Stock Market Prediction Model by Combining Numeric and News Textual Mining

Kranti M. Jaybhay
ME (CSE) - II
Computer Science Department
Walchand Institute of Technology,
Solapur University, India

Rajesh V. Argiddi
Asst.Prof.Dept.CSE
Computer Science Department
Walchand Institute of Technology,
Solapur University, India

S.S.Apte, PhD.
Head Prof.Dept.CSE
Computer Science Department
Walchand Institute of Technology,
Solapur University, India

ABSTRACT

This paper proposes a novel method for the prediction of stock market closing price. Many researchers have contributed in this area of chaotic forecast in their ways. Data mining techniques can be used more in financial markets to make qualitative decisions for investors. Fundamental and technical analyses are the traditional approaches so far. ANN is a popular way to identify unknown and hidden patterns in data is used for share market prediction.

A multilayered feed-forward neural network is built by using combination of data and textual mining. The Neural Network is trained on the stock quotes and extracted key phrases using the Backpropagation Algorithm which is used to predict share market closing price.

This paper is an attempt to determine whether the BSE market news in combination with the historical quotes can efficiently help in the calculation of the BSE closing index for a given trading day.

KEYWORDS

Stock Market, Data Mining, Artificial Neural Network, Back propagation algorithm, Key phrases extraction algorithm.

1. INTRODUCTION

Data mining is the process of sorting through large amount of data and extracting the relevant information from it. Many financial analyst and business intelligence people make use of this information for knowledge discovery by applying various methodologies. Financial news articles from newspapers are the examples of such type of web data. This data is normally in the form of text as well as numeric values.

Forecasting stock market prices has always been challenging task for many business analyst and researchers. In fact, stock market price prediction is an interesting area of research for investors.

For successful investment lot many investors are interested in knowing about future situation of market. Effective prediction systems indirectly help traders by providing supportive information such as the future market direction. E.g. if the direction of a selected stock during 24 hours is predicted to be "up", buying the stock would be a profitable trading action.

Data mining techniques are effective for forecasting future by applying various algorithms over data. Web is rich textual information resource such as financial news even that is unmanageable to one. But one can use this abundance textual information to get datasets of various companies.

This project aims at predicting stock market by using financial news and quotes in order to improve quality of output. We are combining data mining time series analysis and machine learning algorithms such as **Artificial Neural Network** which is trained by using back propagation algorithm.

Recently, the web has rapidly emerged as a great source of financial information ranging from news articles to personal opinions.

Financial information sources on the Web.

1. www.yahoo.finance.com

2. www.stockwatch.in

3. Financial Times (www.ft.com) maintain excellent electronic versions of their daily issues.

4. Reuters (www.investools.com),

This rich variety of on-line information and news make it an attractive resource from which to mine knowledge. Data mining and analysis of such financial information can aid stock market predictions.

Numerical stock quotes collected from yahoo/finance are available in structured manner, but we have to employ some techniques to parse textual news information. The preprocessor unit of this system is developed with a priori domain knowledge. Information about **Indian stock market** is collected from above websites released daily.

2. RELATED WORK

B. Wüthrich et al, in 1998, analyzed news articles, collected from five popular financial websites, available before the opening of the Hong Kong stock market with several text mining techniques (k-nearest-neighbor and different types of neural networks) [1].

He presented techniques and developed facilities for exploiting especially textual financial news and analysis results to produce periodically forecast from stock market. Mittermayer extended B Wüthrich research through NewsCATS system for Intraday forecast of stock market [2]. Robert P. Schumacher and Hsinchun Chen research examines a predictive machine learning approach for financial news articles analysis using several different textual representations [3]. Manisha V. Pinto and Kavita Asnani focused more on key phrases extraction algorithm KEA for structuring news documents [5] [7].The Efficient Market Hypothesis (EMH) states that at any time, the price of a share fully captures all known information about the share [6].

Adebiyi Ayodele has given a hybridized approach for prediction by combining fundamental and technical analysis [8]. Zabir Haider Khan applied ANN technique to predict the Bangladesh Stock Exchange market index values with reasonable a degree of accuracy [9]. Neural networks are used to predict stock market prices because they are able to lead nonlinear mappings between inputs and outputs.

