International Journal of Computer Engineering & Technology (IJCET)
Volume 10, Issue 3, May-June 2019, pp. 20-30, Article ID: IJCET_10_03_003
Available online at
http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=10&IType=3
Journal Impact Factor (2019): 10.5167 (Calculated by GISI) www.jifactor.com
ISSN Print: 0976-6367 and ISSN Online: 0976–6375

STOCK MARKET PREDICTION USING MACHINE LEARNING METHODS

Subhadra Kompella
Department of Computer Science and Engineering, GITAM (Deemed to be University), Visakhapatnam, India.

Kalyana Chakravarthy Chilukuri
Department of Computer Science and Engineering, MVGR College of Engineering (A), Vizianagaram, India

ABSTRACT

Stock price forecasting is a popular and important topic in financial and academic studies. Share market is an volatile place for predicting since there are no significant rules to estimate or predict the price of a share in the share market. Many methods like technical analysis, fundamental analysis, time series analysis and statistical analysis etc. are used to predict the price in tie share market but none of these methods are proved as a consistently acceptable prediction tool. In this paper, we implemented a Random Forest approach to predict stock market prices. Random Forests are very effectively implemented in forecasting stock prices, returns, and stock modeling. We outline the design of the Random Forest with its salient features and customizable parameters. We focus on a certain group of parameters with a relatively significant impact on the share price of a company. With the help of sentiment analysis, we found the polarity score of the new article and that helped in forecasting accurate result. Although share market can never be predicted with hundred per-cent accuracy due to its vague domain, this paper aims at proving the efficiency of Random forest for forecasting the stock prices.

Keywords: random forest, prediction, time series analysis.

Cite this Article: Subhadra Kompella and Kalyana Chakravarthy Chilukuri, Stock Market Prediction Using Machine Learning Methods, *International Journal of Computer Engineering and Technology*, 10(3), 2019, pp. 20-30.
http://www.iaeme.com/IJCET/issues.asp?JType=IJCET&VType=10&IType=3

1. INTRODUCTION

Stock market prediction has been an area of interest for investors as well as researchers for many years due to its volatile, complex and regular changing nature, making it difficult for reliable predictions. Stock market prediction is the act of trying to determine the future value of company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield a significant profit. The efficient-market

hypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable. Others disagree and those with this view point possess several methods and technologies which supposedly allow them to gain future price information.

Predicting how the stock market will perform is one of the most difficult things to do. Intrinsic volatility in the stock market across the globe makes the task of prediction challenging. There are so many factors involved in the prediction – physical factors vs. technical, rational and irrational behavior, etc. All these aspects combine to make share prices volatile and very difficult to predict with a high degree of accuracy. Using features like the latest announcements about an organization, their quarterly revenue results, etc., machine learning techniques have the potential to unearth patterns and insights we didn't see before, and these can be used to make unerringly accurate predictions.

In this paper, we considered historical data about the stock prices of a publicly listed company to implement machine learning algorithms in predicting the future stock price of a company, starting with simple algorithms like averaging and linear regression.

Forecasting and diffusion modeling, although effective can't be the panacea to the diverse range of problems encountered in prediction, short-term or otherwise. Market risk, strongly correlated with forecasting errors, needs to be minimized to ensure minimal risk in investment. Stock Market prices can be predicted based on two ways: Current prices of the stocks and both the prices and the news headings.

Current prices of the stocks – Generally prices vary from day to day on a fixed amount or at a constant rate. These are the type of general mutual funds where amount if invested, will be compounded manually. This is not of specific interest as there is nothing use of a machine to guess the future price. Just a calculator is enough.

Both the prices and the news headings – The prices subjected to these will change from time to time as based on their actions. Suppose a company launched a product which hit the market and got very connected to people. Obviously, sales will be increasing for that company and the investor who invested in that particular company will be profitable. For these type of calculations, we need some tool to be effective. These predictions are performed using several traditional methods such as Traditional Time Series, Technical Analysis Methods, Machine Learning Methods, and Fundamental Analysis Methods. The selection of the above methods is based on the kind of tool used and the data upon which the tool is implemented.

