Research Article

Design of Stock Market Analysis and Prediction System using Social Media Mining

Aarti Ghodke and Nilesh Alone

Dept. of Computer Engineering R.H.Sapat College of Engineering Management Studies and Research

Abstract

The main objective of this paper is to find the best model to predict the value of the stock market. During the process of considering various techniques and variables that must be taken into account, we found out that techniques like support vector machine is not suitable for these applications. In, this system we are going to present and review a more feasible method to predict the stock movement with higher accuracy. The first thing we have taken into account is the dataset of the stock market prices from previous year. The dataset was preprocessed and tuned up for real analysis. We also proposed the sentiment analysis system for the social media data for getting the information about the views of the users of the particular stocks. This will represent the overall ranking and the quality of the stocks in market. This also helps the user to get the knowledge about the stock market. Hence, our system will also focus on data preprocessing of the raw dataset. Secondly, after pre- processing the data, we will review the use of Naives Bayesianon the dataset and the outcomes it generates. In addition, the proposed system examines the use of the prediction system in real-world settings and issues associated with the accuracy of the overall values given. The system also presents a machine learning model to predict the longevity of stock in a competitive market. The successful prediction of the stock will be a great asset for the stock market institutions and will provide real-life solutions to the problems that stock investors face.

Keywords: Stock exchanges, prediction, statistics, sentiment analysis, machine learning, pattern recognition

Introduction

Quantitative dealers with a great deal of cash from securities exchanges purchase stock subsidiaries and values at a modest cost and later on selling them at a significant expense. The pattern in a securities exchange forecast is definitely not another thing but then this issue is continued being talked about by different associations. There are two sorts to break down stocks which financial specialists perform before putting resources into a stock, first is the principal investigator, in this examination speculators take a gander at the inherent estimation of stocks, and execution of the business, economy, political atmosphere, and so on to choose that whether to contribute or not. Then again, the specialized examination is an advancement of stocks by the methods for contemplating the measurements created by advertising action, for example, past costs and volumes.

Lately, expanding the noticeable quality of ML in different businesses has illuminated numerous dealers to apply ML methods to the field, and some of them have delivered very encouraging outcomes. The principal reason for the forecast is to decrease the vulnerability related to speculation basic leadership. Securities exchange pursues the irregular walk, which infers that the best expectation you can have about tomorrow's worth is the present worth. Unquestionably, the anticipating stock files are extremely troublesome in light of the market unpredictability that needs an exact figure model. The securities exchange records are exceptionally fluctuating and it influences the speculator's conviction. Stock costs are viewed as an exceptionally unique and defenseless to fast changes on account of hidden nature of the money related space and to some degree in view of the blend of known parameters (Earlier day's end value, P/E proportion, and so forth.) and the obscure components (like Political decision Results, Gossipy tidbits, and so on.). There have been various endeavors to anticipate stock costs with ML. The focal point of each examination undertaking changes a ton in three different ways. (1) The focusing on value change can be close term (not exactly a moment), present moment (tomorrow to a couple of days after the fact), and a long haul (months after the fact), (2) The arrangement of stocks can be in restricted to under 10 specific stock, to stocks specifically industry, to for the most part all stocks. (3)

The indicators utilized can extend from worldwide news and economy pattern to specific qualities of the organization, to simply time arrangement information of the stock cost. The plausible financial exchange forecast target can be the future stock cost or the instability of the costs or market patterns. In the expectation, there are two sorts like a sham and an ongoing forecast which is utilized in the securities exchange forecast framework. In Sham's expectation, they have characterized some arrangement of rules and foresee the future cost of offers by computing the normal cost. Progressively expectation, mandatory utilized the web and saw the present cost of portions of the organization. Computational advances have prompted the presentation of ML procedures for prescient frameworks in money related markets. In this paper, we are utilizing an ML system i.e., Naive Bayesian to foresee the securities exchange and we are utilizing Java and Python language for programming.

LITERATURE SURVEY

Data mining (the analysis step of the "Knowledge Discovery and knowledge Mining" method, or KDD), knowledgebase subfield of engineering science, is that the machine method of discovering patterns in gain knowledge sets involving strategies at the intersection of computing, machine learning, statistics, and information systems.

