EMP-SA: Ensemble Model based Market Prediction using Sentiment Analysis

Anuradha Yenkikar, Manish Bali, Narendra Babu

Abstract*: Predicting stock market trend is an extremely complicated task and calls for extensive study and insights into the context at hand. Primary requirement for any investor is to assess this trend to help invest for maximizing his returns. The advances in Machine learning and data analytics in particular have changed the way investors can approach this matter. Sentiment analysis or Opinion mining can be carried out by taking into consideration public sentiments regarding the stock market conditions and to understand the ups and down of this most volatile sector. In this paper, public sentiments from Twitter along with news feed related to the stock market conditions for predicting the nature of market is considered to analyse the stock market trend. The data is collected from twitter and various news sites to generate a gross sentiment score regarding the market. The gross sentiment score is used to find a correlation between market price and sentiments to train the proposed models for prediction using Linear and robustness regression techniques such as Ordinary Least squares (OLS), RANSAC, Theil-Sen estimator, Huber Regression and Ridge regression. Ensemble method is used to achieve reliable and better prediction accuracy instead of a single method. Ensemble method combines models and carries out majority voting among them to produce one final model to increase prediction accuracy. The obtained results reveal that public opinion does make a significant impact on market behaviour with the prediction accuracy between 65-91% depending on the dataset.

Index Terms*: Ensemble method, Machine Learning, Opinion mining, Sentiment Analysis.

I. INTRODUCTION

Price fluctuations are commonly associated with stock markets. The efficient market hypothesis is one of the most popular hypotheses in this domain which enumerates that factors that will have a significant impact on a company's stock value are news, current events and product releases. Financial market movements depend on them [2].

Prices on the stock market generally follow "a random walk pattern and cannot be predicted with more than 50% accuracy due to unpredictability in news and current events" [1]. With the advent of internet and easy access to social media, it has made it very easy for users to access or share their thoughts, opinions, ideas and emotions with others regarding any context.

And these emotions shared by users can be a valuable asset in understanding the general overview of public regarding any product, company, service etc. One such platform is twitter which is worldwide used by users to express their opinion or emotions. As per numbers available, there are more than 300 million users who actively tweet more than 500 million tweets on a daily basis. This is a valuable source of data for researchers since twitter can be considered as a sizable corpus [11]. Each tweet, which is 280 characters in length allows public to voice their opinions on any topic, albeit concisely. Tweets are a great source to exploit information and make predictions [12]. And twitter provides their API to access these tweets in real-time. Studies have presented twitter as an important source for conducting research and a strong tool for making forecasts. News websites on the other hand such as Reuters also provides their news archives which can be valuable for improving the accuracy of sentiment score calculation. A classic example of how sentiments can affect market can be taken from this account - on April 2013, a tweet posted by Associated Press (AP) mentioned "Breaking: Two explosions in the White House and Barak Obama injured". This false tweet from a hacked account of AP resulted in immediate drop in Dow Jones Industrial Average (DJIA) index. And after the tweet was declared fake, DJIA quickly recovered [6] [7]. This shows how news can affect markets drastically and is the basis of our consideration in this paper along with public sentiments from twitter, sample as shown in Table 1. The paper is outlined as follows. Section 2 captures related work and section 3 discusses the tools and techniques used in this research. Section 4 outlines our contribution followed by section 5 which discusses data collection and the pre-processing techniques adopted, including the part to find the sentiment score and correlation between market price and sentiment score. In section 6 results in terms of the mean squared error is presented. The paper is concluded in section 7 with the results obtained and discussion for future research.

II. RELATD WORK

Social media sites can be excellent source of information for sentiment analysis.

Revised Manuscript Received on 30 July 2019.
* Correspondence Author

Anuradha Yenkikar*, Research Scholar Department of Computer Science, Ramaiah University, Bangalore, (Karnataka), India.

Manish Bali, Adjunct Professor, Department of Computer Science, IT-ADT University Pune, India.

Narendra Babu, Associate Professor, Department of Computer Science Engineering, Ramaiah University, Bangalore, (Karnataka), India.

Retrieval Number: B2197078219/2019©BEIESP
DOI: 10.35940/ijrte.B2197.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

And this sentiment analysis can be used for many purposes such as designing business strategies and making business decision or for end users it can be a better way for knowing which products to be used or which are best for them.

