Web Sentiment Analysis: Comparison of Sentiments with Stock Prices using Automatic Linear Modeling

A. Pappu Rajan
Research Scholar ,Department of Computer Science
St.Xavier's College
Palayamkottai, Tamil Nadu , India

S. P. Victor
Research Guide, Department of Computer Science
St.Xavier's College
Palayamkottai, Tamil Nadu , India

***Abstract*— Recent day's web is a dominating and versatile tool. Over the web all people are discussing about all matters like personal, politics, world economy, trade market, share market, product and their quality. Based on the discussion or opinion other people will take decision. The recent researcher focusing on opinion mining because that will give mood of other person on the searched field. Opinion Mining is a process of extraction of knowledge from the opinion of others about some particular topic or a problem. Many way opinion mining is helping to the business people for their decision making. In this paper objective is how overall sentiments are positively correlated with the stock prices and they are selected to make predictions, using Automatic Linear Modeling, on the future stock price movements.**

Keywords—Opinion mining, sentiment analysis, Automatic Linear Modeling

I.INTRODUCTION

In early 1990s all sentiment related tasks only depend on Natural Language processing tasks even though lot of sub systems are available which can be accommodate ranging from small piece of information to huge volume of data . All this problems solved by single system called opinion mining or sentiment mining. In this work, we have calculated sentiment score and then the score was compared with stock price of the selected companies from our research. Our previous work was done for calculating sentiment score. After that calculated sentiment score is going compared with stock price of the company which will help us to predict accuracy of the system. In this paper deals how automatic linear modelling can be used to compare sentiments with the stock prices for better accurate prediction stock movements.

II.REVIEW OF LITERATURE

Manuel Sandoval et al. [2012], generalized linear models are a powerful tool to measure relationships between variables, as they can handle non normal distributions without altering the properties of variables involved. When applied to risk factor analysis, they can help determine the most important factors contributing to the incidence, prevalence or acquisition of a particular Medical condition. The macro was built in such a case. Several factors, both numerical and categorical, were tested using forward selection and defined criteria for entering the model and for keeping the variable in the model.

Ratner, [2012]. researchers often collect a data set with a large number of independent variables and each of them is a potential predictor of the dependent variable. The problem of deciding which subset(s) of the large pool of potential predictors to include in a linear regression model is; therefore, very common and arguably the hardest part of regression modeling

Chuan-Ju Wang et al. [Chu, 2013], they tried to explain the importance of sentiment analysis for financial risk and report. They used regression and rank techniques in sentiment lexicon for analyzing the relations between sentiment and financial risks. The system was also found financial sentiment on risk prediction.

Pedro Flores et al.[2012], first of all, the problem in finding a good linear model for a series of data is that it requires the determination of how many and which terms are the most appropriate to solve this problem. Secondly, it is necessary to know in which intervals the coefficients of the linear expression are, and finally, to find the values for these coefficients that minimize the quadratic error for the whole history. In this way we have to consider a problem of non-linear optimization with variables in real intervals, whose limits shall be specified. This problem has multiple local minimums; therefore, proper optimization techniques are required.

III. COMPARISON OF SENTIMENTS WITH STOCK PRICES

A.Motivation and Objective

From the review of literature, a number of approaches are used to improve the accuracy using sentiment score with other techniques. Here we are trying to find the relationship between the Twitter sentiment and stock prices to determine whether the sentiments are good indicators of future stock price

movements using ALM. In this research, we are going to predict the Closing prices of stock by using the Opening prices of stock and the overall sentiment as the predictors. The ALM is an automatic model which combines the predictors in such a way that the target value is predicted to its optimum level. The predictors are given an importance level automatically and they are combined to predict the target variable.

