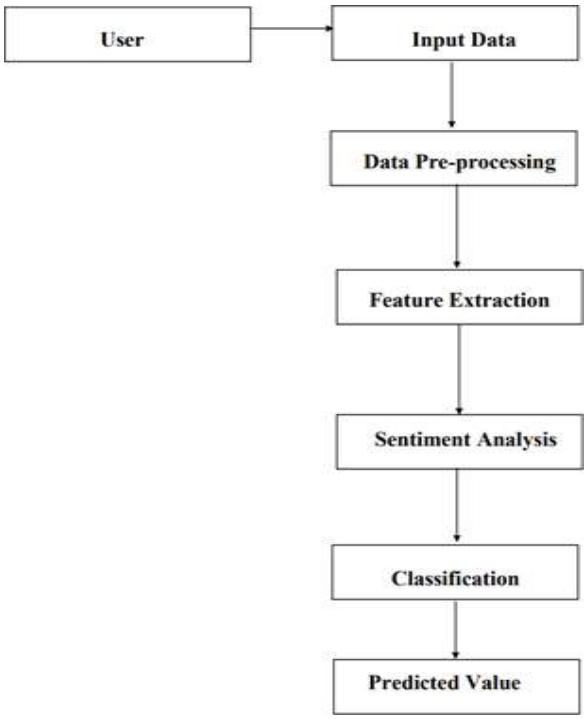

Figure 1. High level structure

Figure 1 illustrates the high-level structure of the Stock market prediction using twitter sentiment analysis. The user is given the option to get the predicted stock price of respective company in the stock market. The user needs to input the name of the company whose stock price has to be predicted. The user can also view the active stocks in the market and also the weekly analysis of the stock market.

In this project we use two main datasets. We are fetching data from twitter, but for accuracy purpose we are collecting data from newspapers and yahoo finance. So, we have taken data from newspaper headlines related to stock of that company and also live twitter data from twitter. These are the two dataset we have taken. From this data we have removed all the special characters including emojis, hashtags (#) and @. These are not necessary for sentiment, so these special characters are removed and we have considered only the plain sentences. When we are performed sentiment analysis in machine learning, the tweets are classified in to three classes: positive, negative and neutral. If you are investing in stock market there is Bullish and Bearish market behaviour. Bullish means the market is going up and Bearish means the market is going down. Neutral is something which is some certain sentences are there which are kind of neutral, either they are positive nor they are negative. Those are like middle sentences, very rare cases that happens we will get like this sentence. Naïve Bayes classifier takes all these data and perform sentiment analysis on this. It fetches the lexical file data line by line and twitter, newspaper headline data what we are taking and it will fetch together. It will classify these in to three classes positive, negative and neutral and we will get the output in dictionary format in python.

2. Procedure for Workflow of The Project

Step 1: Loading the main class which is responsible for training and prediction.
Step 2: Data processing of live twitter data fetched via respective tweets through API.
Step 3: Pre-processing of the fetched data which is done to remove special characters, stop words and perform tokenization.
Step 4: Perform sentiment analysis of the obtained new tweet data using naïve bayes Classifier.

Step 5: Respective company stock data (open, close, adj close) factors taken into consideration along with sentiment analysis data is fed to XGBoost algorithm.
Step 6: Stock price of respective company displaying in between time window of 30 Minutes. Thus, the predicted value will be obtained.

