Comparative analysis of Stock Market Prediction Algorithms based on Twitter Data

R. Sai Venkata Ramana
Assistant Professor
Annamacharya Institute of Technology and Sciences (Autonomous),
Tirupati

M. Reddy Durga Sree
Assistant Professor
Annamacharya Institute of Technology and Sciences (Autonomous),
Tirupati

Ramakrishna Gandi
Assistant Professor
Annamacharya Institute of Technology and Sciences (Autonomous),
Tirupati

A. Sankar Reddy
Assistant Professor
Annamacharya Institute of Technology and Sciences (Autonomous),
Tirupati

ABSTRACT

Stock market prediction is considered as one of the most promising research area that is attainning the attention of various researchers. The vital information which is available for access is assumed to have predictive relationships to the future stock returns. The present work gives information to the investors so that the decision could be made better during the purchase of stocks. The factors that contribute towards the decision are the historical prices of stocks and tweet comments regarding the same. The proposed method uses four methods for predicting the stock market status, namely, Linear Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), and Random Forest (RF) approaches. When evaluated with standard datasets, experimental results concluded that the SVM based prediction has significant predicting performance than the other methods. The proposed work gives a comparison of factors in order to decide the purchase.

General Terms

Classification, Prediction, Stock Market Forecasting

Keywords

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

1. INTRODUCTION

Prediction of the stock market is a challenging task due to the fundamental nature of the financial domain and other parameters such as the closing price of the previous day, the P/E ratio. Many works have been done in devising various algorithms for the prediction that varies from applying data mining techniques to machine learning algorithms. Various factors such as physical, physiological, rational and irrational factors play a vital role in the prediction of stock markets and individual stocks and thus, one gains the knowledge both economically and financially. Prediction of the stock market thus connects several people and investors with modern technologies. Internet being the prime and primary source of providing information to the public has a significant impact on the stock markets since the decision-making process is found to be critical. Techniques and tools have been devised to predict future trends and behaviors.

There are two stock prediction methodologies and they are

1. Technical Analysis
2. Fundamental Analysis

Technical analysis is a technique which is used to determine the stock price. This is based on the history of the stock. Time series analysis is used in this technique.

In fundamental analysis, decisions are based on the company's past performance, earnings forecast. This deals with the company and the actual stock is considered only up to some extent.

However, deploying conventional strategies in prediction does not assure the reliability of the prediction. Hence this work focuses on implementing machine learning algorithms in the prediction of the stock market. From various studies, it is observed that the machine learning techniques have the prospectus of discovering the patterns and insight which could be utilized to make precise, accurate predictions.

2. FINDINGS FROM THE EXISTING APPROACHES

The following section deals with the existing methods of stock prediction approaches.

Qasem et al. [1] work aid the investors in the stock market to finalize the best time for buying or selling stocks and this depend on the knowledge obtained from the previous stock history. Decision tree classifier is used for taking the decision. The proposed model is based on the Cross Industry Standard Process for Data Mining (CRISP-DM).

Nirbhey Singh Pahwa et al. [2] discussed various machine learning algorithms for applications. Linear regression and logistic regression could be used for predicting and analyzing the stock and suggested to use Support Vector Machines to achieve results. Also, the tools that were used for the implementation of machine learning algorithms were discussed.

A trial has been done in the prediction of Karachi Stock Exchange (KSE) [3] and the proposed algorithm has been tested on Saudi Stock dataset for TASI company. Data has been crawled from KSE for six months and several machine learning classifiers have been deployed to forecast the future volume of these companies. The authors have implemented Ada-boost, Multilayer perceptronand Bayesian network. Among the classifiers, it has been proven that Ada- boost provides better results for both KSA when compared with the other two classifiers

Historical data, technical indicators and optimization of Least squares support vector machines (LS- SVM) with Particle Swarm Optimization (PSO) algorithm is used for prediction of daily stock prices. To compare the results, the Levenberg-Marquardt (LM) algorithm is used with LS-SVM and LS-SVM-PSO models [4]. Six input vectors denote the historical data and derived technical indicators and one output denotes the next price. A machine learning algorithm has been proposed by the authors that combine the PSO and LS-SVM. Various indicators such as money flow index, moving average convergence/divergence, exponential moving average and stochastic oscillator are included. Optimization of LS-SVM has been considered as the global optimization algorithm. Also LS-SVM free parameters such as C (cost penalty), ϵ (insensitive-loss function) and γ (kernel parameter) are chosen with the help of the PSO algorithm. The proposed LS-SVM-PSO model overcomes the over- fitting problem that occurred in ANN. The lowest error value is achieved through the LS-SVM-PSO when compared with single LS-SVM and ANN-BP.

