Stock Market Analysis from social media and News using Machine Learning Techniques

Radhika Baheti[1], Gauri Shirkande[2], Sneha Bodake[3], Janhavi Deokar[4], Archana. K[5]
[1,2,3,4,5]*Department of Computer Engineering, DIT, Pimpri, Pune*

Abstract **- Stock market analysis and prediction is a major factor of profit and growth for investors in the business of any field. Investors check the performance of a company before deciding to purchase its stock, to avoid buying stocks which can be risky. Prediction plays an important role in the business of the stock market which is a very complicated and challenging process. Correct prediction of stocks can lead to huge profits for the sellers and the brokers. Prediction of stocks can be done by carefully analyzing the history of the respective stock market. In this paper, we use different machine learning algorithms on social media and financial news data for stock prediction. We can perform feature selection and spam tweets reduction to improve performance and quality of prediction. Random forest classifiers are found to be more consistent and accurate.**

Index Terms **- Feature selection, Forest Classifier, Machine Learning, Random, Sentiment analysis, Stock market prediction.**

I.INTRODUCTION

Stock market is playing a vital role in the prosperity of many businesses and also in the GDP of a country. If the stock market rises, then countries' economic growth would be high. We can say that the stock market and country growth is tightly bound with the performance of the stock market. As the stock market is too uncertain, there is no surety that the investments made in the market would bear some profits rather it may incur some losses as well. Many factors have been found out that affect the stock prices out of which the historical data has been the most prominent one. However, it was observed that solely historical data does not give the predictions accurately. Factors like social media and daily financial news also affect market values of stocks at once in a positive or negative manner. These all factors must be considered for accurate stock market prediction. Therefore, an automated analyzing system is necessary for investors, as this system will evaluate stock trends automatically using such large amounts of data from social media and news. This automated system can be built using machine learning techniques. Introduction of Machine learning to the area of stock analysis has improved the efficiency and accuracy of the measurements. The main part of machine learning is the dataset used. In this project, supervised machine learning is employed on a dataset obtained. The model is then tested on the remaining dataset. Discovering algorithms that are more effective in predicting stock market trends using external data, like financial news and social media data, is very important as this will contribute to correct stock prediction which will increase investors' profits and make the decision-making process quite easier.

II. RELATED WORK

Prediction of stock prices is a very challenging and complicated process because price movement just behaves like a random walk and time-varying. From the literature survey, it was observed that the application of machine learning techniques to stock market prediction is being undertaken throughout the world. Machine learning techniques are proving to be much more accurate and faster as compared to current prediction techniques.

Significant work has been done for stock prediction earlier which is beneficial for our work related to historical, social media, and financial news data.

For the prediction of the stock market with historical data, contribution by [1] Edgar P. Torres P.1(&), Myriam Hernández-Álvarez1, Edgar A. Torres Hernández2, and Sang Guun Yoo are studied which forecasts and predicts future stock marketdata using artificial intelligence techniques, specifically machine learning algorithms.

For referring to social media related data such as Twitter data, work by [2] Chakraborty, P., Pria, U. S.,

Rony, M. R. A. H., & Majumdar, M. A. (2017) is observed to be very useful. They predicted the future movement of the stock market of the United States by analyzing Twitter sentiment about the stock market. While studying financial news, contribution by [3] Dang M, Duong D (2016) is found very useful as they proved the correlation between the financial news and the stock prices with quite a high accuracy at 73%. Contribution by [4] Wasiat Khan, Mustansar Ali Ghazanfar, Muhammad Awais Azam, Amin Karami, Khaled H. Alyoubi, Ahmed S. Alfakeeh4 is referred for gaining more information about stock market prediction using social media and financial news.

