ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

Comparison of Historical Data and Tweets based Stock Market Prediction Algorithms

V. NARAHARI
Assistant Professor
Narahariv@gmail.com@gmail.com

M. BALACHANDRA
Assistant professor
Balachandra.1202@gmail.com

K. RANGASWAMY
Assistant Professor
Rangaswamy.alts@gmail.com

Abstract:

One of the issues in research that has attracted the attention of a variety of scholars is making accurate predictions about the stock market. It is presumed that the essential information that can be accessed at any time has some kind of predictive link to the future stock returns. Investors are provided with information by the current work so that they may improve their decision-making processes throughout the process of purchasing stocks. The past prices of stocks and tweets that remark on those prices are the elements that go into the decision-making process. The stock market state may be predicted with the use of the suggested technique by using the Linear Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF) approaches. The experimental findings came to the conclusion that the SVM-based prediction had much better predicting performance than the other approaches when those results were examined using standard datasets. The study that has been suggested provides a comparison of several criteria that will be considered when making the purchase.

Keywords: Stock Market prediction, Prediction Framework, Linear Regression, Naïve Bayes, SVM, Random forest classifier.

Introduction

The fundamental nature of the financial sphere, in addition to other characteristics such as the closing price of the previous day and the price-to-earnings ratio, make it difficult to make accurate predictions about the stock market. The development of several algorithms for the prediction has been the subject of a great deal of research and effort. These algorithms range from those that use data mining strategies to those that are based on machine learning. The

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

forecast of stock markets and individual stocks depends on a number of different aspects, some of which are physical, some of which are physiological, some of which are rational and some of which are illogical. As a result, one may get information about economics and finance. Therefore, many individuals and investors are connected to contemporary technology via the practise of stock market prediction. Since the decision-making process is seen as being very important, the fact that the Internet is the most important and major means of disseminating information to the public has a huge influence on the stock markets. In order to forecast future behaviours and patterns, several methods and techniques have been developed.

There are two approaches to predicting stock prices, and they are: 1. Technical Analysis 2. Fundamental Analysis

2. Fundamental Analysis

Technical analysis is a method that is used in the process of determining the price of a stock. This conclusion was reached after researching the background of the stock. In order to complete this method, time series analysis is used.

The conclusions made in fundamental analysis are based on the previous performance of the firm as well as its profit estimate. This relates to the firm, and real shares of stock are only taken into consideration to a limited degree.

However, using traditional methods to make a forecast does not guarantee that the prognosis will be accurate. As a result, the primary emphasis of this study is placed on the application of machine learning algorithms to the forecasting of the stock market. According to the findings of a number of research, it has been shown that the methods of machine learning have the potential to uncover patterns and insights that might be used to produce precise and accurate predictions.

Observations made using previously used methods

The following section discusses the several techniques of stock prediction that are currently available.

The research conducted by Qasem et al. [1] assists investors in the stock market in determining the optimal moment for purchasing or selling stocks, which is dependent on the information received from the study of the stock's historical data. For the purpose of making the choice, a

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

decision tree classifier was used. The Cross Industry Standard Process for Data Mining serves as the foundation for the model that has been developed (CRISP-DM).

Several different machine learning methods and their applications were addressed by Nirbhey Singh Pahwa et al. [2]. For the purposes of forecasting and assessing the stock, linear regression and logistic regression might be used, and it was proposed that support vector machines be employed to attain the desired results. In addition, the tools that were used for the execution of the machine learning algorithms were dissected and spoken about.

A test has been run in the prediction of the Karachi Stock Exchange (KSE), and the suggested method has been evaluated using data from the Saudi Stock Exchange for the TASI corporation. For the last half a year, data has been crawled from the KSE, and multiple machine learning classifiers have been put into use in order to provide projections about the future volume of these firms. The authors have developed the Ada-boost algorithm, as well as Bayesian network and Multilayer perceptron. It has been shown that, when compared to the other two classifiers, Ada-boost offers superior results for both KSA. This is the case when comparing Ada-boost to the other classifiers.