3. PREDICTION METHODS

The prediction of the market is without doubt an interesting task. Many methods are used to accomplish this task. These methods use various approaches, ranging from highly informal ways (e.g. the study of a chart with the fluctuation of the market) to more formal ways (e.g. linear or non-linear regressions). These techniques are categorized as Technical Analysis Methods, Fundamental Analysis Methods, Traditional Time Series Prediction Methods, and Machine Learning Methods. The criterion to this categorization is the type of tools and the type of data that each method is using in order to predict the market. What is common to these techniques is that they are used to predict and thus benefit from the market's future behavior [9].

3.1 Technical Analysis Methods

Technical analysis is the method of predicting the appropriate time to buy or sell a stock pricing. The idea behind technical analysis is that share prices move in trends dictated by the constantly changing attributes of investors in response to different forces. The technical data such as price, volume, highest and lowest prices per trading period is used for charts to predict future stock movements [9].

Chartists or technical analyst extracts trading rules from these charts that are useful in the financial market environment. Technical analysts believe that the market is only 10 percent logical and 90 percent psychological.

This is a very popular approach used to predict the market, which has been a work of critic. The major point of this criticism is that the trading rules extraction from the study of charts which is highly subjective therefore different analysts might extract different trading rules by studying the same charts.

It is possible to use this methodology to predict the market on daily basis; still we will not follow this approach on this study due to its subjective characteristic.

3.2 Fundamental Analysis Techniques

This technique uses the principle of the firm foundation theory for the selection of preferred stock. Fundamental data is used by analysts to apply this method of prediction to have clear idea about the market or firm for investment. The growth, the dividend payout, the interest rates, the risk of investment, the sales level, tax rates and so on are the variables used to determine the 'real' value of the asset that they will invest in [9]. Main objective of this technique is the calculation of an intrinsic value of an asset. For this they apply simple trading rules of investments. If the intrinsic value of the asset is higher than the value it holds in the market, invest in it. If not, consider it is a bad investment and avoid it. Analysts for this technique believe that the market is 90 percent logical and 10 percent physiological. In Adebiyi Ayodele has combined this technique with technical analysis method [8]. This type of analysis is not suitable for the objectives of our study. The reason for this is that the data it uses in order to determine the intrinsic value of an asset does not change on daily basis. Thus fundamental analysis helps for prediction of the market only in a long-term basis.

3.3. Traditional Time Series Prediction

This method analyzes historical data and attempts to approximate future values of a time series as a linear combination of these historic data. Regression models have been used to predict stock market time series. There are two basic types of time series analysis simple regression or univariate and multivariate regression [9].

A set of factors that mainly influence the time series under the prediction is formed. These variables are the self explanatory variables x_{it} of the prediction model. They are mapped with values of time series y_t such that a pair $\{x_{it}, y_t\}$ is formed. The linear combination of x_i that approximates in an optimum way y is defined. Univariate models use one explanatory variable while multivariate models uses more than one variable. Multivariate regression is used in the work of Pesaran and Timmermann (1994) [11].

S&P 500 and the Dow Jones predictions of the excess returns time series is done by Pesaran and Timmermann. They applied this prediction technique on monthly, quarterly and annually pattern. The data they used was from Jan 1954 until Dec 1990 [11].

3.4. Machine Learning Methods

Inductive learning is the main criterion for Machine learning. These methods use a set of samples to generate an approximation of the underling function that generated the data. The approximation or drawing a conclusion from various sample data which is presented to the model is the main objective. The Nearest Neighbor and the Neural Networks Techniques are methods that have been applied to market prediction. Artificial Neural Networks is a popular way of prediction of stock market on daily basis [9].

3.5 Comparative Study of Prediction Techniques

Table 1: Comparative Study

Criteria	Technical Analysis	Fundamental Analysis	Traditional Time Series Analysis	Machine Learning Techniques
Data Used	Price, volume,highest,lowest prices	Growth, dividend payment, sales level, interest rates, tax rates etc.	Historical data	Set of sample data
Learning methods	Extraction of trading rules from charts	Simple trading rules extraction	Regression analysis on attributes is used	Inductive learning is used
Type of Tools	Charts are used	Trading rules	Simple Regression and Multivariate analysis used for time series.	Nearest neighbor and Neural Networks are used
Implementation	Daily basis prediction	Long –term basis prediction	Long –term basis prediction	Daily basis prediction