III. MODEL ARCHITECTURE AND TERMINOLOGY USED

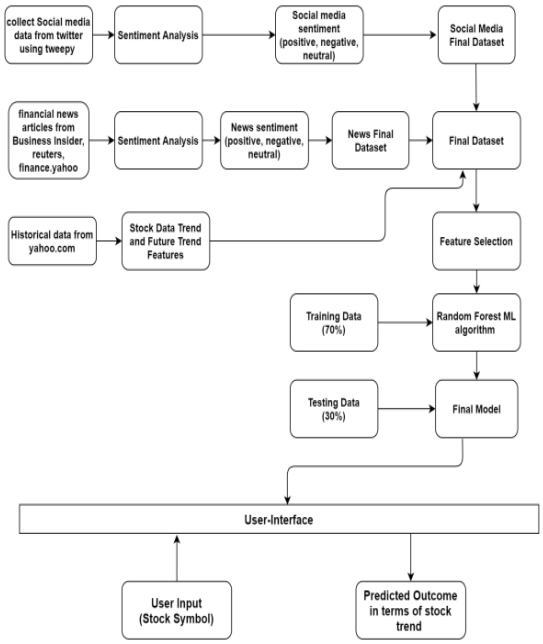


Fig. 1 Architecture Model

Our System Architecture includes steps like data collection, data preprocessing, sentiment analysis and model training. In data collection, historical data, financial news and social media data i.e. tweets are collected using libraries like yfinance, snscrape, scrapy from data sources like yahoo finance, business insider, business standard and twitter. Data preprocessing consists of handling missing values, removing duplicates, emojies, hashtags and stopwords followed by lemmatization. Sentiment analysis is important step in dealing with financial news and social media data. Sentiment analysis helps in classifying data into positive, negative and neutral sentiment. For this purpose popular python library VADER is used. Final dataset is prepared by merging preprocessed data of historical data, financial news and social media data. Random Forest ML algorithm is used for creating model. Further model is trained, tested and then used for trend prediction.

IV. PROPOSED METHODOLOGY

4.1 DATA COLLECTION

In Machine Learning, data collection includes the procedure of collecting, measuring and analyzing accurate insights for research using standard validated techniques. In stock market prediction, data collection is the primary and most important step for further analysis. Data required for prediction is collected from different resources to improve the accuracy from various aspects. Sources such as websites having historical data, financial news and posts from social media are used for the data collection. Twitter is used for collecting data from social media since Twitter is used to keep track of political views, to detect consistency and inconsistency between statements and actions at the government level. Twitter data is collected using snscraper library. Yahoo finance is used for historical data for large amounts of data. The dataset of historical data comprises the following attributes including open, close, low, high and volume. Open, close, low and high are various types of prices for the stock at different times with nearly direct names. The volume is the number of shares that are bought and sold in a particular time period. Financial news websites such as Business Insider, Business standards and Finviz are used to gather recent news related to the company using web scraping.

4.2 DATA PREPROCESSING

Data preprocessing is a technique of data mining that involves transforming raw data into an understandable format. Data from the real world is often incomplete, inaccurate, missing in certain ways, and is likely to contain many errors. Data preprocessing is a renowned method of resolving such problems. Data preprocessing involves steps such as data cleaning, data transformation and data reduction. The data can have many unnecessary and inconsistent information. To handle this, data cleaning is done. Data transformation is used to transform the data into the

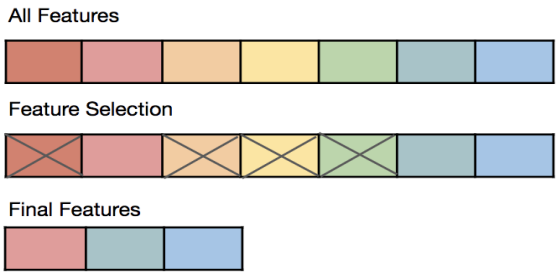
appropriate formats suitable for data mining processes. While working with large amounts of data, analysis becomes difficult. In order to avoid this, we use data reduction techniques. It aims to improve the storage efficiency of data and reduce the storage costs due to the irrelevant data.

4.3 SENTIMENT ANALYSIS

Sentiment analysis is the process which determines whether a piece of text is positive, negative, or neutral. Sentiment analysis is analysis of words which indicates the social sentiment of a commodity or brand and also helps to determine the social opinion on the rise and fall of the market value of shares of any particular company. Sentiment analysis has gained importance because of the availability of large amounts of data on social media and news platforms. This data from social media can be mined for analyzing opinions of users for stock prediction. For sentiment analysis of this huge volume of textual data, data mining and machine learning carry huge importance. Researchers working on machine learning have carried out research on mining opinions of users of these platforms.

4.4 FEATURE EXTRACTION

Feature Extraction is a technique of extracting the features that are useful when you have a large data set and need to reduce the number of resources without losing any important or relevant information. Feature extraction helps to reduce the amount of redundant data from the data set. The reduction of the data helps to build the model with less machine's efforts and also increases the speed of learning and generalization steps in the machine learning process. New reduced set of features should be able to summarize most of the information contained in the original set features.



Commonly used technique to reduce the number of features in a dataset is Feature Selection. The main difference between Feature Selection and Feature Extraction is that feature selection aims to rank the importance of existing features in the dataset and discard less important ones (no new features are created).

4.5 MACHINE LEARNING CLASSIFIER

Classification is the process of which predicts the class of given data points. Classes are also called as targets, labels or categories. Classification predictive modeling is the task which approximates a mapping function (f) from input variables (X) to discrete output variables (y). The input data for classification can range from the text, images, documents to time-series data. For our model it will be a stock symbol. These classifiers are trained and tested on the final data sets to identify future stock market trends. Random forest classifiers are used for prediction purposes.