For the purpose of predicting daily stock values, historical data, technical indicators, and optimization of least squares support vector machines (LS-SVM) using the Particle Swarm Optimization (PSO) method are used. The Levenberg-Marquardt (LM) method is used with the LS-SVM and LS-SVM-PSO models [4] in order to analyse and compare the outcomes. The historical data and derived technical indicators are represented by six input vectors, and the resulting price is represented by a single output vector. The authors have presented a new machine learning method that combines the PSO with the LS-SVM in order to improve accuracy. Included in this package are a number of different indicators, including the money flow index, moving average convergence/divergence, exponential moving average, and stochastic oscillator. There has been some discussion on whether or not optimization of LS-SVM should be the global optimization method. With the assistance of the PSO method, LS-SVM free parameters such as C (cost penalty), (insensitive-loss function), and (kernel parameter) are selected. The over-fitting issue that arose in ANN may be solved with the help of the LS-SVM-PSO model that was presented. When compared with ANN-BP and a single instance of LS-SVM, the LS-SVM-PSO algorithm has the lowest error value possible.

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

Because it has such a significant bearing on the findings, it is necessary for the data that has been gathered for the subsequent procedure to be subjected to both pre-processing and post-processing [5]. The authors developed a model, which has been put into action, in order to lower the risk. Time series, neural networks, and hybrid approaches are all included into the model that the authors utilised. It has been demonstrated that the Recurrent Neural Network (RNN) performs better than the Artificial Neural Network (ANN) for prediction, and when comparing the Layered Recurrent Neural Network (LRNN) to the Feed-Forward Neural Network (NN), it has been observed that the LRNN requires fewer iterations but spends more time overall. The experimental results have shown this to be the case. The outcomes of the Feed Forward Neural network have been improved as a direct consequence of the data being preprocessed with the assistance of the WSMPCA algorithm.

A significant amount of effort has been put into the conception of prediction models, the majority of which have been on linear statistical models [6]. However, since variance is the driving force behind the movement of stocks and other assets, linear methods are suboptimal, and nonlinear models are more likely to generate a smaller forecast error. This is because variance is the underlying concept. In recent years, several studies have been conducted with the goal of developing methods of machine learning that can accurately anticipate stock prices. Implementing a Support Vector Machine in order to make a prediction as to whether the price of the specified stock will be low or high on a specific day is the purpose of the task. For the purpose of determining the daily closing prices for each stock from 2007 to 2014, the suggested model makes use of factors such as the momentum of the particular stock, the current price volatility, and the technology sector. The obtained historical data has been evaluated in order to provide a forecast on the future direction of prices. It has been shown that the suggested model may reach the prediction accuracies with certain parameters over a lengthy period of time.

Machine Learning methods like as LR, RF, and Multilayer Perceptron (MLP) [7] are used in the process of making a prediction of the data of the New York Times for the next ten years. MLP has been shown to be superior to the other two algorithms because, in a certain range, the fluctuation between the projected price and the actual price is negligible when compared with Logistic Regression and random forest. This has led to the conclusion that MLP is the superior method. The performance of prediction has revealed that Random Forest performs better than logistic regression, while MLP performs better than any other method.

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

The forecast of the stock market is dependent on a vast variety of characteristics, each of which plays an important part in the contribution that variations in supply and demand make [8]. In addition to the numerical analysis of stock trends, time series data and neural networks are trained to uncover and analyse the different patterns from recent trends. This is done in conjunction with the research of recent trends. The writers thought of doing a textual analysis of it by researching the general opinion of the public using internet news sources and blogs. In order to provide more accurate predictions about stock prices, a merged hybrid model has been implemented.