4. SYSTEM DESIGN

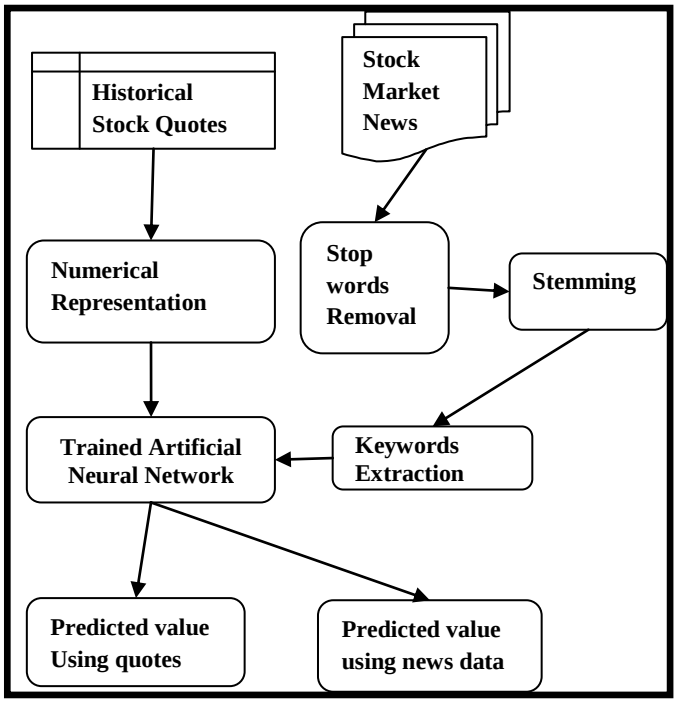

Figure.1 System Architecture

4.1 SYSTEM OVERVIEW

4.1.1 Collection of Stock quotes and News:

We are collecting the stock information once in day.

1. BSE Stock Index Dataset: The data released over financial websites is in both the forms like numerical quotes and textual format news data.

2. It is collected from yahoo/finance; they are available on sites at all the time as in .csv file.

3. Historical prices of the stock quotes and the daily published news are collected.

4.1.2 Numerical Representation

Stock quotes are normalized by scaling its units so they occur in small range of 0.0. to 1.0 [6] .These normalized values are input values for each attribute in the training tuples so as to speed up learning phase.

4.1.3 News Documents Preprocessing

The Preprocessor unit is used to process unstructured news documents. It is fed up with a priori domain knowledge such as .txt files of stop words such as a, an, the, of etc.

The preprocessor unit involves following steps

I. Stop Words Removal: Its first step in conflation algorithm, input to this is .txt file containing stop words list like articles, conjunctions, prepositions etc. as well the URL of any financial news website. This website will provide the news data.

II. Stemming: To fetch the exact grammatical root format of word the stemming process is used. It trims the words e.g. buying to be reduced to 'buy' as root.

III. Key Phrases Extraction: Preprocessed news articles are generating keyword phrases corresponding to the given .txt file global list of topmost keyword phrases.

These key phrases are initiated with some weight. KEA algorithm is used to obtain single global list of keywords phrases [8].

4.1.4 Prediction Module

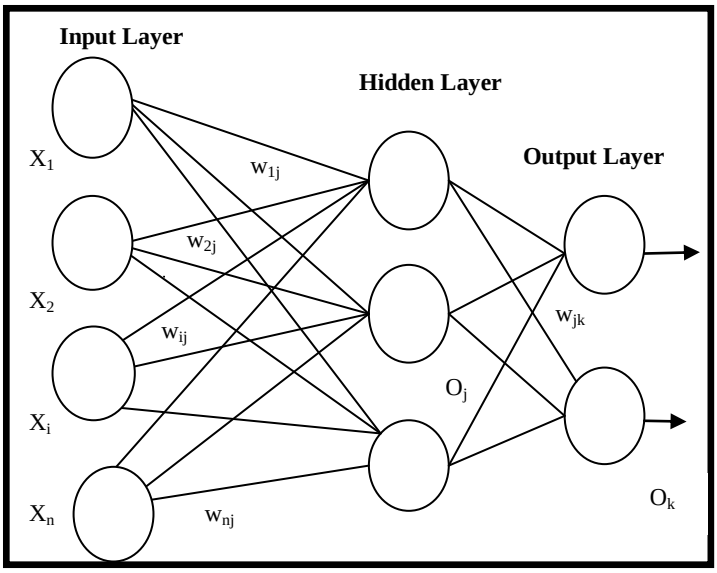

Figure.2. Multilayered feed-forward Neural Network [6]

I. Training Artificial Neural Network to Predict Stock Market

The prediction module compromises trained multilayered feed forward neural network. Backpropagation is the learning algorithm for neural network [8]. Roughly speaking, a neural network is the set of connected input/output units in which each connection has weight associated with it. During learning phase, the network learns by adjusting the weights so as to able to predict the correct class label of the input tuples.

