Stock Market Prediction using Artificial Neural Network & Text Mining

Jibendu Kumar Mantri, Amiya Kumar Sahoo, Sipali Pradhan, Debabrat Dehury

***Abstract:** The art of prediction of stock market volatility has always been a most challenged interdisciplinary research problem among scientist due to its highly non- linear nature of market flow. This paper tries to analysis the historical data of BSE Sensex using extreme volatilities estimators, GARCH, ANN and new proposed Text Mining approach for stock market predictions. Finally experimental results illustrates that the new proposed Text model can able to predict the volatilities of the stock price better than other models.*

Keywords: MLP, GARCH, Text Mining

I. INTRODUCTION

Computation finance is a challenging subject in the field of Applied Computer Science which solves the practical problems of finance with Mathematical proofs considering the study of data analysis and algorithms of Soft computing methods. Prediction of stock market is an emerging task in this subject which leads to the extension of research works range of computation methods from personal computer to super computer and mainframe. Now-a-days it is observed that financial markets are considered to have important role in economic conditions for countries. Hence the main aspects of stock market are to model and estimate its volatility predictions caused by its dynamic fluctuations in stock prices which is a measure of risk.

II. LITERATURE SURVEY

Based upon these challenges , many attempts have been made for understanding the volatility fluctuations of stock market and its prediction by using the mathematical models with the study of historical data of stock prices[5][6][10]and [12]. Latter on, more analysis have been initiated for analysis and prediction using artificial neural network(multilayer perceptron) due to universal functions approximation and robust nature of Artificial neural network for prediction[1][3][4][8]and[9].Also dynamic financial forecasting is studied by using integration of Artificial neural network with Fuzzy and induced Fuzzy associations rules[7] and [11].

III. METHODOLOGY

Over the last decades, many attempts have been made for understanding volatility fluctuations and its predictions using Parkinson model(1980), Rogers and Satchell model(1980)

Garman and Klass model(1991) ,GARCH model and Neural Network model etc of data mining concepts. However, these concepts suffer from various drawbacks and therefore results are very difficult to understand and illustrating inaccurate predictions as it is a non-linear, nonparametric system in nature. Hence, attempts have been made here to analysis its nature for future predictions using the historical data of Indian stock market (BSE Sensex) for 2008 to 2018 considering 2698 rows of data (daily opening, high, low and close prices of stock market data) using extreme volatilities estimators Parkinson model, Roger Satchell model, Garman & Klass model, GARCH model and also specially with New proposed Text Mining approach which exhibits better accuracy enabling the researcher to think and analysis further using this New proposed model with the concepts of Signal processing, Natural Language Processing etc for predictions. The daily opening, high, low and close prices of stock market data.

According to Parkinson model the calculation of volatility is expressed as

$$\sigma_{PV} = k.sqrt(\frac{1}{n}\sum In(\frac{H_t}{L_t})^2$$

Where, k = 0.601 or k = 1
H_t = Intraday High Price
L_t = Intraday Low Price
n = no. of observations
σ PV = Volatility of return

After Parkinson, Garman and Klass (1981) extended his work to calculate the volatility

$$\sigma_{GKV} = \sqrt{\frac{1}{n}\sum\left[(0.5)\left[In\left({}^{H_t}\!/\!{}_{L_t}\right)\right]^2 - [2In(2)-1]\left[In\left({}^{C_t}\!/\!{}_{O_t}\right)\right]^2\right]}$$

Where, n = No. of observations
H_t, L_t, C_t, and O_t are values of intra-day price i.e high, low, close and open respectively.

After him ,Roger and Satchell (1991) described another most robust, volatility estimator model as

$$\sigma_{RSV} = \sqrt{\frac{1}{n}\sum\left[\left[In\left({}^{H_t}\!/\!{}_{C_t}\right)\right]\left[In\left({}^{H_t}\!/\!{}_{O_t}\right)\right]+\left[In\left(\frac{L_t}{C_t}\right)\right]\left[In\left({}^{C_t}\!/\!{}_{O_t}\right)\right]\right]}$$

Manuscript published on January 30, 2020.
* Correspondence Author

Dr.Jibendu Kumar Mantri*, Associate Professor, PG Department of Computer Application, North Orissa University, Baripada, India, Email: jkmantri@gmail.com

Amiya Kumar Sahoo, Senior Lecturer in the Department of Computer Science and Engineering in Aryan College of Engineering, Bhubaneswar

Dr. Sipali Pradhan, Department Computer Science and IT from North Orissa University, Baripada, Odisha.