Nishanth et al. [9] presented three distinct algorithms for the purpose of conducting data analysis. These algorithms include the Recurrent Neural Network (RNN), the Long Short-Term Memory (LSTM), and the Gated Recurrent Unit (GRU). Through a process of self-learning, the hidden pattern and data dynamics are going to be uncovered, and that is the purpose of the task that has been presented. Experiments have been conducted on the data that was gathered from the banking and automotive industries as part of this endeavour. This study uses a method known as sliding window analysis with data overlap. It has been discovered that there is no direct control shared by the two parts of the economy. The LSTM model was used to produce the best possible results in an investigation of the interrelationships between the numerous businesses operating in the same market sector. In the beginning, the objective is to investigate and forecast the data based on a variety of trends and cycles since this has the potential to result in a profit for the investors.

E. Chong and colleagues [10] provided a comprehensive review and discussion of the limitations of using deep learning algorithms to analyse and forecast the stock market. In this work, high-frequency intraday stock returns were employed as input data, and the impacts of three different unsupervised feature extraction approaches, including principal component analysis, autoencoder, and limited Boltzmann machine, were investigated. The future behaviour of the market may be predicted with the use of these strategies. According to the findings, it was found that the extra information was collected from the remaining section of the autoregressive model, which led to an improvement in the prediction performance. The estimate of the covariance is made more accurate whenever the predictive network is used in conjunction with the covariance-based market structure study.

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

Problem Definition

The capability of DNNs to extract features from a huge quantity of raw data without requiring the user to have previous knowledge of predictors is one of the most significant benefits of these models. Because of this, deep learning is well suited for predicting the stock market, which is characterised by a high degree of complexity and nonlinearity due to the multitude of variables that influence stock prices. Instead of just dumping a massive raw dataset, if there are factors that have strong evidence that they may be predicted, it is possible that utilising such factors will yield better performance than dumping the dataset. However, these variables may also be used as part of the input data for deep learning, and we can then use deep learning to determine the link between the factors and stock prices.

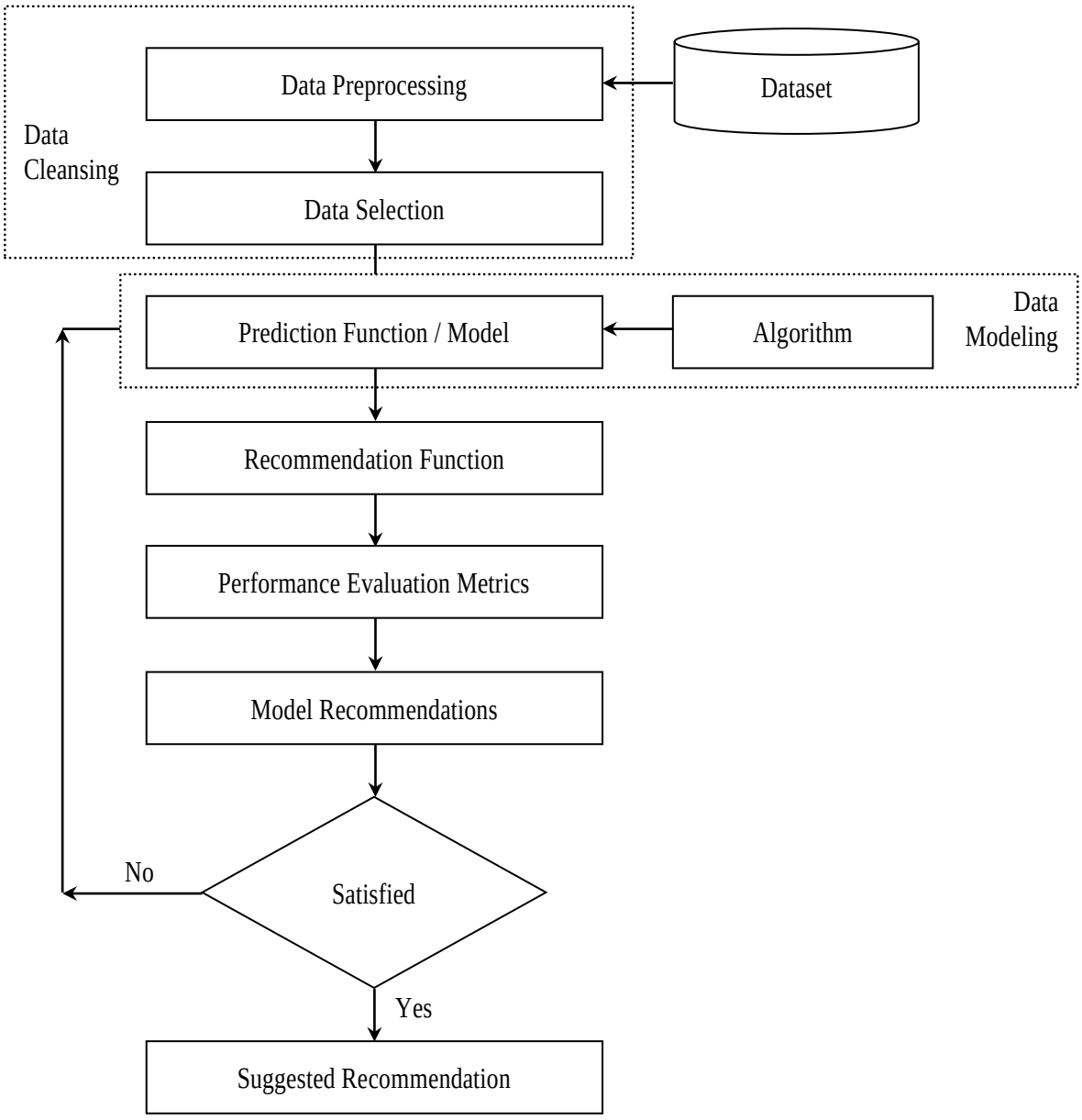
The research may be expanded by using methods that can track the erratic shifts that take place on the stock market. Evaluating the connection between different sectors is also possible, which gives us the opportunity to determine whether or not there are any hidden characteristics that will correlate the performance of different sectors that, at first look, seem to be independent of one another. The stock market may be predicted using the framework model that was provided, as illustrated in figure 1 below. This prediction is based on an analysis.

