International Journal of Research in Engineering, Science and Management
Volume 5, Issue 6, June 2022
https://www.ijresm.com | ISSN (Online): 2581-5792

Stock Market Prediction Using Machine Learning and Twitter Sentiment Analysis: An Implementation Survey

Likitha Thripuranthakam[1*], Shreya B. Gohad[2], H. N. Sujay Singh[3], G. Charan[4], Rashmi Amardeep[5]
[1,2,3,4]*B.E. Student, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, India*
[5]*Associate Professor, Department of Information Science and Engineering, Dayananda Sagar Academy of Technology and Management, Bengaluru, India*

***Abstract*: Sentiment analysis has received a variety of interest in the ultimate decade in particular on the availability of statistics from social media web sites due to this quite a few researchers are showing greater interest on this subject. Many researchers believe that the mood of people or sentiment expressed by using them on social media has an impact on economic markets actions. This paper plans to notice the effect of opinion communicated by means of Twitter on the financial exchange and afterward anticipate the moves of the financial exchange for the next days. For extracting tweets data from Twitter, we will be using Tweepy and for the verification of extracted information, we will use Yahoo Finance. To predict we intend to apply three machine algorithms Autoregressive integrated moving average, long short-term memory, and Linear Regression. The expectation is to apply authentic stock information in relationship with opinion examination of information features and Twitter tweets information, to anticipate the future pace of a stock of interest.**

***Keywords*: Sentiment analysis, stock market prediction, machine learning, twitter.**

1. Introduction

Prediction of stock market movements may be a very challenging task because of non-linear and dynamic nature of stock markets. Now a days the employment of social media has reached another level. The information about public emotions has become considerable on social media. Social media is rebuilding into sort of an optimal stage to share public sentiments about any subject and comprises of huge impact on typical assessment. Techniques which use machine learning will give more accurate, specific and easy to predict stock market movements. This case makes Twitter kind of a corpus with valuable data for researchers. each tweet is of 140 characters long and speaks popular opinion on a subject concisely. The data exploited from tweets are very beneficial for generating predictions. Sentiment analysis of twitter data and sentiment classification is that the task of judging opinion in a very piece of textual content as positive, negative or neutral. During this task a method for predicting stock movements is developed using Twitter tweets about various enterprise. Sentiment analysis of the accumulated tweets is employed for prediction model for locating and analyzing correlation among contents of news articles and stock prices and so making predictions for future prices are developed by the use of machine learning.

2. Experimental Survey

A. System Flow Diagram

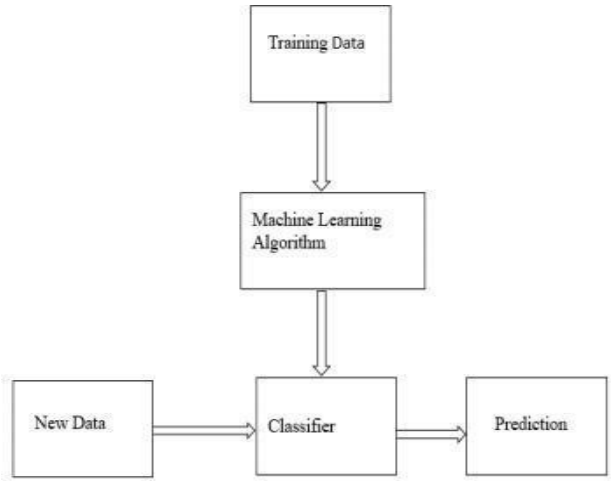

Fig. 1. System flow diagram

System Flow Diagram is basically a graphical and sequential representation of the major steps involved in a systematic process. A SFD (System Flow Diagram) shows what kind of information will be input to and output from the system, where the data will come from and go to, and where the data will be stored.

Steps of how block diagram works:

1. Collection of Data-sets from the twitter which ever we need to check whether it is fake or true news.
2. Clumping of data is nothing but data-preprocessing of data which is taken from real-time datasets of twitter.
3. Feature selection: Splitting the data and find the features which has to be checked.
4. Apply NLP (Natural Language Processing) algorithm to dataset and find the accurate one of the kind.