Debabrat Dehury, M. Phil. Scholar, Department of Computer Applications, North Orissa University, Baripada, Odisha.

But, Banarjee and Sarkar [2] expressed that the Indian stock market illustrates the volatility clustering and GARCH models can have the ability of better prediction of the market volatility for which the present study enables us to use GARCH (4,4), GARCH (5,5) to estimate the volatility prediction of BSE SENSEX.

T. Bollerslev and S.J.Taylor described the GARCH model in 1986. They explained that the conditional variance by GARCH (q, p) model as defined below:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \left(\alpha_i . u_{t-i}^2\right) + \sum_{j=1}^{p} \left(\beta_j . \sigma_{t-j}^2\right)$$

Where, σ_t^2 = conditional variance

With constraints

$\alpha_0 > 0;\ \alpha_i \geq 0$ i= 1,2,………..….q

$\beta_j \geq 0,$ j= 1, 2,………………...….p

for a positive conditional variance.

Thus, the volatility is expressed as a function of α_0, as a constant, u_{t-i}^2 expresses about volatility from the previous period 't$_{-i}$' (the ARCH model) and σ_{t-j}^2 is the forecast variance of previous periods (The GARCH model).

Now, the unconditional variance is expressed by GARCH (q, p) as

$$Var(u_t) = \frac{\alpha_0}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j}$$

Thus, the GARCH model requires $\alpha_1 + \beta_1 < 1$ for GARCH(1,1) for unconditional variance.

If $\sum \alpha_i + \sum \beta_i = 1$, then it is known as Integrated GARCH for unit root invariance' concept.

If $\sum \alpha_i + \sum \beta_i \geq 1$, the unconditional variance of u_t is not defined and it is known as the variance of non-stationarity.

Use of Artificial Neural Networks to Stock market forecasting have become very popular research over the last few years due to characteristics of non-linearity existence in stock market data. This main application of Artificial neural networks concepts to data analysis is the MLP-multilayer perceptron model. To be able to solve these nonlinearly separable problems here, a number of neurons are connected in the layers of the architecture to build a multilayer perceptron. Each of the perceptrons of this model is used to identify small linearly separable sections of the inputs. The Outputs of the perceptrons model are linked into another perceptron to produce the final output.

Algorithm

a. To Initialize weights (to small random values) and transfer function

b To give input variables

c. To weight adjustments from output layer with backtraking

$$w_{ij}(t+1) = w_{ij}(t) + \eta \delta_{pj} o_{pi}$$

$w_{ij}(t)$ known as the weights from node i to node j at time t, η is the gain term, and δ_{pj} is error term for pattern p on node j.

Output layer units

$$\delta_{pj} = \mathbf{k} o_{pj}(1 - o_{pj})(t_{pj} - o_{pj})$$

Hidden layer units

$$\delta_{pj} = \mathbf{k} o_{pj}(1 - o_{pj}) \sum \delta_{pk} w_{jk}$$

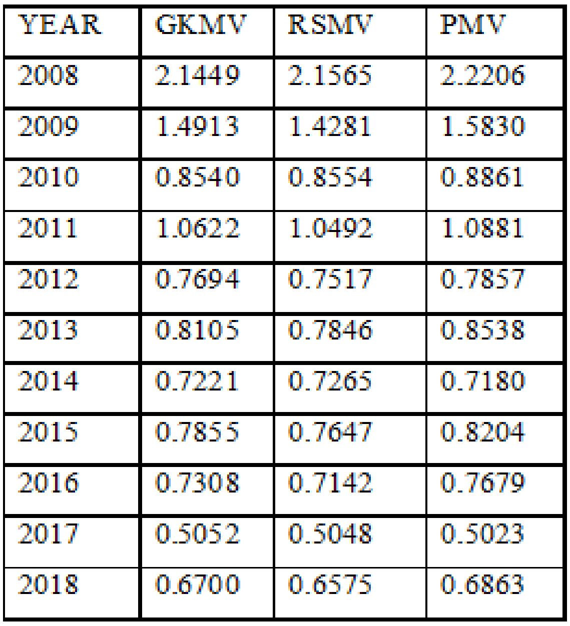
Here this paper tries to emphasize on the new application of our proposed concept of Text Mining on stock market. This proposed model tries to describes to use the numeric data (High, Low, open and close price) of stock market in textual form (Upper, middle upper, middle, lower middle and lower) with certain limitations as so that the text data will be used for analysis for prediction with better accuracy.