Data that has been collected for the further process must undergo pre-processing and post-processing since it produces a high impact on the results [5]. The authors worked a model which is implemented to reduce the risk. The model used by the authors' implements Time series, Neural networks and hybrid techniques. From the experimental results, it has been proven that the Recurrent Neural Network (RNN) performs better than Artificial Neural Network (ANN) for prediction and when comparing Layered Recurrent Neural Network (LRNN) with Feed-Forward Neural Network (NN), it is observed that LRNN takes less iteration but consumes more time. Data preprocessed with the help of WSMPCA algorithm and due to this, results of Feed Forward Neural network have been enhanced.

Many research works have been done on devising prediction models that have been focused on linear statistical models [6]. However, the variance being the fundamental principle behind the stocks movement and other assets make linear techniques suboptimal and also nonlinear models likely to produce a lower predictive error. Nowadays, researches have been focused on the formulation of machine learning techniques for stock price prediction. The objective of the work is to implement a Support Vector Machine to forecast if the given stock's price is less or high on a particular day. The proposed model takes parameters, namely the momentum of the individual stock, current price volatility and the technology sector for computing the daily closing prices for each stock from the years 2007 to 2014. Historical data thus collected has been analyzed for the prediction of price direction. The proposed model is proved to obtain the prediction accuracies with specific parameters in the long- term.

Prediction of data of NY Times of 10 years is implemented with the help of Machine Learning algorithms like LR, RF and Multilayer Perceptron (MLP) [7]. MLP is proved to be better than the other two algorithms since, in a particular range, the variation between the predicted price and the actual price is minimal when compared with Logistic Regression and random forest. Performance of prediction has shown better results in Random Forest rather than logistic regression but inferior to MLP.

Stock market prediction depends on a large number of attributes that plays a vital role in the contribution of the changes in the supply and demand [8]. Time series data and neural network are trained to discover and study the various patterns from recent trends in addition to the numerical analysis of the stock trends. The authors considered the textual analysis of it by investigating the public sentiment from online news sources and blogs. A merged hybrid model is deployed, which is used for stock prediction more accurately.

Nishanth et al. [9] proposed three different algorithms, such as Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) for data analysis. The objective of the proposed work is to determine the hidden pattern and data dynamics through the self-learning process. The work has experimented on the data that has been retrieved from the automobile and bank sectors. This work adopts the sliding window approach with data overlap. It is found that there is no direct control between the two sectors. The interrelation between the various companies in the same sector is analyzed and the best result has been achieved through the LSTM model. At the outset, the aim is to examine and predict the data based on different trends and cycles since it might lead to the gain for the investors.

E Chong et al. [10] presented a detailed analysis of the drawbacks of deep learning algorithms for the stock market analysis and prediction. This study has used high-frequency intraday stock returns as input data and the effects of three unsupervised feature extraction methods such as principal component analysis, autoencoder and the restricted Boltzmann machine. These methods are used to forecast the future of market behavior. From the results, it has been observed that the additional information has been retrieved from the remnant part of the autoregressive model and thus the prediction performance has been improved. When the predictive network is implemented on the covariance-based market structure analysis, covariance estimation is improved.

3. PROBLEM DEFINITION

The ability to derive characteristics from a large number of raw data without depending on previous knowledge of predictors is one of the key benefits of DNNs. This makes profound knowledge especially ideal for stock market forecasting where multiple variables in a dynamic, non-linear way influence stock prices. If there are variables with good predictability proof, it may be possible to manipulate them more efficiently than merely dumping a huge raw dataset. However, these variables can also be used as an input to deep learning and deep learning can define the association between the factors and market prices.

The research can be generalized to capture the irregular shifts on the financial exchange by means of techniques. The relationship between different sectors can also be measured, helping us to decide whether there are secret parameters which correlate the output of different sectors which are first glance independent. The proposed structure model, as seen in Figure 1 below, offers an overviewof the stock market.

Figure 1 shows the proposed Stock Market Prediction (SMP) framework. It consists of phases namely, (1) Data Cleansing, (2) Data Modeling, (3) Recommendation Function, (4) Performance Evaluation, and (5) Suggestions and Recommendations.

The first phase, data cleansing, deals with preprocessing and feature selection of data from the dataset. To transform the raw data into processable and useful data, the preprocessing phase is carried out in the first phase. Later the data selection part deals with the selection of data concerning time stamps, including, trend, seasonality, stationality and cycles. Concerning SMP, the other factors to be considered are opening stock price, closing stock price, highest stock price, and lowest stock price, followed by the ordering of data based on the date because date field plays a vital role in predicting the market value(s) in SMP. Further,the data visualization also helps in sort, view and studies the properties of data over the period, which helps to build a robust prediction model.