Multilayered Feed-Forward network

This neural network consists of an input layer, one or more hidden layer, and one output layer.

Input layer: This layer is made of units; input to the network corresponds to the attributes measured for each training tuple. They are fed to this layer simultaneously. These inputs passes through input layer and weighted and simultaneously fed to the next layer i.e. hidden layer.

Hidden Layer: The outputs of the input layer are inputs to this hidden layer. The number of hidden layer is arbitrary; in practice only one hidden layer is used. The weighted outputs of the hidden layer are inputs to the next or output layer, which actually emits the networks prediction for given tuples.

Output Layer: This layer actually emits the networks prediction for given tuples. Multilayer feed-forward networks are able to model the class prediction as a nonlinear combination of the inputs. For given hidden units and enough training samples, can closely approximate any function.

II. Backpropagation Algorithm in Neural Network [6]

The study used three-layer (one hidden layer) multilayer feed-forward neural network model trained with backpropagation algorithm.

A.Selection of Multilayered feed-forward network:

A network is built with three layers (one hidden layer) is used. Historical stock prices of different companies were obtained from published stock data on the Internet. The learning function or the activation function used is sigmoid function,

$$O_j = \frac{1}{1+e-1j}$$ // output unit

B. Error Backpropagation:

To update the weight and bias the errors of the network's prediction are backpropagated. For the output unit j, and error of the hidden layer are calculated as,

$$Err_j = O_j (1 - O_j) (T_j - O_j);$$

$$Err_j = O_j (1- O_j) \sum_k Err_k w_{jk}$$

Where T_j is the target value of the prediction class.

C. Weight and Bias updation:

In each iteration weight and bias values are updated to reflect the errors.

$$\Delta w_{ij} = (l) Err_j O_i ;$$ //weight increment

$$w_{ij} = w_{ij} + \Delta w_{ij ; }$$ //weight update

Where l is the learning rate which is pre-specified

Algorithm: Neural network learning for prediction using Backpropagation Algorithm [6].

Input:

= D, a data set consisting of the training tuples and their associated target values;

= l, the learning rate;

=network, a multilayer feed-forward network.

Output: A trained neural network.

Method:

1. Initialize all weights and biases in network;
2. While terminating condition is not satisfied {
3. for each training tuple X in D
//propagate the inputs forward:
4. for each input layer unit j{
5. $O_j = I_j$; //output of an input unit is its actual input value
6. for each hidden or output layer unit j{
7. $I_j = \sum_i w_{ij} O_j + \theta_{j;}$ // compute the net input of the unit j with respect to the previous layer, i
8. $O_j = \frac{1}{1+e-1j}$;} // compute the output of each unit j
9. //Backpropagate the errors;
10. for each unit j in the output layer
11. $Err_j = O_j (1 - O_j) (T_j - O_j)$; //compute the error
12. for each unit j in the hidden layers, from the last to the first hidden layer
13. $Err_j = O_j (1- O_j) \sum_k Err_k w_{jk ;}$ // compute the error with respect to the next higher layer, k
14. for each weight w_{ij} in network {
15. $\Delta w_{ij} = (l) Err_j O_i$; //weight increment
16. $w_{ij} = w_{ij} + \Delta w_{ij ; }$ //weight update

17. for each bias θ_j in network {

18. $\Delta\theta_j$ = (l) Errj; // bias increment

19. $\theta_j = \theta_j + \Delta\theta_j$ };// bias update

20.}}

5. IMPLEMENTATION

This research uses three layered Neural Network. The numerical representation of the stock quotes and the key phrases from news articles are taken as input units to the input layer of multilayered feed-forward network.

The network configuration details are as follows:

I. Neural Network

1. Input Layer: In this configuration the layer has around 25 units. Neurons 1-5 are stock quotes and neurons 6 to 25 Boolean values i.e. Presence/Absence of the top 20 global key phrases in the key phrases extracted for the news articles corresponding to a given trading day.