IV. RESULT AND DISCUSSION

At first the numeric data (BSE Sensex- from 2008-2018) is collected from yahoo finance.

Considering these numeric data (open, high, low & close prices of stock market), Table-1 reveals the analysis result of prediction of volatilities using the formulas described by Parkinson (1980), German & Klass (1980) and Roger & Satchell (1991).

Table-1 Yearly Volatility by Extreme Estimators BSE – SENSEX 2008 – 2018



YEAR	GKMV	RSMV	PMV
2008	2.1449	2.1565	2.2206
2009	1.4913	1.4281	1.5830
2010	0.8540	0.8554	0.8861
2011	1.0622	1.0492	1.0881
2012	0.7694	0.7517	0.7857
2013	0.8105	0.7846	0.8538
2014	0.7221	0.7265	0.7180
2015	0.7855	0.7647	0.8204
2016	0.7308	0.7142	0.7679
2017	0.5052	0.5048	0.5023
2018	0.6700	0.6575	0.6863

German-Klass Model Volatility	= GKMV
Roger - Sachet Model Volatility	= RSMV
Parkinson Model Volatility	= PMV

Also Table 2 reveals the predicated year wise volatilities of GARCH (4,4) and GARCH(5,5) from 2008 to 2018 by using the formula of GARCH model described in III (Methodology). Also, this Table 2 illustrates the numeric comparison of predicated volatilities of GARCH (4,4), GARCH (5,5) models with the estimation values of extreme volatility estimator models.

Table – 2 Year wise volatility estimations using Extreme volatility estimators and GARCH

YEAR	GKMV	RSMV	PMV	GARCH (4,4)	GARCH (5,5)
2008	2.1449	2.1565	2.2206	2.33	2.39
2009	1.4913	1.4281	1.5830	1.74	2.11
2010	0.8540	0.8554	0.8861	0.83	0.90
2011	1.0622	1.0492	1.0881	1.29	1.29
2012	0.7694	0.7517	0.7857	0.95	0.95
2013	0.8105	0.7846	0.8538	0.82	0.85
2014	0.7221	0.7265	0.7180	0.78	0.78
2015	0.7855	0.7647	0.8204	0.94	0.94
2016	0.7308	0.7142	0.7679	1.10	1.09
2017	0.5052	0.5048	0.5023	0.56	0.56
2018	0.6700	0.6575	0.6863	0.64	0.66

Though Artificial neural network model is better than GARCH model due to its architecture, regularization techniques, but our proposed new Text mining approach is designed to use the numeric data in textual form so that the text data which will be used for stock market prediction with better accuracy in future to think for buy or sell the share price in stock market. Hence after designing the new database and analysing, the yearly prediction values are calculated from daily data from 2008-18 which is summarized in Table-3 with that of predicated values of Neural network. Again all these text mining predicted values are again compared with that of extreme value estimators which are illustrated in Table-4 exhibiting better accuracy. The graphical representation of better accuracy of Text mining is illustrated in Figure-1.

Table-3 Year-wise Volatility of BSE – SENSEX 2008 to 2018 using Text Mining

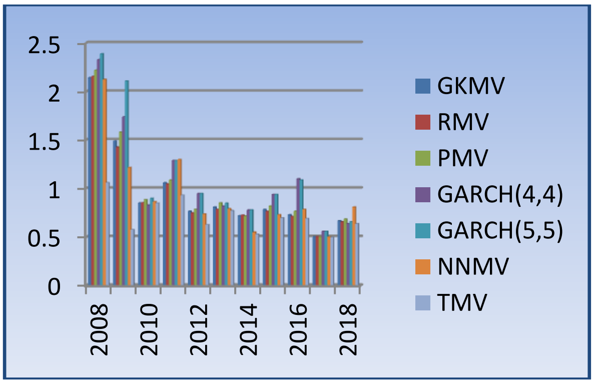
YEAR	TMV	NNMV
2008	1.0609	2.12598
2009	0.578	1.21912
2010	0.8501	0.86591
2011	0.932	1.30004
2012	0.628	0.73816
2013	0.7721	0.79434
2014	0.53094	0.55128
2015	0.6981	0.73217
2016	0.6924	0.78571
2017	0.4973	0.50374
2018	0.63874	0.81003