The general goal of the mining technique is to extract data from AN data set and rework it into a transparent structure for extra use. Except for the raw analysis step, it involves information and knowledge management aspects, knowledge pre-processing, model and reasoning issues, power metrics, quality issues, postprocessing of discovered structures, visualization, Andonline change.[2]Data mining methodology is meant to confirm that the information mining effort results in a stable model that with success addresses the matter it's designed to resolve. Various data processing methodologies are projected to function blueprints for a way to prepare the method of gathering information, analyzing data, disseminating results, implementing results, and monitoring improvements [9].To build the model that analyses the stock trends mistreatment the choice tree technique, the CRISPDM (Cross Industry Standard Process for data mining) is used. This methodology was projected within the mid-1990s by an ECU pool of corporations to function a generic normal method model for data processing.[3]Economic conditions greatly deteriorated in the Great Recession. Unemployment increased drastically, making the Great Recession the worst "labor market downturn since the Great Depression". The S&P500 index dropped 38.49% in 2008 and then increased by 23.45% the next year[1].

Every year except 2011, the index had a doubledigit increase in price.We wished to check however the SVM model, which has had such success in previous literature, would work in such an abnormally volatile market.Although Rosillo, et al found that SVM has better accuracy in high-volatility markets than other types of markets, their study used simulated markets, whereas we used historical data from the Great Recession period [14]. We focus specifically on the technology sector. Focusing on a sector as opposed to the broad market allows us to test the model on companies that are similar to each other, making our results relatively standardized.We use the NASDAQ100 Technology Sector Index (NDXT) because of the general technology sector index.The index consists of technology giants like Microsoft and Apple along with smaller companies like Akamai and NetApp.[4]Stock market prediction involves predicting the longer-term worth of company stock or alternative monetary instruments listed on AN exchange.Various types of trading can be done in the stock market. It could be short term trading or even longterm trading but if someone can predict the value or class of that entity, it can yield a very good return for the investment done. Before the evolution of the digital world, predictors continued to use paperwork methods like fundamental and technical analysis. Various useful technical indicators like SMA, EMA, MACD found to be very useful but with the advent of computer technologies and algorithms, prediction moved into the technological realm. Analysts started building prediction system using Neural Network, Support Vector Machine, Decision Trees, Hidden Markov Model. Prediction accuracy improved using an algorithmic approach. This survey covers various traditional as well as evolutionary data mining techniques used for stock market prediction.[5]

Classification is a type of supervised learning (machine learning) in which some decision is taken or prediction is made based on information which is currently available and the procedure of carrying out classification is a formal method which is used for constantly making such judgments in different and new situations. The formation of a classification method from a data set for which the true classes are known is also known as pattern recognition, supervised learning or discrimination (to differentiate it from unsupervised learning in which the classes are always inferred from the data). Classification is used in many situations like the most difficult situations arising in science, industry, and commerce can be determined by classification or decision problems that use complex and often very extensive data.[6]

The model on the Stock market one-day ahead movement prediction using disparate data sources was proposed to evaluate the performance of the expert system, the researchers present a case study based on the AAPL (Apple NASDAQ) stock. They believed that their expert system had an 85% accuracy in predicting the next-day AAPL stock movement, which outperforms the reported rates in the literature. To predict stock movements, the researchers Bin Weng, Mohamed A. Ahmed and Fadel M. Megahed, propose a data-driven approach that consists of three main phases. They also populated additional features

(i.e. summary statistics) in an endeavor to uncover a lot of vital predictors for stock movement.

Based on the evaluation, they select an appropriate model for real-time stock market prediction. The results of the study suggest: [7]

Kartik Sharma, Akhilesh Rao, conducted a study on Stock Market Analysis. The proposed system was an attempt to reconcile computed sentiments alongside traditional/more common data mining. This will be accomplished with the help of 2 types of datasets. Firstly, historical data from Google Finance will be mined to garner traditionally available predictions. Two independent predictions are then combined to generate a final output which will be used to predict the next day's opening price. Datasets consisting of historical data as well as recent headlines were mined to ascertain stock price movement. The main aim of the system was to predict stock price movement more accurately by emulating instinctual reasoning by implementing sentiment analysis. It helped the proposed system ascertain sentiment analysis as the better companion to the traditional data mining approach instead of employing a neural network in cases that called for the supervised approach. It was noted however that a neural network worked extremely well in situations that called for the unsupervised approach. The research concludes that unsupervised and supervised learning depends on methods that help to find the results in a better way. The combinational study was done to get better accuracy. Further, optimizations can be done in sequence to get improved results.[8]