2. Hidden Layer: only one hidden layer with 30 units is used.

3. Output Layer: Only one output unit is used as the closing price approximation value.

4. Learning Rate: A learning rate 0.2 is used.

5. Terminating condition: When the error is below threshold value.

II. Prediction Module

It comprises of the neural network trained using the back – propagation algorithm. Given a day's open index, day's high, day's low, volume traded and the adjusted close values (all are in normalized form) along with the stock news data, the predictor module will predict the closing index value for a given trading day. The above specified inputs correspond to the data that is observed after stock market closes every day. The predicted value is then de-normalized to obtain the actual close index value.

6. RESULTS AND ANALYSIS

The system evaluation on the stocks from India's Bombay Stock Exchange is carried out. For given day's open index, day's high, day's low, volume and adjacent values along with the stock news textual data, our predictor will predict the closing index value for given trading day.

Our predictive model is evaluated on BSE market on the financial historical stock data over the training period of November 2008 to August 2012. The news data is collected from the financial web sites http://www.finance.yahoo.com, http://reuters.com and www.Stockwatch.in. The news data is collected once in day. The stock quotes corresponding to each trading day were downloaded from http://finance.yahoo.com.

The accuracy of the system is measured as the percentage of the predictions that were correctly determined by the system. For instance, if the system predicts an uptrend and the index indeed goes up, it is assumed to be correct, otherwise, if the index goes down or remains steady for an uptrend, it is assumed to be wrong.

Following stock dataset in Table No.2 is taken as sample training data over the period of 25 days. Corresponding rates file is also provided along with this. Predictions using stock quotes are shown in the Table No.3.

Whenever the desired predictions using quotes are varying from actual one, we rebuilt neural network by considering the news data of that day.

KEA helps us to give global list of key phrases [7]. Sample list of such key phrases is given in Table No.4.

Performance analysis of the prediction model is shown in Table no.5.

Table 2: Dataset

Date	Open	High	Low	Close	Volume	Adj Close
10/2/2012	274.7	274.7	274.7	274.7	0	274.7
10/1/2012	268	276.2	268	274.7	9,420,100	274.7
9/30/2012	264	272	264	264.45	1376453	264.45
9/29/2012	264	272	264	264.45	1376453	264.45
9/28/2012	264.2	272	263.15	267.55	11,540,200	267.55
9/27/2012	263.9	267.3	258.5	259.75	902422	259.75
9/26/2012	267	268.05	262.35	263.1	658708	263.1
9/25/2012	274.25	276	266.65	268.85	879132	268.85
9/24/2012	274	279.65	272	273.4	1685233	273.4
9/23/2012	273	277.1	271.5	275.35	1438708	275.35
9/22/2012	273	277.1	271.5	275.35	1438708	275.35
9/21/2012	273	277.1	271.5	275.35	1438708	275.35
9/20/2012	268	278.55	268	272.55	1836363	272.55

9/19/2012	268	278.55	268	272.55	1836363	272.55
9/18/2012	276	278.35	272.8	274.05	929964	274.05
9/17/2012	276	289.3	274.45	277.65	3100094	277.65
9/16/2012	263	272	263	270.3	1651263	270.3
9/15/2012	263	272	263	270.3	1651263	270.3
9/14/2012	263	272	263	270.3	1651263	270.3
9/13/2012	264	264.95	257.8	259.1	1080808	259.1
9/12/2012	251	263.8	251	262.8	1971547	262.8
9/11/2012	246	250.9	245.6	249.6	1125614	249.6
9/10/2012	246	249.75	242.75	248.85	916967	248.85
9/9/2012	244.15	246.5	244.15	245.7	126186	245.7

Table 2: Result

Date	Actual BSE	Predicted BSE	Actual PIR	Predicted PIR	RMS Error
10/2/2012	1115.1	1132.24	3.25	3.17	0.0189
10/1/2012	1132.99	1132.88	3.25	3.16	0.0247
9/30/2012	1136.52	1136	3.25	3.16	0.0231
9/29/2012	1137.14	1140.68	3.25	3.15	0.0209
9/28/2012	1141.69	1141.84	3.25	3.15	0.0214
9/27/2012	1144.98	1141.91	3.25	3.13	0.0226
9/26/2012	1146.98	1144.94	3.25	3.14	0.0224
9/25/2012	1136.22	1143.41	3.25	3.11	0.021
9/24/2012	1145.68	1144.13	3.25	3.12	0.023
9/23/2012	1148.46	1143.05	3.25	3.11	0.0247
9/22/2012	1136.03	1144.73	3.25	3.11	0.0202
9/21/2012	1150.23	1146.34	3.25	3.11	0.0244
9/20/2012	1138.04	1146.12	3.25	3.1	0.0196
9/19/2012	1116.48	1148.04	3.25	3.1	0.0151
9/18/2012	1091.76	1145.73	3.25	3.1	0.0183
9/17/2012	1096.78	1145.1	3.25	3.1	0.0168
9/16/2012	1092.17	1145.05	3.25	3.13	0.018
9/15/2012	1097.5	1143.38	3.25	3.13	0.0166
9/14/2012	1084.53	1145.39	3.25	3.19	0.0202
9/13/2012	1073.87	1141.97	3.25	3.19	0.0228
9/12/2012	1089.19	1137.53	3.25	3.21	0.017
9/11/2012	1103.32	1133.02	3.25	3.21	0.0151
9/10/2012	1097.28	1131.45	3.25	3.24	0.0156
9/9/2012	1063.11	1125.53	3.25	3.24	0.0203