4.5.1 RANDOM FOREST CLASSIFIER

Random forest classifier is a machine learning algorithm which is a type of supervised machine learning. Random Forest classifier that contains a number of decision trees on various subsets for the given dataset and then it takes the average of the dataset to improve the predictive accuracy. Instead of relying on one decision tree, the random forest takes the prediction from each of the trees and is based on the majority votes of predictions It predicts the final output.

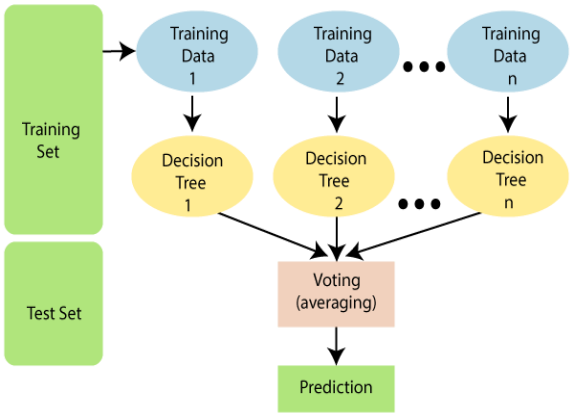


Fig. 2 Random Forest Classifier

The process of working can be explained in the following steps and diagram:

Step:1- Select data points of random K from the training set.

Step:2- The decision trees built should be associated with the selected data points (Subsets).

Step:3- Choose the number N for decision trees, you want to build.
Step:4- Repeat 1st and 2nd step.
Step:5- Find the predictions of each decision tree for new data points and assign the new data points to the category that wins the majority votes.

V. EVALUATION METRICS

Performance evaluation of classifiers is done with accuracy primary classification metric and three within-class classification metrics namely, precision, recall, and F-measure. Accuracy can be calculated by Accuracy = Number of correct predictions/Total number of predictions.
Precision is the skill of the model to classify samples accurately,
Precision = TP /(TP + FP)
where, the true positive rate is TP and the false positive rate is FP.
Recall shows the skill of the model to classify the maximum possible samples,
Recall = TP/(TP + FN)
where FN is the false negative rate of the algorithm.
F-measure describes both precision and recall and
F-measure = 2 × (Precision × Recall) / (Precision + Recall).

VI. IMPLEMENTATION DETAILS

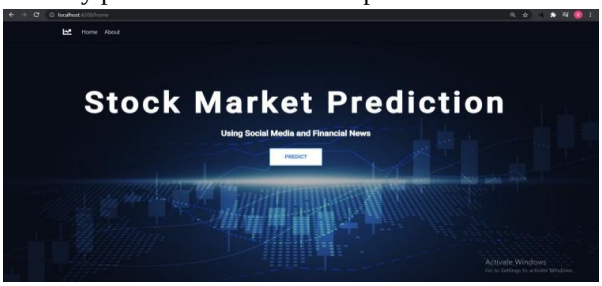
Implementation includes Random Forest Classifier model which provides us trend of company's stock. As an input to the model, we need to provide preprocessed data and relevant steps are performed to prepare it.

6.1 System Description:
Input: The user will select the company to get the expected next day predicted value of stock price.
Output: Mainly provides suggestions to Buy or Sell the stock. Also gives the predicted value of the next day for further investment purposes.Further, it returns the associated close price graphs for the reference purpose along with few sentiment pie charts that show the impact of social media on the stock value.

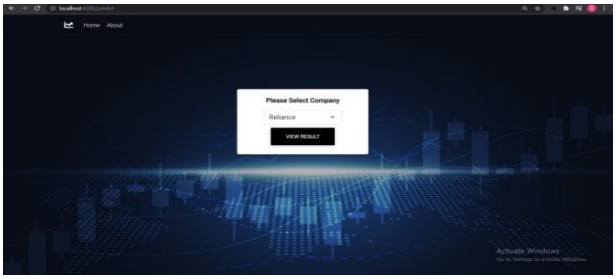
6.2 Software Requirements:
Operating System : Windows 10
Technology : Python
Web Technologies : Angular
Python Version : Python 3
Development platform: Google Colaboratory

6.3 Hardware Requirements:
Operating System: Windows Operating System
Processor: Intel Core i5 9th generation
Ram: 1GB or more
High Internet speed

6.4 User Interface:
The user will be provided with Graphical User Interface to select the company to get the expected next day predicted value of stock price.