Technical Analysis Methods: Method of guessing the correct time to purchase stock pricing. The reason behind technical analysis is that share prices move in developments uttered by the repetitively altering qualities of investors in answer to different forces. The technical data such as price, volume, peak and bottom prices per trade-off period is used for graphic representation to forecast future stock activities.

Fundamental Analysis Techniques: This practice uses the theory of the firm foundation for preferred-stock selection. Data of fundamental analysis can be used by forecasters for using this tech of prediction for having a fully clear idea about the market or for investment. The growth, the bonus payout, the IR, the risk of investing so on are the standards that will be used to get the real value for an asset in which they could finance in the market. The main target of this process is to determine the inherent value of strength.

Traditional Time Series Prediction: Past data is used here and it uses this data to find coming values for the time series as a linear grouping. Use of Regression depictions has been used for forecasting stock market time series. Two rudimentary types of time series are simple and multivariate regressions.

Machine Learning Methods: These types of methods use samples of data that is needed for creating hope for the underlying function that had produced all of the other data. Taking out a deduction from different samples which are given to the model is the main aim for this. The Nearest Neighbor and the Neural Networks Practices have been used for forecasting of the market. Random forest type of models is being used as this model is involved in fields where we deal with risk.

Sentiment Analysis: Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It is also known as opinion mining, deriving the opinion or attitude of a speaker. Some domains in which sentiment analysis is used are-

- Business: In marketing field, companies use it to develop their strategies, to understand customers' feelings towards products or brand, how people respond to their campaigns or product launches and why consumers don't buy some products.
- Politics: In the political field, it is used to keep track of political view, to detect consistency and inconsistency between statements and actions at the government level. It can be used to predict election results as well!
- Public Actions: Sentiment analysis also is used to monitor and analyze social phenomena, for the spotting of potentially dangerous situations and determining the general mood of the blogosphere.

Sometimes known as opinion mining, sentiment analysis is the process of contextually mining text to identify and categorize the subjective opinions expressed by the writers. Normally it is used to determine whether the writer's attitude towards a particular topic or product, etc. is positive, negative, or neutral. It is also often used to help them understand the social sentiment of their brand, product or services while monitoring online conversations. In the context of a Twitter sentiment analysis, at its simplest, sentiment analysis quantifies the mood of a tweet or comment by counting the number of positive and negative words. By subtracting the negative from the positive, the sentiment score is generated. The process of reducing an opinion to a number is bound to have a level of error. For example, sentiment analysis struggles with sarcasm. But when the alternative is trawling through thousands of comments, the trade-off becomes easy to make. A little sentiment analysis can get you a long way when you're looking to gauge overall Twitter sentiment on a topic. This is especially true when you compare the sentiment scores with other data that accompanies the text.

Regression: Regression is a way of describing numerical relationship between a variable to predictor variables that is the outcome. The dependent variable is also referred to as Y which is plotted on the vertical axis (ordinate) of a graph. The predictor variable(s) is (are) also referred as independent prognostic or explanatory variable denoted by X. The horizontal axis (abscissa) of a graph is used for plotting X.

Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent (target) and independent variable (predictor). This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. For example, the relationship between rash driving and the number of road accidents by a driver is best studied through regression.

Regression analysis is an important tool for modeling and analyzing data. Here, we fit a curve/ line to the data points, in such a manner that the differences between the distances of data points from the curve or line are minimized.

If we want to estimate growth in sales of a company based on current economic conditions having the recent company data which indicates that the growth in sales is around two and a

half times the growth in the economy, regression analysis is extremely useful. Using this insight, we can predict future sales of the company based on current & past information.

There are multiple benefits of using regression analysis. They are as follows:

- It indicates the significant relationships between the dependent variable and independent variable.
- It indicates the strength of the impact of multiple independent variables on a dependent variable.
- There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of the regression line). Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represent by the following equation.

odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

ln(odds) = ln(p/(1-p))

$$\text{logit}(p) = \ln(p/(1-p)) = b_0+b_1X_1+b_2X_2+b_3X_3. \ldots\ldots\ldots\ldots\ldots\ldots +b_kX_k$$

where p is the probability of the presence of the characteristic of interest.