The framework that has been presented for the Stock Market Prediction (SMP) is shown in Figure 1. It is broken down into many stages, which are as follows: (1) Data Cleansing, (2) Data Modeling, (3) Recommendation Function, (4) Performance Evaluation, and (5) Suggestions and Recommendations.

In the first stage, known as "data cleaning," the preprocessing and feature selection of the dataset's data are the primary focuses. The first phase consists of carrying out the preprocessing step in order to convert the raw data into data that can be processed and that is of value. In the next stages, the data selection phase will deal with the selection of data pertaining to time stamps, such as trend, seasonality, stationarity, and cycles. Regarding SMP, the other factors that need to be considered are the opening stock price, the closing stock price, the highest stock price, and the lowest stock price, followed by the ordering of data based on the date because the date field plays an essential role in predicting the market value(s) in SMP. Other factors that need to be considered include opening stock price, closing stock price, highest stock price, and lowest stock price. In addition, the data visualisation assists in the process of sorting, viewing, and

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

researching the qualities of data over a period of time, which helps in the construction of a reliable prediction model.



ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

Figure 1 Proposed Framework for Stock Market Prediction

The prediction function or model phase deals with the process of condensing the enormous data into the human understandable and accessible form. In turn, based on the history of events, the forthcoming value might be predicted. The fundamental components of predictive models are data, statistical modeling and assumptions. Based on the requirements, the concern algorithm is chosen for the prediction model. The recommendation function consists of an engine, helps in learning the machine from the positive (accepted) and negative (not accepted) feedbacks. The categories of recommendation functions or engines are (1) Collaborative Filtering, (2) Content-based Filtering, and (3) Hybrid – Systems. Based on the requirements of the data to be recommended, the corresponding recommendation engine is adopted. Finally, the performance of the recommendation engine is evaluated using the following metrics.