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

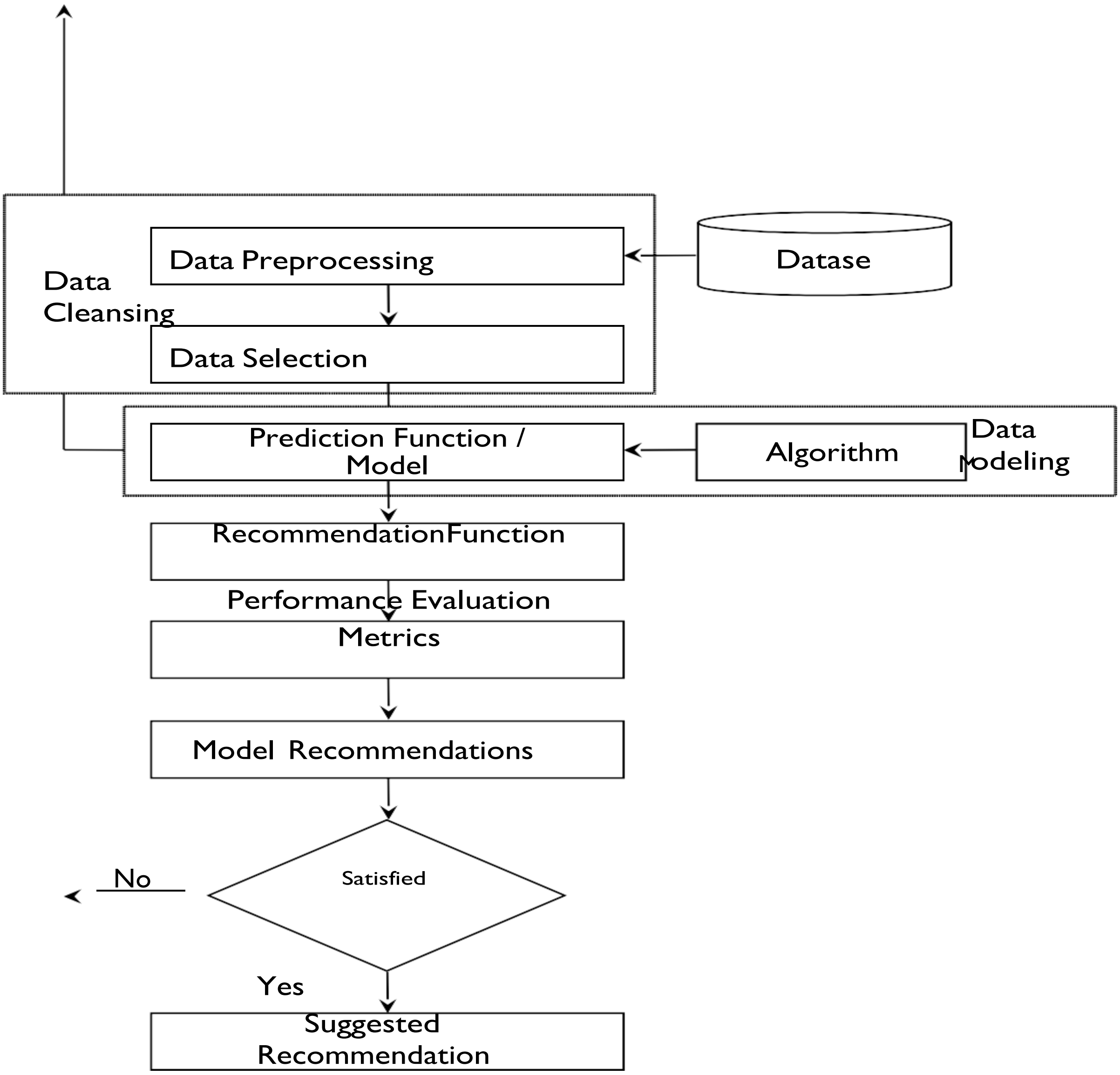

Fig 1: Proposed Framework for Stock Market Prediction

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
packagesRead 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
Applying the model (prediction
        function)Logistic Regression
        model Naïve Bayes
        Random Forest
        Model
        SVMGausiana
    Performance Comparison
```

Fig 2: Pseudo-code for the proposed method

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.

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

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

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

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 mandatory to 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: Stopwords are 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 preprocessed is given as input in order to parse the related text and a 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 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 the n-gram of the tweet and a 0 denotes its

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

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

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

The entropy in a node N can be calculated as follows

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

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.

4. EXPERIMENT 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 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.The heading of a section should be in Times New Roman 12- point bold in all-capitals flush left with an additional 6-points of white space above the section head. Sections and subsequent sub- sections should be numbered and flush left. For a section head and a subsection head together (such as Section 3 and subsection 3.1), use no additional space above the subsection head.

Table 1. Comparison features of classifiers

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 2. Detailed information of the test data evaluation with the unigram model

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

5. 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.

6. REFERENCES

[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, Neeha Khalfay, Vidhi Soni, Deepali Vora,” 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 Awais Azam, “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 PREDICTIONUSING 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] Eunsuk Chong , 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.