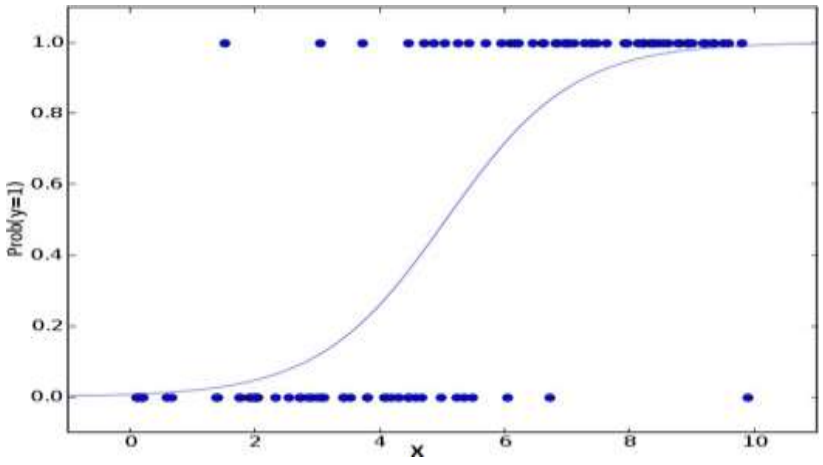


Figure 1 Logistic Regression Example

2. REGRESSION METRICS

Variance score: Variance is a measure of difference between the observed values to that of the average of predicted values, i.e., their difference from the predicted value means.

Mean absolute error (MAE): The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures *accuracy* for continuous variables. The equation is given in the library references. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

Root mean squared error (RMSE): The RMSE is a quadratic scoring rule which measures the average magnitude of the error. The equation for the RMSE is given in both of the references. Expressing the formula in words, the difference between forecast and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable.

The MAE and the RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater the difference between them, the greater the *variance* in the individual errors in the sample. If the RMSE=MAE, then all the errors are of the same magnitude

Both the MAE and RMSE can range from 0 to ∞. They are negatively-oriented scores: Lower values are better.

Mean square error (MSE): It is the average of the square of the errors. The larger value of mean implies larger error. The error, in this case, means the difference between the observed values y_1, y_2, y_3, … and the predicted one's pred(y_1), pred(y_2), pred(y_3), … We square each difference (pred(Y_n) – Y_n)) ** 2, so that negative and positive values do not cancel each other out.

3. RELATED WORK

[1] proposes an approach towards the prediction of stock market trends using machine learning models like the Random Forest model and Support Vector Machine. The Random Forest model is an ensemble learning method that has been an exceedingly successful model for classification and regression. Support vector machine is a machine learning model for classification.

[2] studies stock market prediction on the basis of sentiments of Twitter feeds which was experimented on the S&P 100 index. A continuous Dirichlet Process Mixture model was used to learn the daily topic set. Stock index and Twitter sentiment time-series were then regressed to make a prediction.

[3] analyzes data retrieved from Twitter issued to predict the public mood. A Self Organizing Fuzzy Neural Network is used on predicted mood from the Twitter feeds and Dow Jones Industrial Average values from the previous day to predict the movement of the stock market.

[4] uses the data from financial news articles to predict short-term movement of stock price. The movement of the stock price is classified three different classes representing three different directions ,namely "up ","down" , and "unchanged". A naive Bayesian text classifier is used to predict the direction of the movement of the stock price by deriving a set of indicators from the textual data retrieved from various financial news articles.

[5] presented an overview of artificial neural networks modeling process in predicting stock market price. This paper also discussed the problems encountered in implementing neural networks for prediction the future trends of stock market.

[6] developed a framework for power system short term load forecasting using feature selection model. Along with this the authors also used SVM to forecast load for simple and nonlinear loads.

[7] This paper evaluates the effectiveness of neural network models which are known to be dynamic and effective in stock-market predictions. The models analyzed are artificial neural network (ANN) trained with gradient descent (GD) technique, ANN trained with genetic algorithm (GA) and functional link neural network (FLANN) trained with GA. Experimental results and analysis has been presented to show the performance of different models.

[8] The authors in this paper presented a four layered model involving fuzzy multiagent system architecture to develop an artificial intelligent model to perform tasks like data preprocessing and stock market prediction.