Venkata Sasank Pagoluet et all [3] considered data from twitter for predicting the stock market using user sentiments. In their paper they represented data in the form of Word2Vec and N-grams to calculate the sentiments from the data extracted from Twitter. Supervised machine learning principles are applied to the tweets to find the correlation between stock price and public sentiments. Tejas Mankar et all [4] used two classifiers: Naïve Bayes and SVM for sentiment analysis. Here data is collected with the help of Twitter Search API which enabled them to fine tune their queries for data collection. Data is collected in JSON format which included variety of information and the paper focuses on only Time and Tweet Text. The tweets extracted are pre-processed to remove noise. The processed tweets are used for feature extraction using Naive Bayes and SVM classifier. Each tweet is processed to create feature matrix by unigram technique. Bing Yang et all [5] proposed a way for predicting stock market using ensemble method for deep neural network. In their paper, multiple multi-layer deep neural networks are trained on historical Chinese stock market indices. For improving the training efficiency, sigmoid function is replaced by Leaky Rectified Linear Unit (LReLU). Along with optimization algorithms, back propagation algorithm and Adman algorithm are used to accelerate the training and speed. Finite set of such networks are trained and ensemble of these networks is constructed with the help of a bagging technique. The accuracy prediction for high and low was overall good and up to 75% (approx.) but prediction on close price was unsatisfactory. Sunil Kumar Khatri and Ayush Srivastava [8] extracted sentiments from Twitter and stocktwits. And the data collected is classified in four categories: Happy, up, down and rejected. In their paper they used Feed Forward Neural Network. The network is trained using 75% of data, 15% was used for testing and remaining 10% was used for validation. Buche Arti et all [10] carried out a survey of different methods that have an impact on prediction of stock markets using financial news. It also discussed and presented a general procedural flow of the methods.

Table 1: Sample tweets on companies

Happy to join @Nvidia as a Solution Architect
Google revenues on the rise since 2017, augers well for tech stocks
Pathetic! Service levels @Dell are not worth looking at them again
@Apple is a great company to work for but it better reduce its pricing so more people can try their products

III. TOOLS AND TECHNIQUES

We have used Ordinary Least Squares (OLS) or Linear Regression that fits a linear model to minimize the residual sum of squares between the observed responses in the dataset, and the responses predicted by the linear approximation. However, since the data may consist of outliers that may ultimately effect prediction accuracy, we have also used Robustness Regression which fits a regression model even if the data is corrupt i.e. having error or outliers in the model. Scikit-Learn provide the following four Robust Regression Estimators - RANSAC, Theil-Sen, Huber Regression and Ridge regression which are briefly described below.

1. RANSAC

RANdomSAmple Consensus employs a learning technique to estimate parameters of a model by random sampling of observed data. Since a dataset contains both inliers and outliers, the algorithm uses a voting scheme to find the most optimally fitting result.

2. Theil-Sen Estimator

Theil-Sen Regress or fits a line robustly to sample points in the plane by choosing the median of slopes of all lines passing through pairs of points. Its robustness decreases in high dimension and it starts to mimic OLS.

3. Huber Regression

In Huber regression, classification of a sample as an inlier happens when the absolute error of the sample is less than a certain threshold. However, it differs from the above two estimators in that it does not ignore the effect of outliers and it does this by giving them a lesser weight.

4. Ridge Regression

Multiple regression data that suffers from multi-collinearity can be analysed using Ridge regression. Generally, least squares estimate are used which are unbiased when multi-collinearity occurs. But due to large variance, it may not provide the right values. Therefore, by adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is expected that this would lead to more reliable estimates.

IV. OUR CONTRIBUTION

Sentiment analysis can be carried out by considering public sentiments regarding the market conditions and to understand the ups and down of this volatile sector. However, in this paper we have considered public sentiments along with news feed related to the market conditions to predict the nature of market. We have used a generalized linear model viz. Ordinary Least squares and Robustness Regressors namely Theil-Sen, RANSAC, Huber Regressor and Ridge regression for calculating the sentiment from word list in standalone mode. We then find a correlation between the calculated sentiments by comparing it with stock market price to predict the nature of market. Thus, key contribution of this research is the development of a sentiment/opinion analyser (used to classify the sentiments in tweets extracted) using a novel approach that calculates a correlation between market price and sentiment score to train our models - both in standalone mode and as an ensemble model with majority voting to improve the stock market prediction with more than 50% accuracy to negate the hypothesis proposed in [1].