B.Data Set

Tweets are collected from five companies and also found sentiment score using positive and negative score analysis. The Stock prices of below mentioned companies are collected from the official website of National Stock Exchange (NSE) for a period from 1st November, 2013 to 20th December 2013. The following are the companies according to verticals:

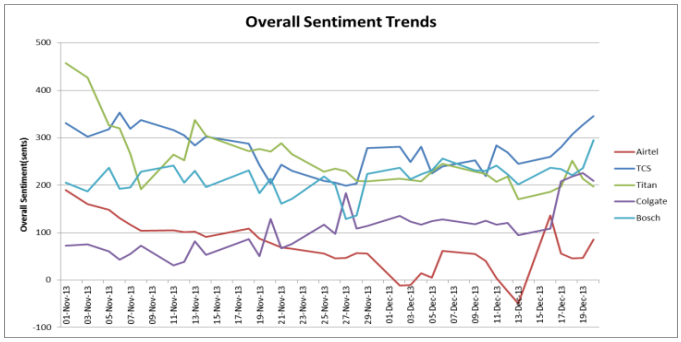
TABLE I LISTS OF COMPANIES

Company Name	**Type of Industry**
Bharti Airtel Ltd.	Telecommunication - Services
Titan Industries Ltd.	Retail
Bosch Ltd.	Automobiles
Tata Consultancy Services Ltd.	Computers – Software
Colgate Palmolive (India) Ltd.	FMCG

C. *Overall Opinion Analysis*

In the previous work, we have found overall sentiment score. The lexical sentiment analysis was performed over five companies. We have analyzed about 2290 tweets per day. So for 50 days from 1st November, 2013 to 20th December, 2013, a total of 1,14,500 tweets were analyzed for the entire project. The following chart represents the trends of all the five companies (i.e.) Airtel, Tata Consultancy Services, Titan Industries, Colgate Palmolive and Bosch by using the overall sentiment in the previous work of our research.

Fig.1
Overall sentiment



D. *Steps for Comparison of Sentiments with Stock Prices*

In opinion mining are various types of sentiment analysis as: word level , feature-level, entity-level, sentence-level, document-level ,. Data set are collected from different companies using Twitter by web crawling. Extracting data from SMN (Twitter) : using Twitter API REST API s are having the following resources : Time lines , tweets search, streaming , direct message, friends and followers ,users, ,suggested users ,favorites , lists, saved searchers, place and Geo , trends , spam reports, OAuth, help. These APIs use the pull strategy for data retrieval. To collect information a user must explicitly request it. Streaming APIs provides a continuous stream of public information from Twitter. These APIs use the push strategy for data retrieval. Once a request for information is made, the Streaming APIs provide a continuous stream of updates with no further input from the user. Opinion Retrieval involves retrieving desired information from bag-of-words or Twitter textual data to measure ad hoc information retrieval effectiveness in the standard way of methodology. A standard binary assessment of either relevant or non-relevant for each query-document pair. Sentiment Extraction: Finding or discovering of target entity. It uses various method to extract the sentiment from sentiment document using unsupervised learning, supervised learning and lexicon based approach. Sentiment Classification: Positive / Negative -Score Analysis: To find weather a piece of text is opinionated or not and to find the polarity of the text. This classification may be binary or multiclass classification. Up to this step we have already completed and calculated sentiment score. Now in this paper we going to do compare these score value with stock price of the company. The above given company stock prices above mentioned companies are collected from the official website of National Stock Exchange (NSE) for a period from 1st November, 2013 to 20th December 2013. The same period we already collected twitter data from the same companies and to using sentiment score analysis technique to calculate sentiment score values. These sentiment score value can be compared with Stock Price and finding the perfect predictors using ALM.