Table 3: Key Phrases

Sr.No.	Key Phrases
1	Hike
2	Jump
3	Flat
4	Buy
5	Loses
6	Down
7	Lower
8	Steady
9	Recession
10	Scam
11	Swoon
12	Slump

Table 4: Performance Table

No. of Inputs	Dataset
No. of Inputs	543
No. of results	543
Correct result	481
Precision Rate	94.34
Recall Rate	91.12

7. CONCLUSION AND FUTURE WORK

Determining the Stock market forecasts is always been challenging work for business analysts. Thus, we attempted to make use of huge textual data to predict the stock market indices. If we combine both techniques of textual mining and numeric time series analysis the accuracy in predictions can be achieved. Artificial neural network is trained to predict Bombay Stock Exchange market future trends.

Financial analysts, investors can use this prediction model to take trading decision by observing market behaviour.

More work on refining key phrases extraction will definitely produce better results. Enhancements in the preprocessor unit of this system will help in improving more accurate predictability in stock market.

8. REFERENCES

[1]. B. Wüthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, and W. Lam, "*Daily Stock Market Forecast from Textual Web Data*", Proceedings of the IEEE International /Conference on Systems, Man, and Cybernetics, IEEE Computer Society Press, Los Alamitos, CA, 1998, pp. 2720-2725.

[2]. Marc-André, Mittermayer, "*Forecasting Intraday Stock Price Trends with Text Mining Techniques*", Proceedings of the 37th Hawaii International Conference on System Sciences - 0-7695-2056-1/04 $17.00 (C) 2004 IEEE.

[3]. Robert P. Schumacher and Hsinchun Chen , "*Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFinText System"* Artificial Intelligence Lab, Department of Management Information Systems The University of Arizona, Tucson, Arizona 85721, USA.

[4] Manisha V. Pinto, Kavita Asnani "*Stock Price Prediction Using Quotes and Financial News***",** International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-5, November 2011

[5] Ramon Lawrence, *"Using Neural Networks to Forecast Stock Market Prices"*, Department of Computer Science University of Manitoba umlawren@cs.umanitoba.ca December 12, 1997.

[6]. Jiawei Han and Micheline Kamber, "*Data Mining concepts and Techniques*", Second Edition 2006.

[7]. Ian H. Witten,* Gordon W. Paynter,* Eibe Frank,* Carl Gutwin† and Craig G. Nevill –Manning, "KEA: *Practical Automatic Key phrase Extraction".*

[8]. Adebiyi Ayodele A., 1Ayo Charles K., 1Adebiyi Marion O., and 2Otokiti Sunday O. *"Stock Price Prediction using Neural Network with Hybridized Market Indicators*", Journal of Emerging Trends in Computing and Information Sciences, VOL. 3, NO. 1, January 2012 ISSN 2079-8407.

[9]. Zabir Haider Khan, Tasnim Sharmin Alin, Md. Akter Hussain, Department of CSE, SUST, Sylhet, Bangladesh ,"*Price Prediction of Share Market using Artificial Neural Network (ANN)*", International Journal of Computer Applications (0975 – 8887) Volume 22– No.2, May 2011.

[10] Y.-Q. Zhang and A. Kandel, "*Compensatory Genetic Fuzzy Neural Networks and Their applications,*" Series in Machine Perception Artificial Intelligence, Volume 30, World Scientific, 1998.

[11] M. Hashem Pearsan and Allan Timmermann, "*Predictability of Stock Returns:Robustnesss and Economic Significance"* , The Journal of Finance, Vol.50,No.4.(Sep.,1995),pp.1201-1228.