V. DATA COLLECTION, PRE-PROCESSING AND SENTIMENT ANALYSIS

In this research, we have extracted tweets from twitter using twitter API and along with it we also collected the historical twitter archive data for one month.

To improve the accuracy of sentiment score we considered collecting news data from news websites by web crawling. To collect news data, we scrapped different news websites with the help of beautiful soup library. The data collected from twitter is in JavaScript Object Notation (JSON) format so we extract the required information by applying regular data processing methods. In case of news websites, we directly captured html website in an html format and carried out pre-processing with some adjustment to extract the required information from the webpage. The complete process flow chart is shown in Fig 1.

1. Acquiring data in raw format

1.1. Fetching through Twitter API

Twitter is a social networking platform with user base of more than 100 million active users. Users share their views, opinion in 280 characters on the platform. These opinions can be used as valuable information for sentiments analysis. Twitter API helps developers to fetch these tweets in JavaScript Object Notation (JSON) form which consist of all the information such as user name, tweets associated with the user, location and time. In our work, we have focused on only three parameters - time, location and tweet posted by users. Even though Twitter provides their API, it has many restrictions regarding fetching of tweets or request limits so we have used archive of historical tweets for one month in training our models.

1.2. Fetching from News websites

News websites can be a great source of information regarding market condition. News can be potentially used to improve our sentiment analysis accuracy as they provide generic public view about market. To get news from websites, we developed a script to scrap websites using beautiful soup library. To scarp website, we fetched the news website in html format, cleaned complete html document and focused our goal towards extracting content from paragraph tags (<p></p>) as they contain most of the essential information required.

2. Extracting relevant information: Data Processing

2.1. Text Processing

The text data gathered may consist of data such as alphanumeric words, numeric text, Unicode characters etc. Such type of data is considered noisy data and it's required to remove this noise in order to keep our sentiment score as accurate as possible. We created regular expression to remove such type of anomalies from text data.

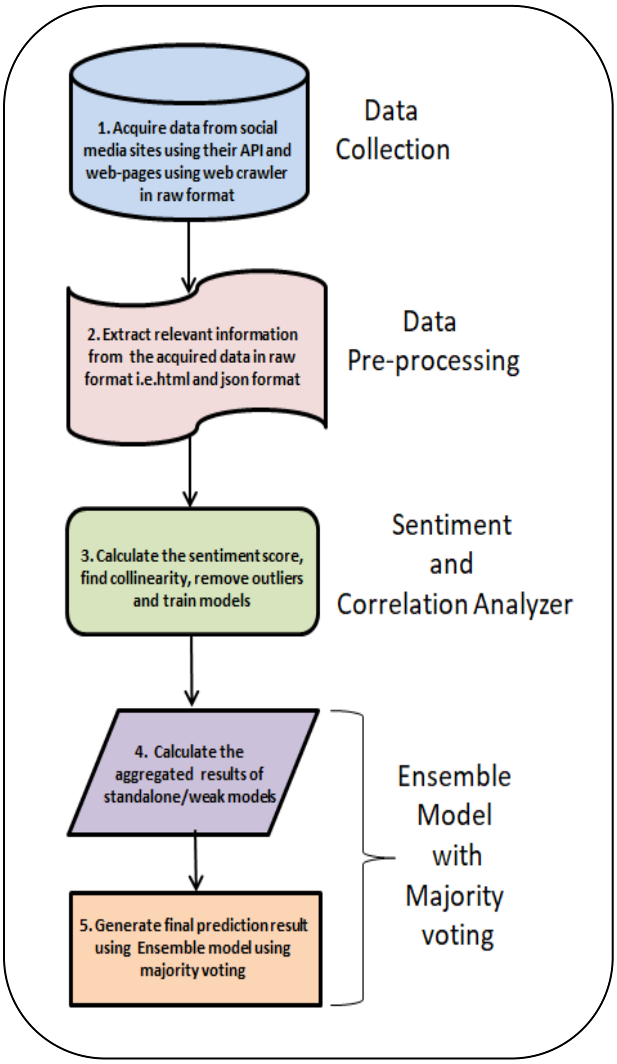
1. Acquire data from social media sites using their API and web-pages using web crawler in raw format
Data Collection
2. Extract relevant information from the acquired data in raw format i.e.html and json format
Data Pre-processing
3. Calculate the sentiment score, find collinearity, remove outliers and train models
Sentiment and Correlation Analyzer
4. Calculate the aggregated results of standalone/weak models
5. Generate final prediction result using Ensemble model using majority voting
Ensemble Model with Majority voting

Fig. 1 Process flow chart of proposed methodology

2.2. Removing stop words

Stop words are those which are frequently occurring in a sentence and doesn't have any meaning on its own e.g. at, the, on etc. Such types are useless for analysing sentiment of text data. We used natural language toolkit (NLTK) which consist of dictionary of such stop words. We compared each word in text with stop words dictionary and then removed it from our text data.