[9] This paper developed a two stage neural network by combining Support vector machines and Empirical Mode Decomposition for predicting stock market. Experimental results proved that the combined model shows better prediction when compared to simple SVM.

4. METHODOLOGY

The implementation of this paper begins with preprocessing the data collected from stock market pickled data set. This preprocessed data is classified using popular machine learning algorithm to calculate the polarity score. In order to prepare the data ready to apply Random forest algorithm, noise in the data is removed by smoothing. The working of random forest algorithm is presented below:

i. Randomly select "k" features from total "m" features, where k << m

ii. Among the "k" features, calculate the node "d" using the best split point.

iii. Split the node into daughter nodes using the best split.

iv. Repeat 1 to 3 steps until the "l" number of nodes has been reached.

v. Build forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees.

Random forest algorithm starts by randomly selecting *k* features from *m* available features. Over the k selected features a point d has to be selected in order to split the features. This process would be executed iteratively to obtain the tree structure with a root node and leaf nodes as the target features to be processed further. This results in *n* number of trees in the generated forest. The algorithm is now tested for its efficiency by measuring the accuracy of predicting the stock price and also by calculating the variance score generated by the algorithm, finally ending up the process by comparing random forest with logistic regression. The experimental results obtained prove that Random Forest algorithm is efficient in predicting the stock price through achieving better score of the regression metrics over logistic regression. The results obtained are plotted in the form of a graph as presented below:

All the calculations are done based upon the four regression values variance score, mean absolute error, mean squared error, mean squared log error.

Sample Python code for variance score:

```
def
   score(y_tru
   e, y_pred):
   import
   numpy as
   np


   y_diff_avg = np.average(np.array([y_true]) - np.array([y_pred]))

   numerator = np.average((np.array([y_true]) - np.array([y_pred]) - y_diff_avg) ** 2)


   y_true_avg = np.average(y_true, axis=0)

   denominator = np.average((y_true - y_true_avg) ** 2)

    nonzero_numerator =
   numerator != 0
   nonzero_denominator =
   denominator != 0
```

valid_score = nonzero_numerator &

nonzero_denominator output_scores =

np.ones(y_true)

output_scores[valid_score] = 1 - (numerator[valid_score] / denominator[valid_score]) output_scores[nonzero_numerator & ~nonzero_denominator] = 0

return np.average(output_scores)

Sample Python Code for SMP using RandomForest

```
                def _set_oob_score(self, X, y):
 """Compute out-of-bag score"""
 X = check_array(X, dtype=DTYPE, accept_sparse='csr')


 n_classes_ =
    self.n_classes_
    n_samples = y.shape[0]


 oob_decision_function
    = [] oob_score = 0.0
    predictions = []


for k in range(self.n_outputs_):
    predictions.append(np.zeros((n_samples,
    n_classes_[k])))


 for estimator in self.estimators_:
unsampled_indices =
    _generate_unsampled_indices(
    estimator.random_state, n_samples)
 p_estimator = estimator.predict_proba(X[unsampled_indices, :],
 check_input=False)


if self.n_outputs_ == 1:
    p_estimator = [p_estimator]
```

```
for k in range(self.n_outputs_):
    predictions[k][unsampled_indices, :] +=
    p_estimator[k]


 for k in range(self.n_outputs_):

if (predictions[k].sum(axis=1) == 0).any():
    warn("Some inputs do not have OOB scores.
    ")


 decision = (predictions[k] /

              predictions[k].sum(axis=1)[:,
    np.newaxis])
    oob_decision_function.append(decision)

 oob_score += np.mean(y[:, k] ==

 np.argmax(predictions[k], axis=1), axis=0)


 if self.n_outputs_ == 1:

 self.oob_decision_function_ = oob_decision_function[0]

 else:

       self.oob_decision_function_ =

    oob_decision_function self.oob_score_ =

    oob_score / self.n_outputs_
```