Bhardwaj, A. et, al. conducted a study on Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty. The main focus of the research work was the importance of sentiment analysis for stock market indicators such as Sensex and Nifty to predict the price of the stock. For implementation purposes, the proposed system fetched the live Sensex and Nifty data values from Timesofindia.com. A python script was run with a sleep count time interval of one second for fetching the data, and values were calculated for a different time interval. After that result is drawn which shows that for a particular time interval the fetched values of Sensex and Nifty remain constant. It was proposed that we should use python scripting language which has a fast execution environment and this will help out the investors to predict what shares money should be invested, it will also help in maintaining the economical balance of the share market.[9]

Problem Definition

To design the system using machine learning for prediction of stocks from share market by using various algorithms that will help the user to predict the stock prices on the basis of historical data. Stock market prediction is basically defined as trying to determine the stock value and offer a robust idea for the people to know and predict the market and the

stock prices. It is generally presented using the quarterly financial ratio using the dataset. Thus, relying on a single dataset may not be sufficient for the prediction and can give a result which is inaccurate. Hence, we are contemplating towards the study of machine learning with various datasets integration to predict the market and the stock trends. The problem with estimating the stock price will remain a problem if a better stock market prediction algorithm is not proposed. Predicting how the stock market will perform is quite difficult. The movement in the stock market is usually determined by the sentiments of thousands of investors. Stock market prediction, calls for an ability to predict the effect of recent events on the investors. These events can be political events like a statement by a political leader, a piece of news on scam etc. It can also be an international event like sharp movements in currencies and commodity etc. All these events affect the corporate earnings, which in turn affects the sentiment of investors. It is beyond the scope of almost all investors to correctly and consistently predict these hyperparameters. All these factors make stock price prediction very difficult. Once the right data is collected, it then can be used to train a machine and to generate a predictive result.
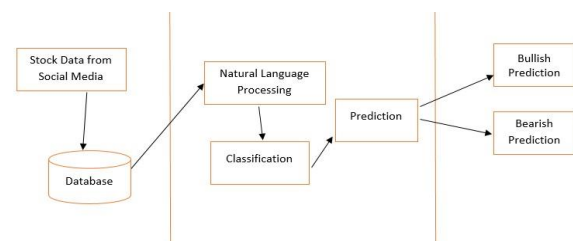
Proposed System

Fig.1 System Architecture

Alpha Vantage is an online community for data analysis and predictive modeling. It also contains dataset of different fields, which is contributed by data miners. Various data scientist competes to create the best models for predicting and depicting the information. It allows the users to use their datasets so that they can build models and work with various data science engineers to solve various real-life data science challenges. The dataset used in the proposed project has been downloaded from Kaggle. However, this data set is present in what we call raw format. The data set is a collection of stock market information about a few companies. The first step is the conversion of this raw data into processed data. This is done using feature extraction, since in the raw data collected there are multiple attributes but only a few of those attributes are useful for the purpose of prediction. So the first step is feature extraction, where the key attributes are extracted from the whole list of attributes available in the raw dataset. Feature extraction starts from an initial state of measured data and builds derived values or features. These features are intended to be

informative and non-redundant, facilitating the subsequent learning and generalization steps. Feature extraction is a dimensionality reduction process, where the initial set of raw variables is diminished to progressively reasonable features for ease of management, while still precisely and totally depicting the first informational collection. The feature extraction process is followed by a classification process wherein the data that was obtained after feature extraction is split into two different and distinct segments. Classification is the issue of recognizing to which set of categories a new observation belongs. The training data set is used to train the model whereas the test data is used to predict the accuracy of the model. The splitting is done in a way that training data maintain a higher proportion than the test data. The Naives Bayesian algorithm utilizes a collection of random decision trees to analyze the data. In layman terms, from the total number of decision trees in the forest, a cluster of the decisiontrees look for specific attributes in the data. This is known as data splitting. In this case, since the end goal of our proposed system is to predict the price of the stock by analyzing its historical data.

Modules Identification:-
The various modules of the project would be divided into the segments as described.

I. Data Collection Data collection is a very basic module and the initial step towards the project. It generally deals with the collection of the right dataset. The dataset that is to be used in the market prediction has to be used to be filtered based on various aspects. Data collection also complements to enhance the dataset by adding more data that are external. Our data mainly consists of the previous year stock prices. Initially, we will be analyzing the Kaggle dataset and according to the accuracy, we will be using the model with the data to analyze the predictions accurately.