2.3. Tokenization

The text data is converted into list of words; this word list is later on used to calculate sentiment score using NLTK.

2.4. Storing in comma separate value(.csv) format

The cleaned text data is stored in .csv format along with the date and time of the context when it was first published. Storing data is .csv format makes task easier for further processing.

3. Analyzing sentiment and finding correlation

We used Natural Language Tool Kit (NLTK) for first calculating the sentiment from word list.

The calculated sentiment (positive, negative, compound and difference between positive and negative) is compared with the stock market price to find the correlation between them as shown in Fig 3. Fig 2 depicts the association between sentiments during a selected time period and Fig 3 shows a relationship between the market during a similar selected time period.

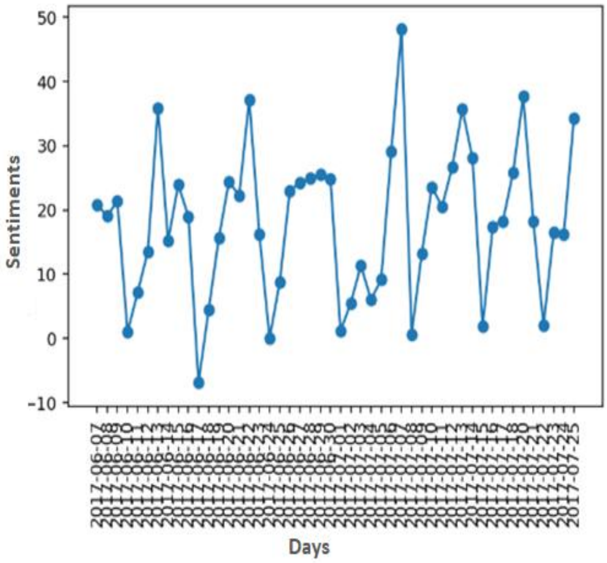

Fig. 2 Sentiments vs. selected time period

They are then combined to find the correlation between selected time period, stock market value and sentiments. Fig 4 shows the correlation between selected period, market value and compound sentiment. Similarly Fig 5, 6 and 7 depict this correlation for positive, negative and difference between positive and negative sentiments respectively.