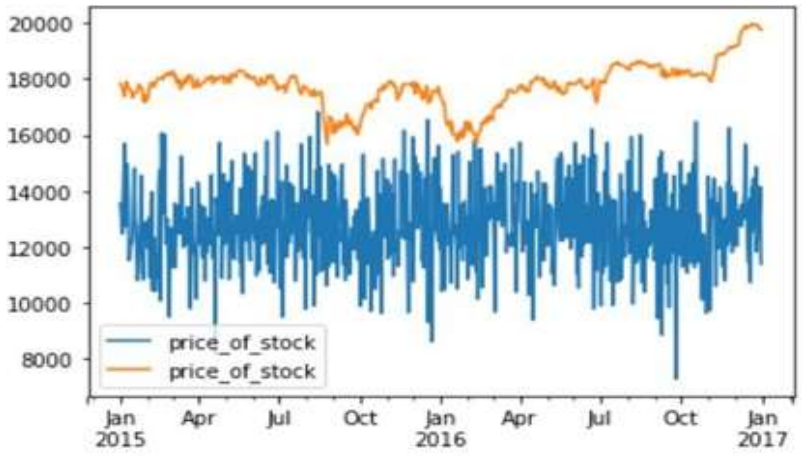
Regression metrics recorded for logistic regression:

EVS : -3.2762106267273907

Mean absolute error : 1595.0595446425434 Mean squared error : 3823664.4016104033 Mean squared log error: 0.013768419455969015

Regression metrics recorded For Random Forest:

EVS : -0.40059426107559504
Mean absolute error : 1139.3280327868883 Mean squared error : 1679163.3107885325



Mean squared log error : 0.004777722468450043

Figure 1: Random Forest without Smoothing

Figure 2: Random Forest without Smooting Labelled

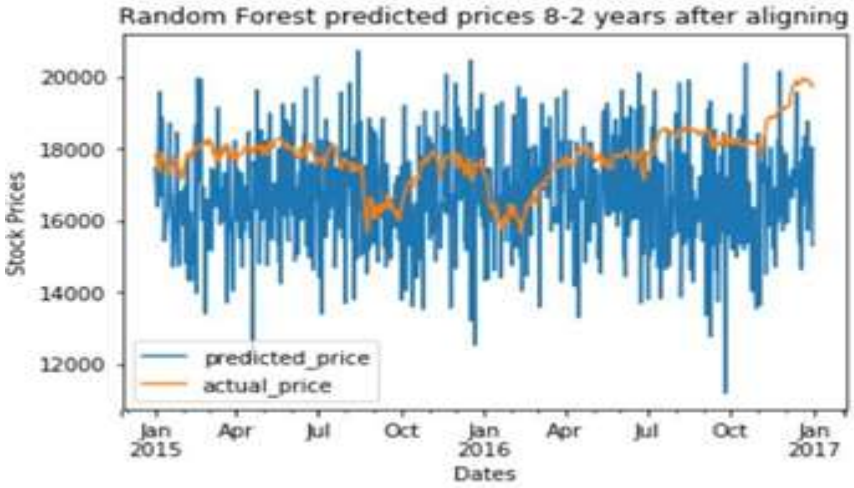


Figure 3: Random Forest after Aligning.