II. Pre Processing Data pre-processing is a part of data mining, which involves transforming raw data into a more coherent format. Raw data is usually, inconsistent or incomplete and usually contains many errors. The data preprocessing involves checking out for missing values, looking for categorical values, splitting the data-set into training and test set and finally do a feature scaling to limit the range of variables so that they can be compared on common environs.

III. Training the Machine Training the machine is similar to feeding the data to the algorithm to touch up the test data. Thetraining sets are used to tune and fit the models. The test sets are untouched, as a model should not be judged based on unseen data. The training of the model includes crossvalidation where we get a well-grounded approximate performance of the model using the training data. Tuning models are meant to specifically tune the hyperparameters like the number of trees in a Naive Bayesian. We perform the entire cross-validation loop on each set of hyperparameter values. Finally, we will calculate a crossvalidated score, for individual sets of hyperparameters. Then, we select the best hyperparameters. The idea behind the training of the model is that we some initial values with the dataset and then optimize the parameters which we want to in the model. This is kept on repetition until we get the optimal values. Thus, we take the predictions from the trained model on the inputs from the test dataset. Hence, it is divided in the ratio of 80:20 where 80% is for the training set and the rest 20% for a testing set of the data.

IV.Data Scoring The process of applying a predictive model to a set of data is referred to as scoring the data. The technique used to process the dataset is the Naive Bayesian Algorithm. Naive Bayesian involves an ensemble method, which is usually used, for classification and as well as regression. Based on the learning models, we achieve interesting results. The last module thus describes how the result of the model can help to predict the probability of a stock to rise and sink based on certain parameters. It also shows the vulnerabilities of a particular stock or entity. The user authentication system control is implemented to make sure that only the authorized entities are accessing the results.

Algorithm Used

1. Naïve Bayesian Algorithm:
A classifier is a machine learning model that is used to discriminate different objects based on certain features. A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

P(A/B) = P(B/A)P(A) / P(B)

The NB algorithm is used for the classification of the data on the basis of the probability of the classes is given. We have to calculate the total probability of the system. The system contains two classes
1. High price and
2. Low Price

Steps: -
1. We collect the data i.e. comments from the Facebook post by using Facebook developers API.
2. Then the comments may have the noise data. So we need to clear the noise data from comments.
3. After the noise removal the comments will contains only the keywords which has some meanings.
4. After that we calculate the total no of comments from the posts.
5. then we classify the comments in two classes.
6. After the classification we calculate the probability of each class with total no of comments in the post. 7. The highest probability of the class is given to the post.

Data sets used for system:- 1. Alpha Vantage:-
Composed of a tight-knit community of researchers, engineers, and business professionals, Alpha Vantage Inc. is a leading provider of free APIs for real-time and historical data on stocks, physical currencies, and digital/crypto currencies. Our success is driven by rigorous research, cutting edge technology, and a disciplined focus on democratizing access to data.
This suite of APIs provides realtime and historical global equity data in 4 different temporal resolutions:

(1) daily,
(2) weekly,
(3) monthly, and
(4) intraday. Daily, weekly, and monthly time series contain 20+ years of historical data.

API key is required for the access these data. This dataset is freely available for users.

Table 1. Data set from Alpha Ventage, Previous Year dataof Stock : Reliance

date	1. open	2. high	3. low	4. close	5. volume
10/22/2019	1611.65	1627.4	1416.45	1535.25	14471640
9/30/2019	1500	1709	1423.85	1608.15	15027045
8/30/2019	1554.7	1617	1441.6	1519.85	14842785
7/31/2019	1565.05	1647.55	1480.65	1560.4	22933279
6/28/2019	1696.85	1735.85	1470.1	1562.15	11135309
5/31/2019	1628	1769.05	1541	1686	14111220
4/30/2019	1673	1723.4	1570	1632.05	20442720
3/28/2019	1422.3	1629.4	1420.3	1617.8	17531454
2/28/2019	1416.3	1457.3	1326	1422.3	16156250
1/31/2019	1516.9	1516.9	1339.4	1423.35	10485774

2. Facebook Post:-

The Facebook Platform is the set of services, tools, and products provided by the social networking service Facebook for third-party developers to create their own applications and services that access data in Facebook. The current Facebook Platform was launched in 2010. The platform offers a set of programming interfaces and tools which enable developers to integrate with the open "social graph" of personal relations and other things like songs, places, and Facebook pages. Allocation the on facebook.com, external websites, and devices are all allowed to access the graph. The Graph API is the core of Facebook Platform, enabling developers to read from and write data into Facebook. The Graph API presents a simple, consistent view of the Facebook social graph, uniformly representing objects in the graph (e.g., people, photos, events, and pages) and the connections between them (e.g., friend relationships, shared content, and photo tags). Access Key is required to use the facebook app. When you register, we will generate a unique App ID for your app. You will use this a lot, since it must be included when making any calls to our APIs. All of our SDKs provide a way for you to easily set this so it will automatically be included with any API calls.