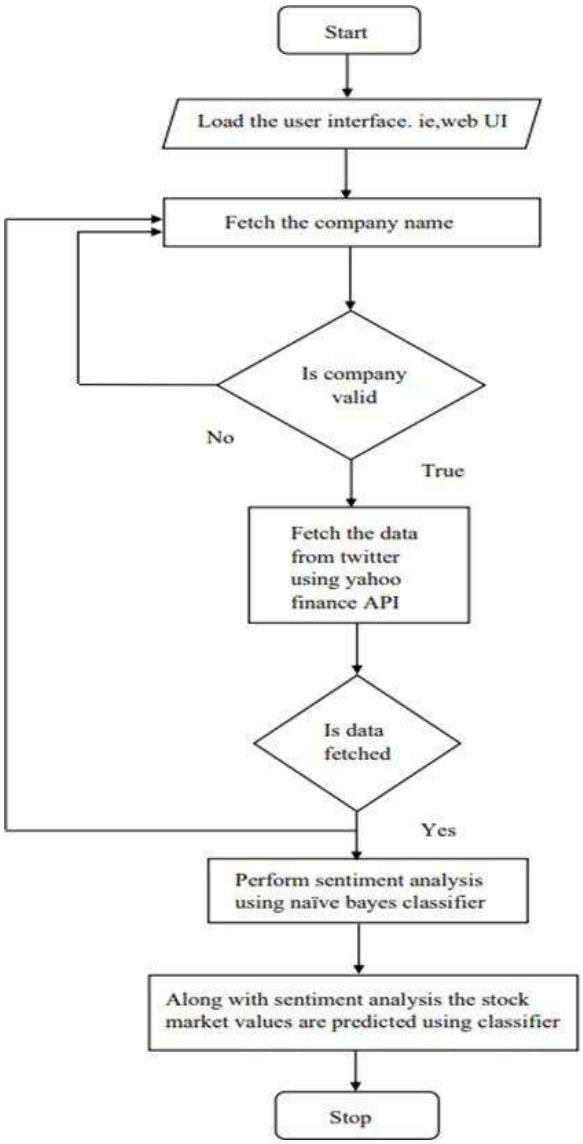

Figure 2. Flowchart for workflow of the project

3. Procedure for User Interface

Step 1: Sign up or login to open the home page
Step 2: Input the name of the company of which the stock price should be predicted.
Step 3: View the stock price by clicking on the check button.
Step 4: Have the option to view the active stocks, analysis of stock data based on week.
Step 5: Logout

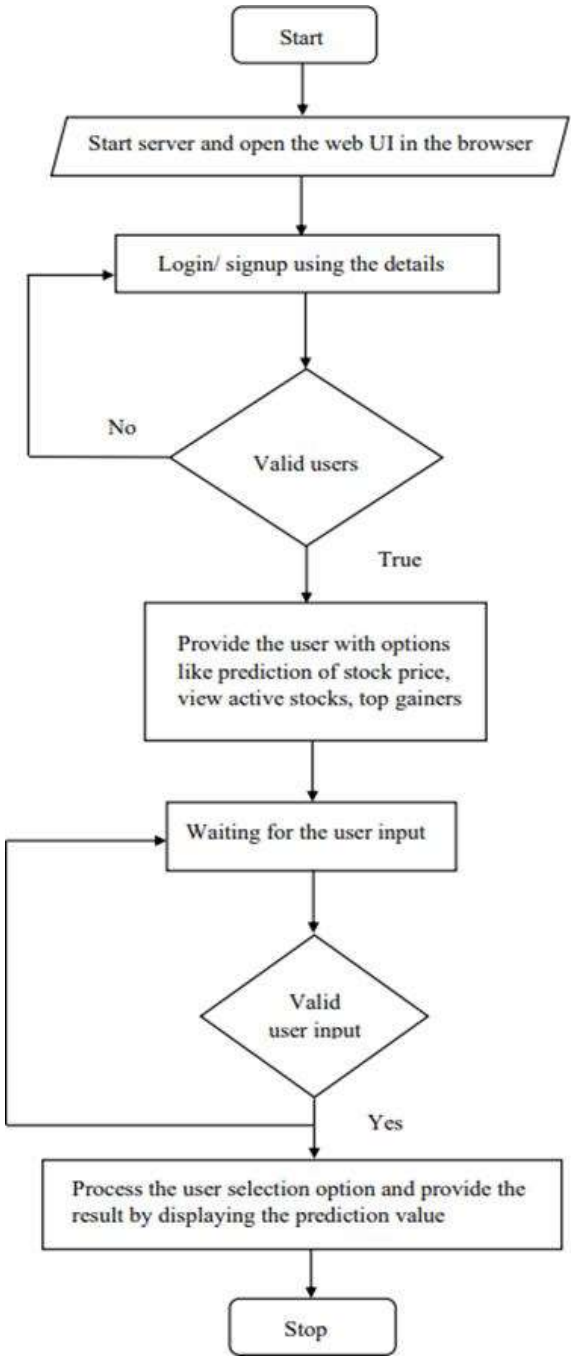

Figure 3. Flowchart for user interface

4. Processing Steps

There are some standard methods involved in this technique. Those are as follows:

4.1 Data Collection

The data collection from twitter, news headline and yahoo finance are collected for analysis.

4.2 Data Preprocessing

Stock data is extracted is not completely understandable because of public holidays and weekends where the stock market does not function. There are missing in the stock value. These empty values can be approximately using simple way. Consider, the stock values on a day is x and the next value present is y with some missing in between. So, the first value is estimated as (y+2)/2 and the same method is used to fill the missing values.

Extracted tweets contains many stop words, unnecessary data like special character, URLs, pictures. These tweets are pre-processed to obtain the emotion of the public. For pre-processing of data, we employ three steps of filtering:

Tokenization: Each tweet is split into individual words called tokens. This process is done to break the text, separated by whitespace character.