(a) Recall
(b) Precision
(c) Root Mean Square Error (RMSE)
(d) Mean Reciprocal Rank (MRR)
(e) Mean Averaging Precision (MAP) at the cutoff
(f) Normalized Discounted Cumulative Gain (NDCG)

Based on the resultant metrics, if the recommendation is satisfactory, the concern suggestions are passed to the users; else those suggestions are submitted to the prediction function or model for better refinements. The performance of the prediction function varies based on the parameters of the adopted algorithm. The k-fold process helps in refining the performance of the recommendation system in the proposed framework. Prediction of the stock market with price history alone does not produce accurate predictions.

Import the required packages

Read the data

Cleaning the data (data preprocessing and data selection)

Removing the punctuations

Removing the column name for ease of access

Converting the headlines into lowercases

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

Applying the model (prediction function)

Logistic Regression model

Naïve Bayes

Random Forest Model

SVMGausiana

Performance Comparison

Figure 2 Pseudo-code for the proposed method

Data cleansing:

Tweets have been collected over a period for further analysis. It is recommended to use both the opinions of the public about the stock and also the reviews about the products and services offered by the company. The initial phase is to pre-process the data since the data thus collected might not be in an understandable format. Stock values might be missing in between the dates. Certain computations are performed to fill all the null values. Tweets posted by many users might consist of unnecessary data. Hence, it is mandatoryto process to preprocess the data in order to signify the public emotions. Preprocessing consists of three phases,namely tokenization, stopwords removal and matching regex for the removal of special characters.

- Tokenization: Individual words based on the space and extraneous symbols like special symbols, emoticons are extracted from the obtained tweets. A group of individual words is formed for each tweet.
- Stopword removal: Stopwordsare categorized as prepositions, articles, adverbs and conjunctions of the English language. These words could be removed from the group of words.
- Matching regex for the removal of special characters:URLs must be substituted by the term URL. Symbols like # and @ must be replaced properly. Intense emotions must be reinstated with proper words. The tweets are classified as positive and negative based on the views posted by the user.

Feature Extraction:

Co-occurring words within a specified window could be obtained through this N-gram representation. Tweets that are preprocessedis given as input in order to parse the related text anda word sequence of length ‘n’ is retrieved from the tweets so that a dictionary is constructed with a group of words and phrases. The tweets are split into bi-gram, tri-gram and N-gram for

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

further analysis. The features to the model are given in the form of a string of 1's and 0's in where 1 denotes the occurrence of then-gram of the tweet and a 0 denotes its absence. All the steps are carried out for the training set and a test set is used to perform classifications and efficiency of the classifier is shown. Classifiers like logistic regression, Support Vector Machine(SVM) Gaussian,Naïve Bayes (NB) and Random Forest are applied.

1. Logistic Regression[9]:

In this technique, one or more dependent variable is used to identify the outcome. A dichotomous variable is used for measuring the outcome. The dependent variable is binary or dichotomous. The relationship between the dependent and a group of the independent variable is described with the help of logistic regression. It could be known easily because the β parameters that best fit are determined and the same is denoted in the below eqn 1.

$$y = \begin{cases} 1, \text{ if } \beta_0 + \beta_1 x + \varepsilon, \\ 0 \text{ , otherwise} \end{cases}$$

The logistic function σ(t) is defined as follows:

$$\sigma(t) = \frac{e^t}{1+e^t}$$

Here 't' takes the numeric value and the above equation could be rewritten as

'y' is the predicted value.

2. **Random Forest**:

Decision trees can be used for various machine learning applications. Irregular patterns are learnt by applying the concept of trees. A slight variation makes the tree to behave differently. The main characteristic is that the decision trees have high variance and low bias. Data is partitioned recursively and when a particular node is reached, and then the split depends on the response given for the question for an attribute. Shannon Entropy or Gini impurity is used as the splitting criteria. The quality of the split in each node is measured with the Gini impurity and it is given as

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

$$g(N) = \sum_{i \neq j} p(w_i) p(w_j)$$

where P(ωi) is the proportion of the population with class label i.

The entropy in a node N can be calculated as follows

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \quad (1)$$

Here c refers to the number of classes considered and (ωi) is the proportion of the population labeled as i.

When all the classes are enclosed in equal part in the node, then entropy will be high. If the entropy is low, then there is only one class in the node. The impurity could be reduced by selecting the best splitting decision at a node. The highest gain in information is the principle behind the best split.