Sentiment Analysis

Sentiments can drive transitory market changes which therefore causes a differentiation between the stock expense and the certifiable estimation of an association's offer. Over a huge extent, in any case, the measuring machine kicks in as the essentials of an association finally cause the value and market cost of its ideas to join. Inclinations are a significant bit of protection trades and separating suppositions reliant on various data sources can give bits of learning on how monetary trades react to different kinds of news in the brief and medium-term. In this way, a novel system—nostalgic assessment—has risen which checks the supposition from data sources or evaluation behind the news to recognize its impact on the business parts. Conclusion examination is another system that has generally been used for protections trade assessment (Bollen et al. 2011). It is the route toward anticipating stock examples by methods for modified assessment of substance corpora, for instance, news sources or tweets express to protections trades and open associations. The sentiment portrayal frameworks are generally isolated into AI approach and jargon based philosophy which is also parceled into dictionary-based or corpus-based strategies (Bhardwaj et al. 2015). Seng and Yang (2017) showed the capacity of using inclination signals from an unstructured book for improving the profitability of models for anticipating insecurity slants in the money related trade.

Table 2: Classes and Keywords

Classes	Keywords
High price	high, gain, target, dividend, bull, bullish, buy, invest, mature, positive, up, profit, going up, high volume, blue chip stocks, high margin
Low Price	Bear, bearish, low, down, stoploss, sell, disinvest, negative, results loss, volatile, low margin

Mathematical Model of The System

Let S represent from system as a set of components as follows:

S: { F, A, DB, I, O, C, R,G} where, I:- Input
O:- Output
R:- Reviews
S:- Stocks
C:- Comments
DB:-Database
A:-User
G:- Graphs

Input: -

I1:- set of input data Stock Data Storage {s1,s2,s3,s4…sn}
I2:- set of input data from Facebook {c1,c2,c3,c4…cn}
I3:- set of input from System {g1,g2,g3,g4…gn}

Output: -

O1:- set of output data from System {s1,s2,s3,..sn}
O2:- set of output data from Sentiment Module
{r1,r2,r3....rn}
O3:- set of output data from Database{d1,d2,d3.dn}

Functions: -

F1:- set of functions on the stock data
F2:- set of the functions of comments on Facebook post
F3:- set of the functions on results to be represent in graphs F4:- set of functions of comments classification.

Results and Discussions

The following parameters to be considered while comparing the propose system with existing system.
1. Accuracy of results: Accuracy is used to describe the closeness of a measurement to the true value. In our project we get the true value related to the Classes defined for the High Price and Low Price related keywords.
2. Efficiency: It is the performance of the system to accomplish something with the least waste of time and efforts. It is usually expressed in percentage. In our system the keyword mentioned in the comment for any stock is efficiently matched with the word dictionary.
3. Throughput: It is the work done by a system in a given period of time. The system work out the process of fetching the comments from the Face book pages in the given period of time.
4. Time Delay: It is a delay between two operations designed to make it safer or more efficient. The time delay may introduce while data fetching and processing by the system.

The system is used for prediction of the Stock Market by two ways, as
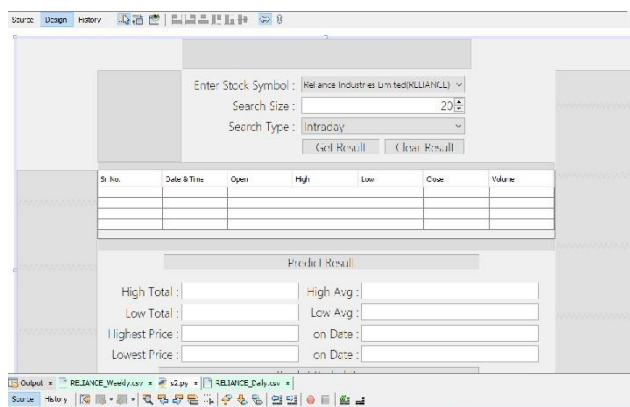1. By using Historical data