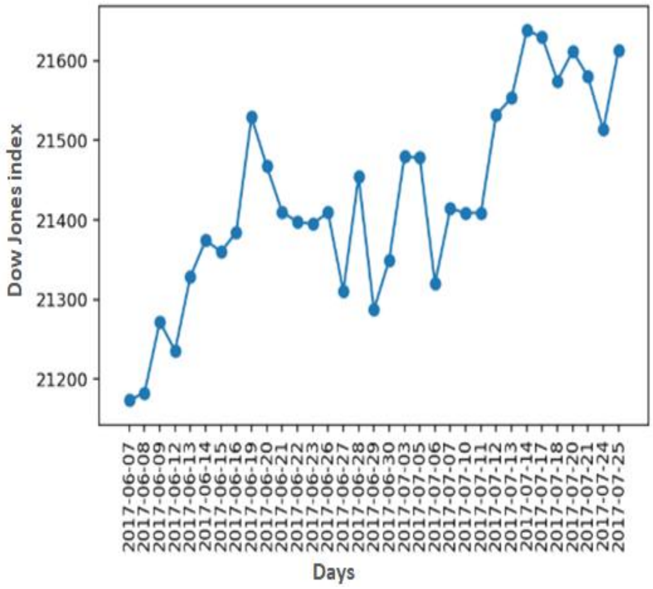

Fig. 3 Market value vs. selected time period

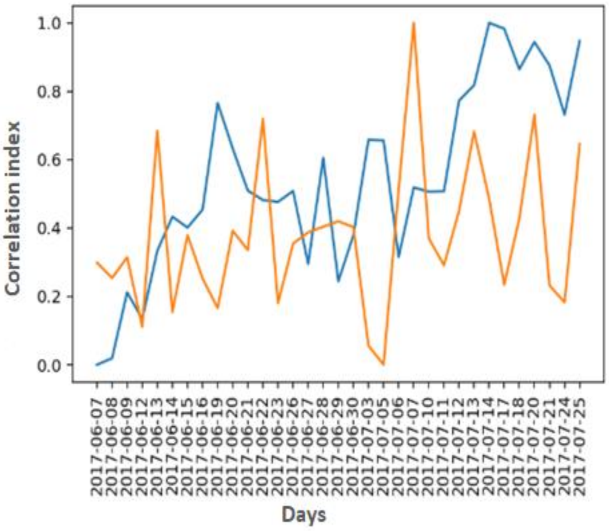

Fig. 4 Compound sentiment, market price and time period correlation

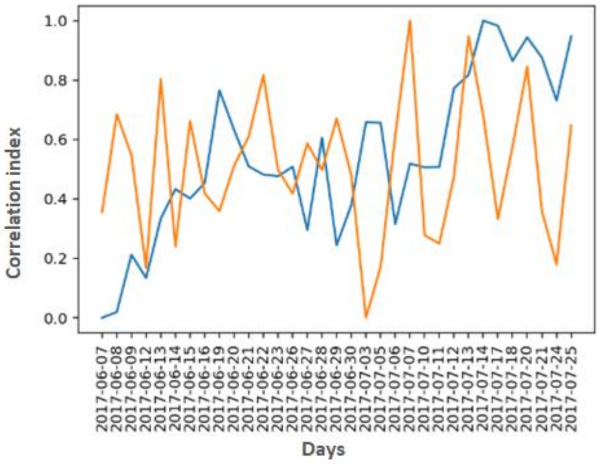

Fig. 5 Positive sentiment, market price and time period correlation

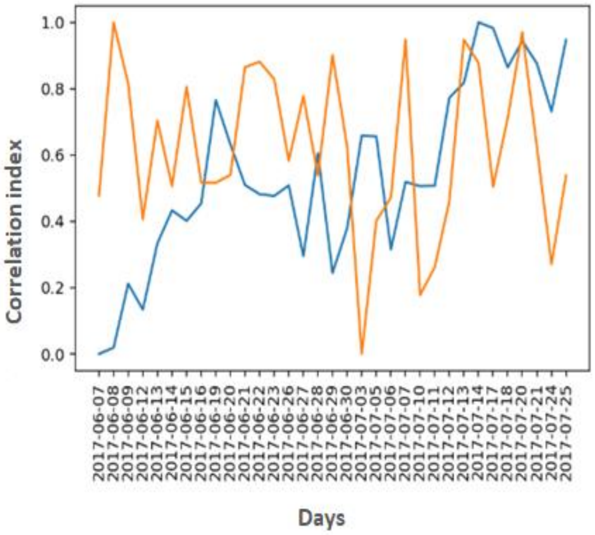

Fig. 6 Negative sentiment, market price and time period correlation

Retrieval Number: B2197078219/2019©BEIESP
DOI: 10.35940/ijrte.B2197.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

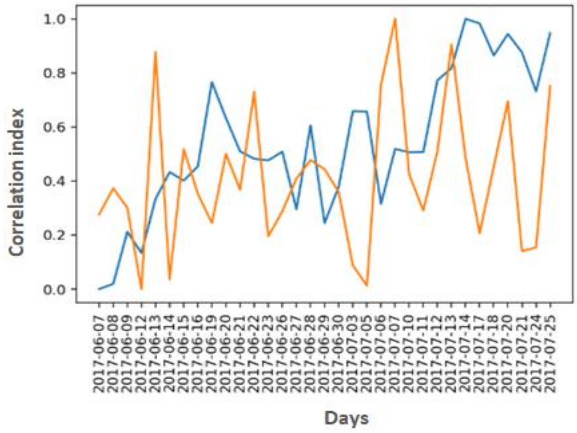

Fig. 7 Positive-negative difference sentiment, market price and time period correlation

We find that when the sentiments (positive/negative) are high then market is most likely to rise/collapse. In other words, the sentiments of each day are compared with the market price of each individual day to find the correlation between them to find how prices have fluctuated with respect to sentiments of that day. We used this relationship between market price and sentiment score to train our models.