Removal of stop words: Words like “a”, “an”, “the”, “he”, “she”, “by”, “on”, etc are not required for sentiment analysis. These are called stop words, which is removed before sentiment analysis process.

Regex Matching: Special characters such as “URL”, “!”, “#”, “@” are all removed and replaced by whitespaces.

4.3 Classification

Use a bag of words containing information on sentiment (positions, negative, neutral) along with sentiment scores. After this, we adopt negation detection measures to differentiate between “good” and “not good”. In this blog we will be trying to do sentiment analysis on twitter dataset and categorizing them into positive, negative and neutral behaviour of people. If the entire review has a positive, joyful attitude on if something is mentioned with positive connections. So, it is considered as a positive statement. If the entire comment has a negative, sad or if something mentioned with negative connections. So, it is considered as a negative statement. If the review expresses no personal opinion in the comments and reviews transmits information.

After the feature extraction we perform sentiment analysis using naïve bayes classifier.

4.4 Stock market prediction

The obtained sentiment analysis data along with stock market data are combined and given as input to the training model. the stock market values are fetched using yahoo finance. The XGBoost classifier evaluates both the data and predicts the stock market value.

IV. CONCLUSION

In this paper we investigated how sentiment analysis of the twitter data is correlated to the prediction of the stock market price for all the companies which are taken. The result obtained after the prediction process clearly specifies that, we have obtained the accurate value which matches with the actual stock price appropriately. The accuracy obtained is 89.8%. Thus, social media such as twitter can be used as a source to predict the stock market price with maximum accuracy. Furthermore, the machine learning model XGBoost provides more accurate values compared with other models. Thus, using sentiment analysis of twitter data and stock data from yahoo finance API, we predict the stock market price which is helpful for predicting future stock price.

In the future, we plan to further improve the work in the following areas. First, our analysis is limited to 16 companies. An expansion to broader set of companies or all Twitter data might yield more

insights into the data, leading to more effective application in stock price prediction. Second, we use the optional sentiment labels provided by Twitter users as the ground truth data for model training. As measured, this data has only 89.8% accuracy, getting better training data is expected to improve the quality of the sentiment analyser. Finally, the current project examines correlation at daily granularity because the stock data are only available at the daily level. It will be interesting to study correlations at a finer granularity such as hourly.

V. REFERENCES

[1]. R. Ahuja, H. Rastogi, A. Choudhuri and B. Garg, "Stock market forecast using sentiment analysis", 2nd International Conference on Computing for Sustainable Global Development, pp. 1008-1010, 2015.

[2]. a. Mittal and a. Goel. "Stock Prediction Using Twitter Sentiment Analysis." Tomx.Inf. Elte.Hu, (June), 2012.

[3]. Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics, 2011.

[4]. W. Antweiler and M. Frank. Do US stock markets typically overreact to corporate news stories? Working Paper, (1998):1–22, 2006.

[5]. Jabaseeli, A. Nisha, and E. Kirubakaran. "A Survey on Sentiment Analysis of (Product) Reviews." International Journal of Computer Applications 47.11, 2012.

[6]. Fazel Zarandi M.H, Rezaee B, Turksen I.B and Neshat E. "A Type-2 Fuzzy Model for Stock Market Analysis.", 2007.

[7]. International Journal of Computer Applications (0975-8887) Volume 121 – No.20, July 2015 Sentiment Analysis on Social Media and Online Review.

[8]. L. A. Gallagher and M. P. Taylor, "Permanent and temporary components of stock prices: evidence from assessing macroeconomic shocks," Southern Economic Journal, vol. 69, pp. 345-362, 2002.

[9]. S. Urolagin, "Text mining of tweet for sentiment classification and association with stock prices," Proceedings of 2017 International Conference on Computer and Applications, pp. 384-388, 2017.

[10]. V. S. Pagolu, K. N. Reddy, G. Panda and B. Majhi, "Sentiment analysis of Twitter data for predicting stock market movements, "Proceedings of 2016 International Conference on Signal Processing, Communication, Power and Embedded System, pp. 1345-1350, 2016.

[11]. B. Qian and K. Rasheed, "Stock market prediction with multiple classifiers," Applied Intelligence, vol. 26, pp. 25-33, 2007.

Cite this article as :

Padmanayana, Varsha, Bhavya K, "Stock Market Prediction Using Twitter Sentiment Analysis", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 4, pp. 265-270, July-August 2021. Available at doi : https://doi.org/10.32628/CSEIT217475
Journal URL : https://ijsrcseit.com/CSEIT217475