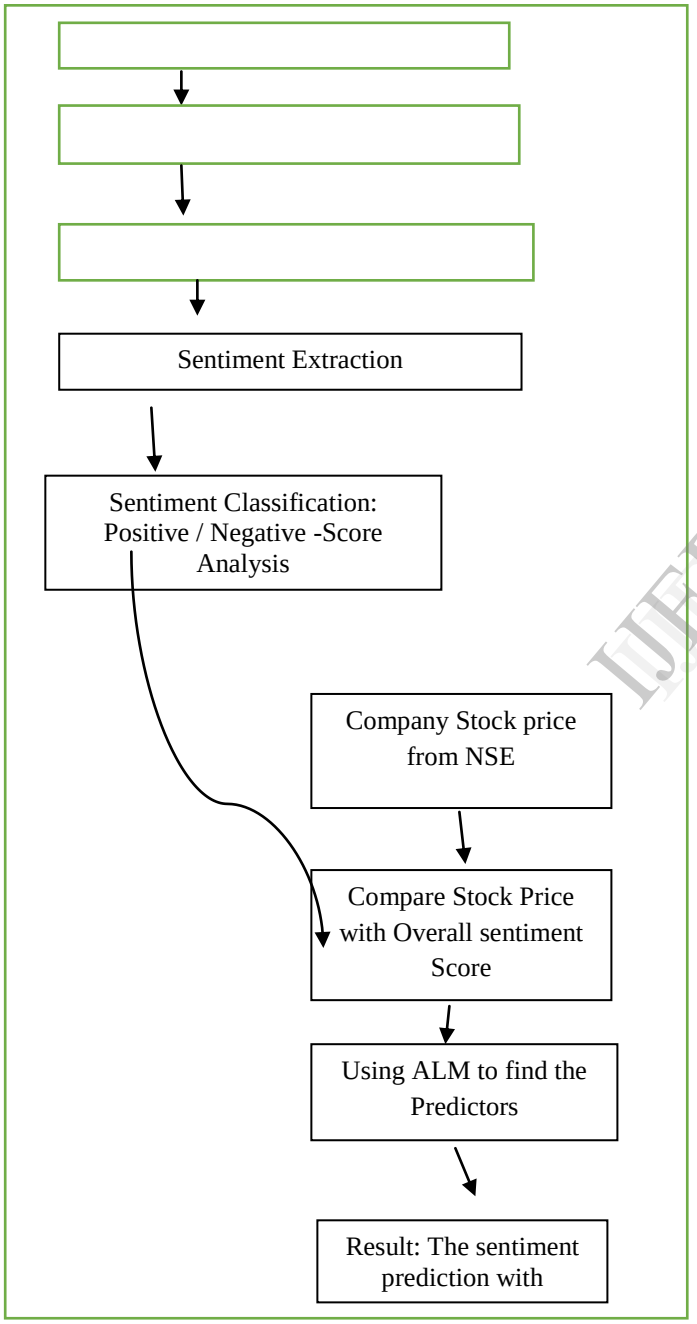
IV. WEB SENTIMENT WITH AUTOMATIC LINEAR MODELING

Given the new features of the linear procedure, it is important for researchers who use regression analysis regularly to take advantage of them. However, a review of the most recent regression literature indicates very few texts ever mention its use. The given that scarcity of coverage in the literature may prevent the features of the new and challenging procedure from being fully utilized; All the problem of statistical tool overcome with new usage of ALM. In general linear models predict relationships between the continuous target and one or more predictors. Which can be easy to formulate mathematical model with scoring that model easily and quickly compare to other model such as neural networks or decision trees using the same data set. The main goal of linear regression is all predictors are combine together and predict values on a single scaled outcome variable. By using Automatic Linear Modeling, we can predict the values of the target variable by using the predictors as the input. In this research, we are going

to predict the Closing prices of stock by using the Opening prices of stock and the overall sentiment as the predictors. The ALM is an automatic model which combines the predictors in such a way that the target value is predicted to its optimum level. The predictors are given an importance level automatically and they are combined to predict the target variable. The model selection method used in this research is Forward Stepwise.

Fig2

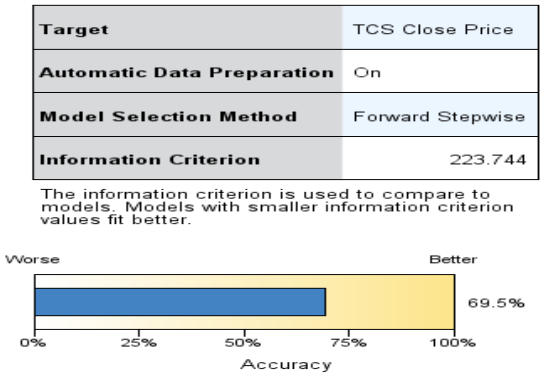
Steps for Comparison of Sentiments with Stock Prices



A.Model Summary of ALM

The first visible model view is a high-level summary of the model and its fit. From the table, we can see that automatic data preparation and model selection were performed as part of the model building process. More detailed results of these operations are visible in other views. The below figure represents the accuracy of the model by using the predictors to predict the target variable. In this case shown model has a higher accuracy of about 69.5%, which is a higher value for this analysis. The method used in performing the Automatic Linear Modeling is Forward Stepwise.