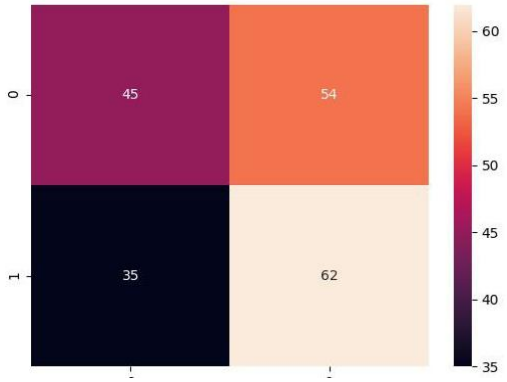
After selecting the company system will mainly provide suggestions to Buy or Sell the stock. Also it will give the predicted value of the next day for further investment purposes. Further, it returns the associated close price graphs for the reference purpose along with few sentiment pie charts that show the impact of social media on the stock value.



VII. RESULTS AND DISCUSSION

7.1 Final Output

7.2 Confusion Matrix



A confusion matrix is an NxN matrix used for evaluating the performance of a classification model, where N is the number of target classes i.e., buy and sell.

VIII. CONCLUSION

This research study is to help stock investors and brokers for investing money in the stock market. Prediction plays a very important role in the stock market. It will determine whether to buy or sell the stock of a particular company with greater accuracy and reliability using machine learning technique.

REFERENCES

[1] Edgar P. Torres P.1(&), Myriam Hernández-Álvarez1, Edgar A. Torres Hernández2, and Sang Guun Yoo. Stock Market Data Prediction Using Machine Learning Techniques

[2] Chakraborty, P., Pria, U. S., Rony, M. R. A. H., & Majumdar, M. A. (2017). Predicting stock movement using sentiment analysis of the Twitter feed.

[3] Dang M, Duong D (2016) Improvement methods for stock market prediction using financial news articles. In: IEEE 3rd national foundation for science and technology development conference on information and computer science (NICS), pp 125–12.

[4] Wasiat Khan, Mustansar Ali Ghazanfar, Muhammad Awais Azam, Amin Karami, Khaled H. Alyoubi, Ahmed S. Alfakeeh4. Stock market prediction using machine learning classifiers and social media, news.

[5] Khan W, Malik U, Ghazanfar MA, Azam MA, Al Youbi KH, Alfakeeh AS (2019) Predicting stock market trends using machine learning algorithms via public sentiment and political situation analysis.

[6] Usmani M, Adil SH, Raza K, Ali SA (2016) Stock market prediction using machine learning techniques. In: IEEE 3rd international conference on ICCOINS, pp 322–327

[7] Vargas MR, dos Anjos CEM, Bichara GLG, Evsukoff AG (2018) Deep learning for stock market prediction using technical indicators and financial news articles. In: IEEE international joint conference IJCNN, pp 1–8

[8] Chen W, Zhang Y, Yeo CK, Lau CT, Lee BS (2017b) Stock market prediction using neural networks through news on online social networks. In: IEEE international ISC2, pp 1–6

[9] Qasem M, Thulasiram R, Thulasiram P (2015) Twitter sentiment classification using machine learning techniques for stock markets. In: IEEE international conference on ICACCI, Kochi, India, pp 834–840

[10] Khatri, S. K., & Srivastava, A. (2016). Using sentimental analysis in the prediction of the stock market investment. 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions).

[11] A. C. Jishag, A. P. Athira, Muchintala Shailaja and S. Thara Predicting the Stock Market Behavior Using Historic Data Analysis and News Sentiment Analysis

[12] 2018 IEEE International Conference on Data Mining Workshops (ICDMW), Zhaoxia WANG, Seng-Beng HO, Zhiping LIN Stock Market Prediction Analysis by Incorporating Social and News Opinion and Sentiment

[13] Li X, Xie H, Chen L, Wang J, Deng X (2014c) News impact on stock price return via sentiment analysis. J Knowl-Based Syst 69:14–23

[14] Wang, F., Zhao, Z., Li, X., Yu, F., & Zhang, H. (2014). Stock volatility prediction using multi-kernel learning based extreme learning machines. 2014 International Joint Conference on Neural Networks.

[15] Lakshmi V, Harika K, Bhavishya H, Harsha CS (2017) Sentiment analysis of twitter data. Int Res J Eng Technol 4(2):2224–2222

[16] Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, Babita Majhi (2016) Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. 2016 International

conference on Signal Processing, Communication, Power and Embedded System (SCOPES)

[17] Joshi, R., & Tekchandani, R. (2016). Comparative analysis of Twitter data using supervised classifiers. 2016 International Conference on Inventive Computation Technologies (ICICT).

[18] Khare, K., Darekar, O., Gupta, P., & Attar, V. Z. (2017). Short term stock price prediction using deep learning. 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology.

[19] Monica Tirea, Viorel Negru Text Mining News System - Quantifying Certain Phenomena Effect on the Stock Market Behavior.