*Corresponding author: likitha.tv@gmail.com

3. Dataflow Diagram

A dataflow diagram is a graphical representation of the "flow" of data through an information system, modeling its process aspects. A DFD is often used as a preliminary step to create an overview of the system without going into great detail, which can later be elaborated. DFDs can also be used for the visualization of data processing. A DFD shows what kind of information will be input to and output from the system, how the data will advance through the system, and where the data will be stored.

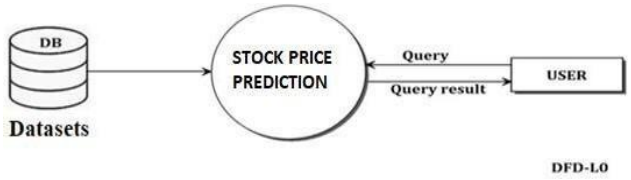

Fig. 2. DFD level zero

DFD Level 1:

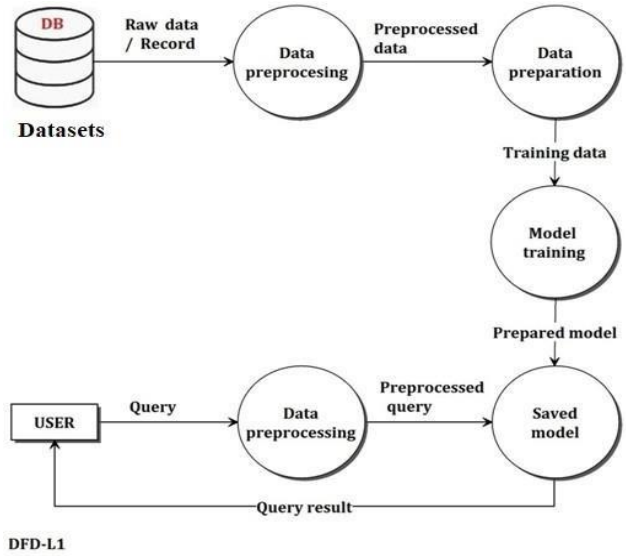

Fig. 3. DFD-L1

4. Implementation

A. Python

Python is a widely used high-level, general-purpose, interpreted, dynamic programming language. Its design philosophy emphasizes code readability, and its syntax allows programmers to express concepts in fewer lines of code than possible in languages such as C or Java. The language provides constructs intended to enable writing clear programs on both a small and large scale. Hence, you can use the programming language for developing both desktop and web applications. Also, you can use Python for developing complex scientific and numeric applications. Python is designed with features to facilitate data analysis and visualization. Python's run time must work harder than Java's. For these reasons, Python is much better suited as a "glue" language, while Java is better characterized as a low-level implementation language. In fact, the two together make an excellent combination.

B. NLTK (Natural Language Tool Kit)

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum. NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language." NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. Natural Language Processing with Python provides a practical introduction to programming for language processing.

C. CSV Files

A CSV file (Comma Separated Values file) is a type of plain text file that uses specific structuring to arrange tabular data. Because it's a plain text file, it can contain only actual text data. The structure of CSV file is given away by its name. CSV files are normally created by problems that handle large amount of data. They are a convenient way of export data from spreadsheets and databases as well as import or use it in other programs.CSV files are very easy to work with programmatically.

5. Methodologies

Data mining methodology is designed to ensure that the data mining effort leads to a stable model that successfully addresses the problem it is designed to solve. Various data mining methodologies have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementation results and monitoring improvements.

Naive Bayes - The naïve bayes classifier is based on prior knowledge of condition that might relate to an event it is based on the bayes theorem there is a strong independence between future assumed. It can easily and fast predict classes of data sets also it can predict multiple classes.

Random Forest algorithm – The random forest model can both run regression and classification module. It divides the data set in to subset and then runs on the data. It can handle thousands of input variables without variable decision. Random forest is often used by financial institution.

A. Logistic Regression

Logistic regression is another supervised learning algorithm which is used to solve the classification problems. In classification problems, we have dependent variables in a binary or discrete format such as 0 or 1. Logistic regression algorithm works with the categorical variable such as 0 or 1, Yes or No, True or False, Spam or not spam, etc. It is a predictive analysis algorithm which works on the concept of probability.