Fig.3. Model Summary for Automatic Linear Modeling



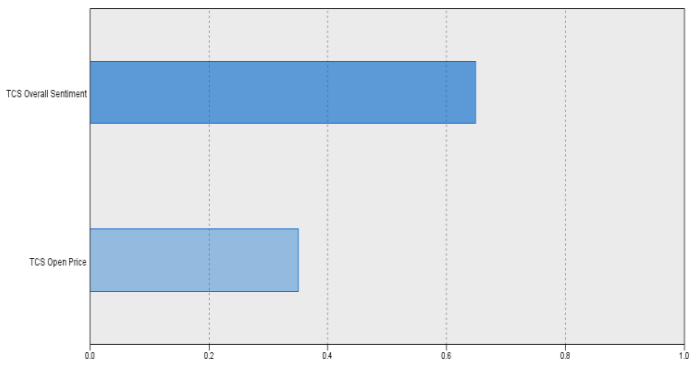
B. *Automatic Data Preparation*

Automatic data preparation provides details of the actions taken during the automatic data preparation step for model building. In general Automatic Linear Modeling is to trim out the outliers. Both the predictors (i.e.) open Price of a company and Overall sentiment of the company has to trim out the outliers. The variables are named with a suffix of transformed to represent that the outliers are trimmed.

C. *Predictor Importance*

Predictor importance shows the predictors in the final model in rank order of importance. In this research the predictor importance in ALM gives a clear picture which shows the importance given to the predictors in predicting the target variable (i.e.) in this case Closing Price of a company.

Fig. 4 Predictor Importance a Company



From the above chart, it is clear that company overall sentiment is given an importance of 0.65 while company opening price is given an importance of 0.35. In this case, the sentiment is given a higher importance in predicting the closing price of a company for the next day. In this work also predict the significance test for the predictors, which shows the p value below 0.05. Thus, both the predictors are well utilized in predicting the target variable.

D.*Predicted value*

The following table represents the actual closing price and the closing price predicted using the Opening price and the overall sentiment of few sample data.

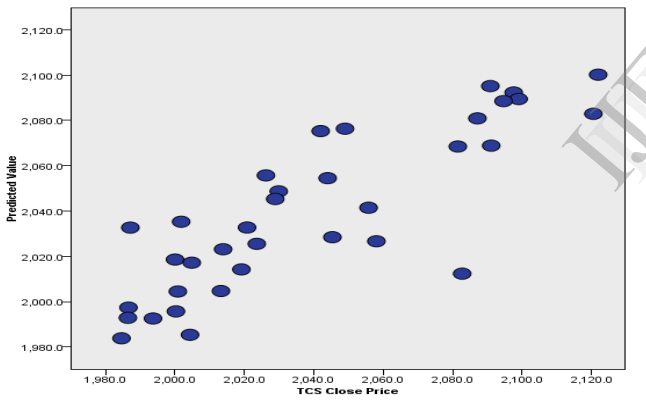
TABLE II PREDICTED VALUES OF COMPANY CLOSING PRICE USING ALM

Date	Company Close Price (Actual Value)	Predicted Value
03-Nov-13	2099.00	2089.44
05-Nov-13	2042.00	2075.28
06-Nov-13	2091.10	2068.85
……	……	…….
20-Dec-13	2120.55	2082.90

E.*Predicted by Observed*

The following figure represents the predicted closing price for a company and the actual closing price for the company.