Fig. 2 Historical data Stock Market data Fetching, Classification and Prediction
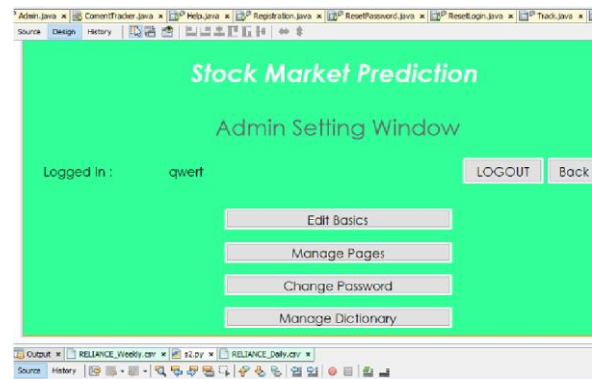
2. Facebook data

Fig. 3 Admin window for Stock Market Prediction

Conclusion

This paper condenses significant methods in ML which are applicable to stock expectation. The paper presents the utilization of direct relapse and strategic relapse for stock forecast and stock examination and this investigation prescribes Naive Bayesian to get precise outcomes. An imperative to this end is the need of the dataset utilized in the forecast to be grouping cordial. The paper outlines the apparatuses which can be utilized for the execution of ML calculations. Every one of the instruments bolsters relapses and grouping calculations, clients can pick any device dependent on their nature and accommodation. The paper proposes a framework to extricate learning from information and playing out an expectation to prompt the shopper for ventures as well as the sentiment analysis of the users from social media data.

Acknowledgement

References

[1] YefengRuan, ArjanDurresi, LinaAlfantoukh,'' Using Twitter trust network for stock market analysis, Elsevier, Volume 145, 1 April 2018, Pages 207-218
[2] Bhagwant Chauhan, Umesh Bidave, Sachin Kale "Stock Market Prediction Using Artificial Neural networks," International Journal of Computer Science and Information Technologies, vol. 5, 2014.

[3] Qasim A.Al-Radaideh, Adel Abu Assaf "predicting stock prices using data mining techniques," International Arab Conference on Information Technology, 2013.

[4] S. Madge, "Predicting Stock Price Direction using Support Vector machines," 2015.

[5] M. Kolambe, "Survey Paper on Stock Market Prediction," International Journal of Innovative Research in Computer and communication engineering, vol. 4, no. 10, 2016.

[6] Neha Khalfay. Deepali Vohra, Vidhi Soni, Nirbhey Singh Pahwa, "Stock Prediction using Machine Learning a Review Paper," International Journal of Computer Applications, vol. 163, no. 5, 2017.

[7] Mohamad A.Ahmad, Fadel M Megahed. Bin Weng, "Stock market one-day ahead movement prediction using disparate data sources," Expert Systems With Applications, pp. 153-163, 2017.

[8] Kartik Sharma, Akhilesh Rao, Sujay sail "STOCK MARKET ANALYSIS," in 6th international conference on latest innovations in science and engineering, new Delhi, 2017.

[9] Aditya Bhardwaj, Yogendra Narayan, the van ran "Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty," in 4thInternational Conference on Eco-friendly Computing and Communication Systems, Chandigarh, 2015.

[10] Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey, 2019, IEEE

[11] A Review on Machine Learning Algorithms, Tasks and Applications Diksha Sharma, Neeraj Kumar, (IJARCET) Volume 6, Issue 10, October 2017, ISSN: 2278 – 1323

[12] Stock Market Analysis: A Review and Taxonomy of Prediction Techniques Dev Shah, Haruna Isah * and Farhana Zulkernine, Int. J. Financial Stud. 2019, 7, 26; DOI:10.3390/ijfs7020026

[13]Machine learning: A review of classification and combining techniques, S. B. Kotsiantis · I. D. Zaharakis · P. E.Pinellas,https://www.researchgate.net/publication/226525180

[14] Stock Prediction using Machine Learning a Review Paper, Nirbhey Singh Pahwa,Neeha Khalfay, International Journal of Computer Applications (0975 – 8887) Volume 163 – No 5, April 2017

[15] Stock Market Prediction Using Machine Learning V Kranthi Sai Reddy, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 10 | Oct 2018