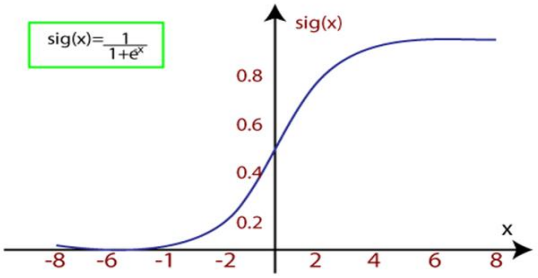

Fig. 4. Graph for Logistic Regression

B. Decision Tree Classifier Algorithm

A decision tree is a flowchart-like tree structure where an internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making.

C. XGBoost

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library.

It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

XGBoost builds upon supervised machine learning, decision trees, ensemble learning, and gradient boosting.

Supervised machine learning uses algorithms to train a model to find patterns in a dataset with labels and features and then uses the trained model to predict the labels on a new dataset's features.

Accuracy:

- Train - 84.24%
- Test – 73.61%

6. Future Scope

Predicting stock market movement through sentimental analysis is a step-by-step process through the literature survey we got to know the basic structure of the whole process. For future work we intend to use of Tweepy to fetch twitter data. To predict we propose to use three algorithms ARIMA, LSTM and Linear Regression. Furthermore, to increase the efficiency more factors such as location of the twitter account, number of retweets and popularity of the twitter user. can be taken into consideration while performing a wistful investigation on the tweets removed from twitter Sentiments from various social media platforms could be incorporated to upgrade the general exhibition of the framework and taking huge informational collections for preparing will assist with expanding the precision of the proposed framework.

7. Conclusion

Predicting movement of Stock Market is an interesting research area. Prior foreseeing Stock Market developments were about arbitrary mathematical expectation in light of the accessible authentic information however with the expansion of ongoing social money individuals' conviction, temperament and response to certain occurrences are additionally taken in thought while anticipating stock development. Similarly, as with the development of web-based entertainment stages and it is made accessible online to expand number of convictions. These are different investigations and explores which proposes that feeling examination of public state of mind got from Twitter channels can be utilized to figure developments of individual stock costs. Through the literature survey it's clear that there are few defined processes which was common almost all the research papers i.e., twitter data extraction, processing of the extracted data, performing a sentimental analysis on the extracted data validating the extracted data using Yahoo Finance and then training a machine learning model by using the same data sets. We propose to use Tweepy to fetch twitter data. To predict we intend to use of three algorithms ARIMA, LSTM and Linear Regression. All this data will then be combined with the sentimental analysis of tweets. And finally, it will recommend whether the price will rise or fall.

References

[1] Saloni Mohan et al" Stock Price Prediction Using News Sentiment Analysis", 2019.
[2] Kesavan M. Karthiraman J et al, "Stock Movement Prediction with Historical Time Series Data and Sentimental Analysis of Social Media Data", 2020.
[3] Niveditha N Reddy et al, "Predicting Stock Price Using Sentimental Analysis Through Twitter Data", 2020.
[4] Rakhi Batra "Integrating Stock Twits with Sentiment Analysis for better prediction of Stock Price Movement", 2018.
[5] Mehar Vijh "Stock Closing Price Prediction using Machine Learning Techniques", 2019.
[6] Arpit Goel "Stock Prediction Using Twitter Sentiment Analysis, Neural Networks", 2019.
[7] Nehal Shah "Stock Market Movements Using Twitter Sentiment Analysis," 2019.
[8] Padmanayana, Varsha, Bhavya K "Stock Market Prediction Using Twitter Sentiment Analysis", IJSRCSEIT, vol 7, Issue 4, pp. 265-270, 2021.
[9] Sai Vikram Kolasani, Rida Assaf "Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks", Journal of Data Analysis and Information Processing, vol 8, pp. 309-319, 2020.
[10] Siti Sakira Kamaruddin, "Conceptual Framework for Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naïve Bayes Classifiers", International Journal of Engineering & Technology, vol. 7, pp. 57-61, 2018.