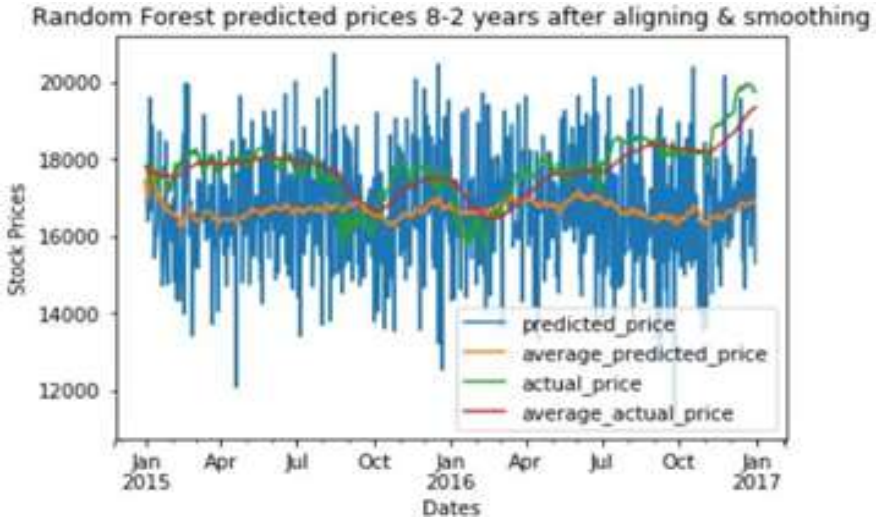


Figure 4 Random Forest with Aligning/Smoothing

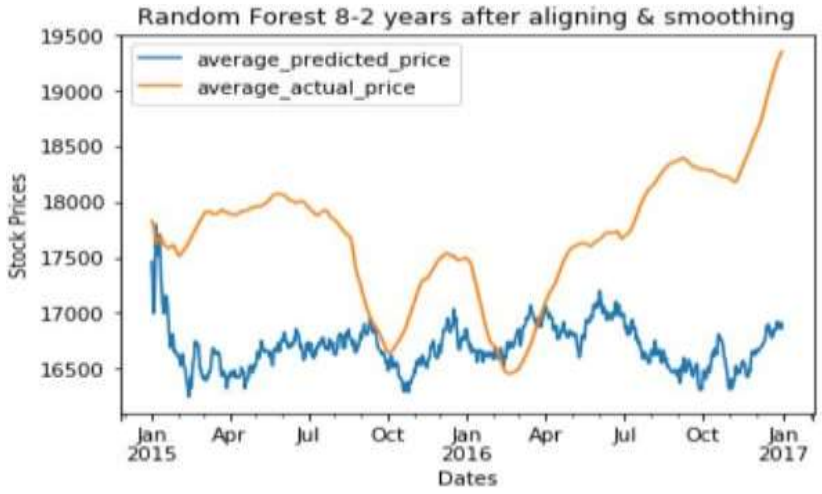


Figure 5 Prices Comparison of Original and Predicted Prices without Applying Algorithm

It is very much evident from the above graphs that for all regression metrics Random Forest algorithm produced better results when compared to Logistic Regression method.

5. CONCLUSION

In this paper, we predicted stock market which takes input as the price of stock and news heading. We used sentiment analysis to calculate the polarity score and then use it further in detecting the type of article has a positive or negative impact towards the stock and those can be used further in the analysis. The obtained scores are used to calculate the prices of stock and to complete those inputs as a set we used the exponential moving average method and saved as the impact of stock is correctly determined. The data after calculating is updated and displayed to the user as a graph. Finally we applied random forest algorithm and compared with logistic regression for efficiency. Variance score of Random forest is better than that of logistic regression. Mean absolute score of Random forest is better than that of logistic regression. Mean squared score of Random forest is better than that of logistic regression. Mean squared log error score of Random forest is better than that of logistic regression. In all, it can be concluded that the random forest algorithm is much efficient compared to logistic regression for the stock market prediction based on sentiment analysis.

REFERENCES

[1] Sahaj Singh Maini,Govinda.K, " Stock Market Prediction using Data Mining Techniques," IEEE International Conference on Intelligent Sustainable Systems, 2017.

[2] Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., &Deng, X, Stock Market Prediction On The Basis Of Sentiments Of TwitterFeeds , 2013.

[3] Mittal, A., & Goel , A., Sentiment Analysis On Twitter Feeds to Discover The Interrelationship Among "Public Sentiment "And The "Market Sentiment, 2012

[4] Gidofalvi, G., & Elkan, C, Predict Short-term Movement Of Stock Price Using Financial News Articles, 2003.

[5] Olivier C., Blaise Pascal University: "Neural network modeling for stock movement prediction, state of art". 2007

[6] Leng, X. and Miller, H.-G. : "Input dimension reduction for load forecasting based on support vector machines", IEEE International Conference on Electric Utility Deregulation, Restructuring and Power Technologies (DRPT2004), 2004.

[7] Nayak, S.C. :"Index prediction with neuro-genetic hybrid network: A comparative analysis of performance", IEEE International Conference on Computing, Communication, and Applications (ICCCA), pp. 1-6, 2012.

[8] Fazel Z.,Esmaeil H., Turksen B.: "A hybrid fuzzy intelligent agent-based system for stock price prediction", International Journal of Intelligent Systems, 2012.

[9] Honghai Y., Haifei L.: "Improved Stock Market Prediction by Combining Support Vector Machine and Empirical Mode Decomposition",Fifth International Symposium on Computational Intelligence and Design (ISCID), 2012