The information gain due to a split can be calculated as follows

$$\Delta I(N) = I(N) - PL * I(NL) - PR * I(NL)$$

where I(N) is the impurity measure (Gini or Shannon Entropy) of node N, PL is the proportion of the population in node N that goes to the left child of N after the split and similarly, PR is the proportion of the population in node N that goes to the right child after the split. NL and NR are the left and right child of N respectively.

3. **Naïve Bayes Classification**:

It is a probabilistic machine learning model and it is mainly used for classification. The principle behind this classification is the Bayes theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

A simplifying conditional independence assumption has been involved in this model. To this class, words that are conditionally independent of each other are given as input. The assumption made in this algorithm does not create an impact on accuracy. The Bayes theorem is stated as per the following equation.

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

$$\text{posterior probability} = \frac{\text{conditional probability} \cdot \text{prior probability}}{\text{evidence}} \quad (3)$$

The naïve bayes aim to maximize the posterior probability for the given training data in order to construct a decision rule for the new data. The naïve assumption to the Bayes theorem is about the independence between the features. Therefore the evidence is partitioned into the independent parts. Assume if there are two events which are considered to be independent, then it is represented as given below

P(A,B) = P(A)P(B)

Consequently, the result thus obtained is

P(y|x_1,...,x_n) = P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}

which can be expressed as:

P(y|x_1,...,x_n) = \frac{P(y)\prod_{i=1}^{n}P(x_i|y)}{P(x_1)P(x_2)...P(x_n)}

Now, as the denominator remains constant for a given input, we can remove that term:

P(y|x_1,...,x_n)\propto P(y)\prod_{i=1}^{n}P(x_i|y)

The probability of given set of inputs for all possible values of the class variable *y* has to be determined and choose the output with maximum probability. This can be expressed mathematically as:

So, finally, the task of calculating P(y) and $P(x_i \mid y)$ must be done. Here P(y) is also called class probability and $P(x_i \mid y)$ is called conditional probability.

The likelihood of the features is assumed to be Gaussian, hence, conditional probability is given by:

P(x_i | y) = \frac{1}{\sqrt{2\pi\sigma _{y}^{2} }} exp \left (-\frac{(x_i-\mu _{y})^2}{2\sigma _{y}^{2}} \right)

4.**Support Vector Machine**: The intention of using this algorithm is to bring out a perfect distinction that classifies the data points with a hyperplane in an N-dimensional space where N denotes the number of features.

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

dataset D is the set of n couples of the element (xi,yi)

$D=\{(xi,yi)|xi\in Rp,yi\in\{-1,1\}\}ni=1$

Given a hyperplaneH_0 separating the dataset and satisfying:

$$w\cdot x+b=0\ w\cdot x+b=0$$

Two others hyperplanes H_1 and H_2 are selected which also separate the data and have the following equations :

$$w\cdot x+b=\delta \quad \text{and} \quad w\cdot x+b=-\delta$$

so that H_0 is equidistant from H_1 and H_2.

However, here, the variable δ is not necessary. So we can set δ=1 to simplify the problem.

$$w\cdot x+b=1 \ \text{and} w\cdot x+b=-1$$

The following constraints must be met by the hyperplane and therefore,the corresponding plane that meets the constraint is selected.

For each vector x_i either :

$w\cdot x_i+b\geq 1$for $x_i$having the class1 or

$w\cdot x_i+b\leq -1$for $x_i$having the class1 -1

When the hyperplane does not meet the above constraints, then the above two constraints are merged into a single constraint as

$yi(w\cdot_{xi}+b)\geq 1$for all$1\leq i\leq n$

The distance between the two hyperplanes must be maximized

Among all possible hyperplanes meeting the constraints, the hyperplane with the smallest width is chosen

Minimize in (w,b)

$\|w\|$

subject to $y_i(w\cdot x_i+b)\geq 1$(for any i=1,…,n)

Experimental Results and Discussion

The present method is evaluated on Intel Pentium I3 Machine with 2 GB RAM on Python 3.6 platform. In this study, the data set description is as follows. The total number of samples is 4102 days from Jan 2007 to Jan 2016. The entire data set is partitioned into two parts, (year less than 2015- 80%) as training data and year greater than 2015- 20% as test data set. A separate learning model with logistic regression classifier, NB, SVM, RF have been constructed on training dataset

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

and evaluation is done on unigram,bigram and trigram model to extract the features. The model has been evaluated based on the accuracy and a comparison is shown in the below Table 1.