Table – 5 Yearly volatility Predictions (2008-18)

YEAR	GKMV	RMV	PMV	GARCH (4,4)	GARCH (5,5)	NNMV	TMV
2008	2.145	2.157	2.2206	2.33	2.39	2.12598	1.0609
2009	1.491	1.428	1.58302	1.74	2.11	1.21912	0.578
2010	0.854	0.855	0.88611	0.83	0.9	0.86591	0.8501
2011	1.062	1.049	1.08812	1.29	1.29	1.30004	0.932
2012	0.769	0.752	0.7857	0.95	0.95	0.73816	0.628
2013	0.811	0.785	0.85387	0.82	0.85	0.79434	0.7721
2014	0.722	0.727	0.71801	0.78	0.78	0.55128	0.53094
2015	0.786	0.765	0.82042	0.94	0.94	0.73217	0.6981
2016	0.731	0.714	0.76799	1.1	1.09	0.78571	0.6924
2017	0.505	0.505	0.50239	0.56	0.56	0.50374	0.4973
2018	0.67	0.658	0.68639	0.64	0.66	0.81003	0.63874

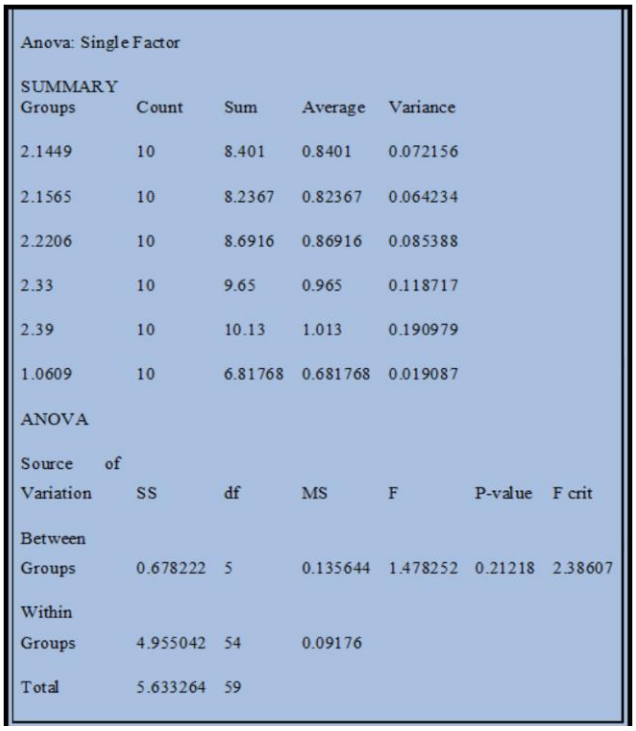
German-Klass Model Volatility	= GKMV
Roger Model Volatility	= RMV
Parkinson Model Volatility	= PMV
Neural Network Model Volatility	=NNMV
Text Mining Volatility	= TMV

Figure-1 Graphical Comparison-Yearly volatility prediction(2008-18) by Parkinson, Rogers and Satchell, German and Klass, GARCH, Neural Network (MLP) models and new proposed Text Mining approach.



Except the numeric and graphical representation, Anova test is conducted again for exhibiting statistically significant among the predicated results of all models. Figure-2 reveals for the analysis that the value of F is less than that of critical value 2.38607 concluding the remarks that there is no difference among the results of predicated volatilities of all models. But the smaller values obtained using Text Mining is better than others as they are compared with other models.

Figure-2 Statically representation Yearly Predictions(2008-18) by Parkinson, Rogers and Satchell, Garman and Klass, GARCH, Neural Network (MLP) models and new proposed Text Mining approach



Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
2.1449	10	8.401	0.8401	0.072156
2.1565	10	8.2367	0.82367	0.064234
2.2206	10	8.6916	0.86916	0.085388
2.33	10	9.65	0.965	0.118717
2.39	10	10.13	1.013	0.190979
1.0609	10	6.81768	0.681768	0.019087

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.678222	5	0.135644	1.478252	0.21218	2.38607
Within Groups	4.955042	54	0.09176			
Total	5.633264	59				

From this above calculation it shows that the new proposed Text Mining approach exhibits better prediction results than all previous methods. Hence, this aspect of text mining to covert the large volume of numeric and noisy data to text and analysis will lead to predict with better accuracy is challenging problem in day today life.