Fig. 5 Predicted by Observed for a company Closing Price



In this research predict Residuals for predicted closing prices of the company. The smooth line represents the normal distribution. The closer the frequencies of the residuals are to this line, the closer the distribution of the residuals is to the normal distribution. The research also predict when overall sentiment is used to predict the closing price, the AICC is 227.410 and when both overall sentiment and Opening price are combined to predict the target variable, the AICC is decreased to 223.744 which is a better form. Thus, sentiment plays a vital role in predicting the closing price of company share price movements. From the above table, the information criterion is very less when opening price and overall sentiment combined together in predicting the value of a share. Thus, Sentiment is an indicator of future price movements of a share.

V.CONCLUSIONS

Sentiment analysis is also used to take importance decision which can be improve the business performance and their potential role as enabling technologies for other sub systems. In this study shows the company sentiment score and their stock price having strong relationship among them. If the overall sentiment for a firm is consecutively declining for three days, then it is bound to decrease for another two days. When we predict the closing prices of a firm by using its opening prices, the information criterion is very high, thus leads to low accuracy in prediction. When we predict the closing prices of a firm by using a combination of opening prices and overall sentiment, the information criterion is low, thus leads to higher accuracy in prediction. For further research we are going use ANN and to the predicted values of closing prices by Automatic Linear Modeling and Neural Networks MLP and RBF are compared and the variances are found to be equal which will lead for better accuracy for prediction

REFERENCES

[1]..Box George E., Jenkins Gwilym M.(1976), Time Series Analysis: Forecasting and Control. Holden-Day, INC. akland, halif. USA .

[2]..Carvalho Alexandre X., Tanner Martin A. (2005), Mixtures-of-Experts of Autoregressive Time Series: Asymptotic Normality and Model Specification ,IEEE Transactions on Neural Networks. Vol. 16, Pp. 39-56.

[3]. Hastie, T., Tibshirani, R & Friedman, J. (2011), The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

[4]. Kantardzic, M. (2011), Data mining: Concepts, models, methods, and algorithms (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

[5]. Leisch, F., & Dimitriadou, E. (2013). Machine learning benchmark problems. Viewed on 10.01.2014 from http://cran.r-project.org

[6].Lynda Pod Cast. (2011), How to use automatic linear modeling, Retrieved from http://www.youtube.com/watch?v=JIs4sMxAFg0

[7].Witten, I. H., Frank, E., & Hall, M. A. (2011), Data mining: Practical machine learning tools and techniques (3rd ed.). Burlington, MA: Elsevier Inc.

[8]. A.Pappu Rajan , S.P.Victor(2012) , Features and Challenges of web mining systems in emerging technology , International Journal of Current Research ,Vol.4, Issue ,07 ,pp.066-070, ,ISSN : 0975-833X

[9].Qiu M.,Yang L., Jiang J. (2013), Mining user relations from online discussions using sentiment analysis and probabilistic matrix factorization. Proceedings of NAACL-HLT, Atlanta, Georgia. 75-833X.

[10]. S. Kogan, D. Levin, B.R. Routledge, J.S. Sagi, and N.A. Smith (2009) , Predicting risk from financial reports with regression. In NAACL '09,pp. 272–280.

[11]. Areerat Songwattana (2008) ,Mining Web logs for Prediction in Prefetching and Caching, IEEE International Conference on Convergence and Hybrid Information Technology, pp. 1006-1011.

[13]. C.R.Kothari (2010), Research Medthodology Methods and Techniques, Second Edition, New Age International Publishers.

[14]. Linear Model(2013),View on 10.March.2014 from http://pic.dhe.ibm.com

[15]. H. Mao, S. Counts, and J. Bollen.(2011), Predicting financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint arXiv.

[16]. Paudel, J. and Laux, J. (2010) A Behavioral Approach To Stock Pricing. The Journal of Applied Business Research –Volume No. 26, Number 3, pp. 99---106

[17]. S. Charles, and L. Arockiam, (2012),Feature Selection: A New Perspective,International Journal of Data Mining and Knowledge Engineering, , Print: ISSN 0974-9683 & Online: ISSN 0974-9578.