Models	Unigram	Bigram	Trigram
LR	0.822	0.857	0.851
SVM	0.851	0.851	0.825
NB	0.820		
RF	0.847	0.839	0.849

Table 1: Comparison features of classifiers.

The detailed information of the test data evaluation with the unigram model is shown in the below table 2.

Models	Positive/negative	Precision	Recall	F1-score	Support
LR	0	0.83	0.80	0.82	186
	1	0.81	0.84	0.83	192
SVM	0	1.00	0.70	0.82	186
	1	0.77	1.00	0.87	192
NB	0	0.81	0.83	0.82	186
	1	0.83	0.81	0.82	192
RF	0	0.92	0.75	0.83	186
	1	0.80	0.94	0.86	192

Conclusion

The prediction of the stock market becomes the goal of the investors and has attracted many types of research to do various research works. A detailed analysis has been done based on the models that have been developed. From the experimental results, it has been proven that the precision obtained through support vector machine model is better when compared with the other algorithms such as Linear Regression model, Naïve Bayes algorithm and Random Forest classifiers.

REFERENCES

ISSN: 2278-4632
Vol-9 Issue-11 No.01 November 2019

[1] Qasem a. Al-radaideh, Adel Abu Assaf, Eman Alnagi, "Predicting Stock Prices using Data Mining Techniques", The International Arab Conference on Information Technology (ACIT'2013)

[2] Nirbhey Singh Pahwa, NeehaKhalfay, VidhiSoni, DeepaliVora," Stock Prediction using Machine Learning a Review Paper" International Journal of Computer Applications (0975 – 8887) Volume 163 – No 5, April 2017.

[3] Mustansar Ali Ghazanfar, Saad Ali Alahmari, Yasmeen Fahad Aldhafiri, Anam Mustaqeem, Muazzam Maqsood, and Muhammad AwaisAzam, "Using Machine Learning Classifiers to Predict Stock Exchange Index", International Journal of Machine Learning and Computing, Vol. 7, No. 2, April 2017

[4] Osman Hegazy, Omar S. Soliman and Mustafa Abdul Salam, A Machine Learning Model for Stock Market Prediction, International Journal of Computer Science and Telecommunications Volume 4, Issue 12, December 2013.

[5] Zahid Iqbal, R. Ilyas, W. Shahzad, Z. Mahmood and J. Anjum, "Efficient Machine Learning Techniques for Stock Market Prediction", Int. Journal of Engineering Research and Applications, Vol. 3, Issue 6, Nov-Dec 2013, pp.855-867

[6] Saahil Madge Predicting Stock Price Direction using Support Vector Machines

[7] Shubham Jain, Mark Kain, "Prediction for Stock Marketing Using Machine Learning", International Journal on Recent and Innovation Trends in Computing and Communication Volume: 6 Issue: 4

[8] Robert Chun, Thomas Austin, "STOCK PRICE PREDICTION USING DEEP LEARNING", https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1639&context=etd_projects

[9] Nishanth C P , Dr. V K Gopal , Vinayakumar R , Lakshmi Nambiar , Dileep G Menon, "Predicting Market Prices Using Deep Learning Techniques", International Journal of Pure and Applied Mathematics Volume 118 No. 20 2018, 217-223.

[10] EunsukChong ,Chulwoo Han , and Frank C. Park," Deep Learning Networks for Stock Market Analysis and Prediction: Methodology, Data Representations, and Case Studies" Article in Expert Systems with Applications · April 2017.