V. CONCLUSION AND FUTURE SCOPE

Here we can conclude that the proposed a new approach on text mining analysis on BSE stock market data (from 2008 to 2018) provides accurate results than the existing models like Roger, Parkinson and German Klass and GARCH,ANN model on stock market prediction which has a great impact in Stock market for decision making. Finally, it is proposed that the new text mining approach can be enhanced further by using DEA, Genetic algorithm, etc to enrich the study of predictions.

REFERENCES:

1. Aiken, M. and M. Bsat. (1999) “Forecasting Market Trends with Neural Networks.” Information Systems Management 16 (4)”, 42-48.
2. Banerjee, A. & Sarkar, S. (2006). Modeling daily volatility of the Indian stock market using intraday data. Working Paper Series No. 588, Indian Institute of Management Calcutta
3. Chiang, W.-C., T. L. Urban and G. W. Baldridge. (1996) “A Neural Network Approach to Mutual Fund Net Asset Value Forecasting.” Omega, Int. J. Mgmt Sci. 24 (2), 205-215.
4. Dunis Ch. L., Jalilov J. (2002), “Neural network regression and alternative forecasting techniques for predicting financial variables“,Vol. 12, Neural Network World, 113-139.
5. German M. & Klass M. (1980) “On the Estimation of security Price volatility from historical data” journal of business, Vol53, 67-69
6. Joshi M.P. & Pandey K (2008), **“**Exploring Movements of Stock Price Volatility in India” The Icfai Journal of Applied Finance **Vol.** 14, No. 3, 5-32.
7. Kuo, R. J., L. C. Lee and C. F. Lee. (1996) “Integration of Artificial Neutal Networks and Fuzzy Delphi for Stock Market Forecasting.” IEEE, June, 1073-1078.
8. Mantri J,K,,Gahan P and Nayak B.B (2010). Artificial Neural Networks –An Application To Stock Market Volatility. International Journal of Engineering Science and Technology,Vol. 2(5), 2010, ISSN: 0975-5462,1451-1460
9. Mitra Subrata Kumar: (2009), “Optimal Combination of Trading Rules Using Neural Networks” International Business Research, Vol 2, No 1 , 86-95.
10. Parkinson, (1980) “The extreme value method for Estimating the variance of the rate of Return”, Journal Business, Vol.53, 61-65
11. Romahi, Y. and Q. Shen. (2000) “Dynamic Financial Forecasting with Automatically Induced Fuzzy Associations.” IEEE, 493-498
12. Roger, L CG, Satchell, SE (1991) “ Estimating variance from High, Opening & Closing Price annals of applied probability,1(4),504-512.

AUTHORS PROFILE

Dr J. K. Mantri, pursed Master of Technology in Computer Sc. from Utkal University in year 2002 and Ph. D degree in Computer Application from Sambalpur University in year 2010. He is currently working as Reader in Department of Computer Application, North Orissa University, India. He is a life member of ISTE and ISCA. He has published 5 books and more than 58 research papers in reputed international journals including Thomson Reuters (SCI and Web of Science) and conferences including IEEE and Springer. His main research work focuses on Cryptography, Software Engineering and Computational Intelligence. He has 28 years of teaching experience.

Amiya Kumar Sahoo, is a Ph.D research scholar in the Department of Computer Applications, North Orissa University, Baripada, Odisha. He is working as a Senior Lecturer in the Department of Computer Science and Engineering in Aryan College of Engineering, Bhubaneswar. He has 12 years of teaching experience.

Dr. Sipali Pradhan, has completed her Ph.D in Computer Science and IT from North Orissa University, Baripada, Odisha. She has completed M.Tech in Computer Science and Engineering from Utkal University after completing B.Tech in Electronics and Communication Engineering from Biju Pattnaik University of Technology Odisha. Her research interest includes Internet of Things and its application in Healthcare.

Debabrat Dehury, is an M. Phil. Scholar in the Department of Computer Applications, North Orissa University, Baripada, Odisha. His research area is Artificial Intelligence, Data Mining and Soft Computing Techniques.