4. Calculating aggregated results of the weak models

The aggregated results generated by all the trained models are estimated in a standalone mode. We basically find the mean square error. But this is found after we have removed any outliers which might affect the results and hence the accuracy.

5. Generate prediction results on the basis of voting

5.1. Ensemble Method

Ensembling depends on the presumption that diverse models trained autonomously are probably going to be useful for various reasons: each model looks at marginally different parts of the data to make predictions, getting some portion of reality however not every last bit of it.” The popular methods are mixture of experts, majority voting ensemble, boosting, bagging and stacking. Current work uses Majority voting. Stacking is based on a heterogeneous set of weak learners. Every model is trained autonomously and final choice is made by a majority vote, averaging the result.

5.2. Results prediction

There are five (05) weak learners considered in this study. The result generated by all the weak learners is combined together using ensemble method and a majority poll is conducted between them to generate the average result. The below scenarios emerge:

Scenario 1: Out of 5 models, 3 are predicting market will rise and remaining 2 are predicting market will collapse. This basically means that the market is more likely to rise as per the majority aggregated result.

Scenario 2: Out of 5 models, 2 models are predicting upward trend and 2 downwards. In this case, the 5th one is the average of other 4 models and considered. This is the reason for using odd number of models so that a tie seldom arises.

The ensemble system model for training, testing and voting is shown in Fig 8.

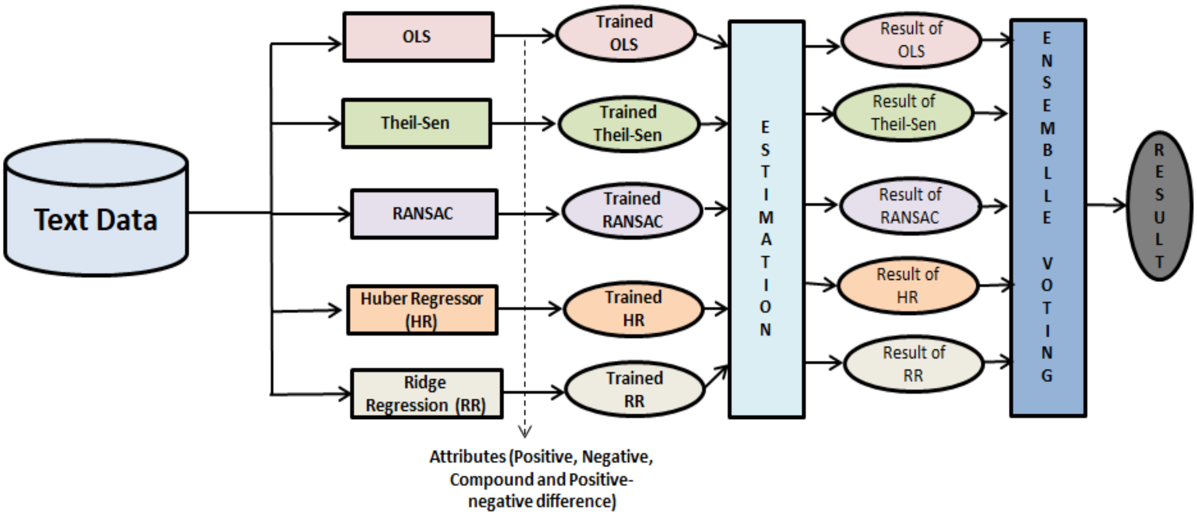

Fig. 8 Ensemble model for training, testing and voting

VI. RESULTS AND DISCUSSION

We have trained our models with consideration of four different features or attributes: Positive, Negative, Compound and difference of positive and negative. Fig 9 illustrates Ensemble model result. The scatter plot depicts test and training data vs. combined prediction average. The results we received are after training & testing different models on above attributes.

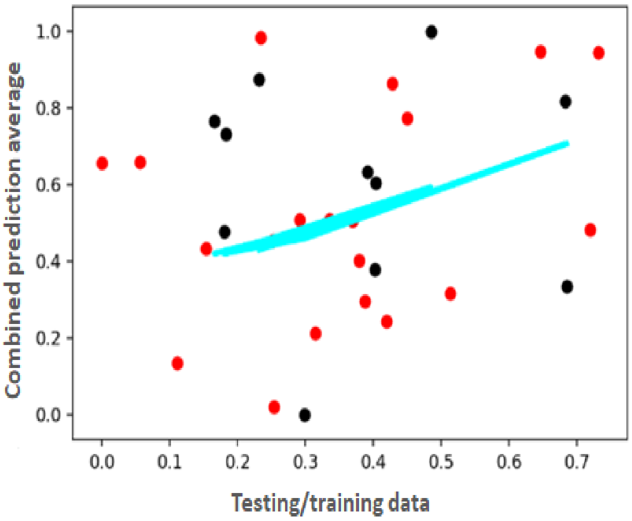

Fig. 9 Ensemble model scatter plot on all attributes

Mean squared error for each model is calculated using the formula:

$$\text{Error} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$$

where n = number of data points; Y_i represents observed value and $\widehat{Y_i}$ represents predicted value.

Table 2 depicts the error rate of individual model corresponding to each attribute that we have considered.

Table 2: Mean squared error in models

Model	Positive	Negative	Compound	Pos_Neg diff.
OLS	0.080	0.097	0.095	0.122
Theil-Sen	0.086	0.106	0.102	0.218
RANSAC	0.204	0.094	0.130	0.471
HuberReg	0.090	0.117	0.102	0.127
Ridge Reg	0.081	0.090	0.103	0.126
Ensemble	0.09 and 65%-91% accuracy			

The results prove beyond doubt that there exists a strong relationship between sentiments shared on twitter and stock prices the next day. Fig 10, 11, 12 and 13 shows the model errors between the normalized sentiment and the normalized market value.

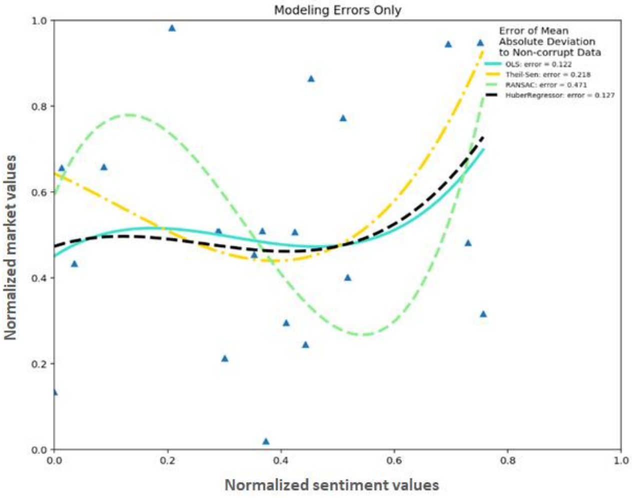

Fig. 10 Positive sentiment vs. normalized market price

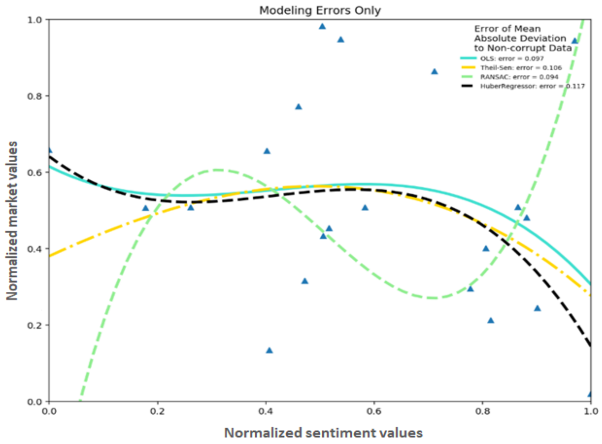

Fig. 11 Negative sentiment vs. normalized market price

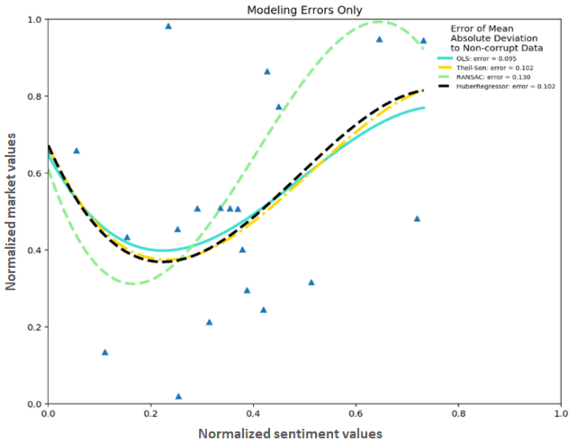

Fig. 12 Compound sentiment vs. normalized market price

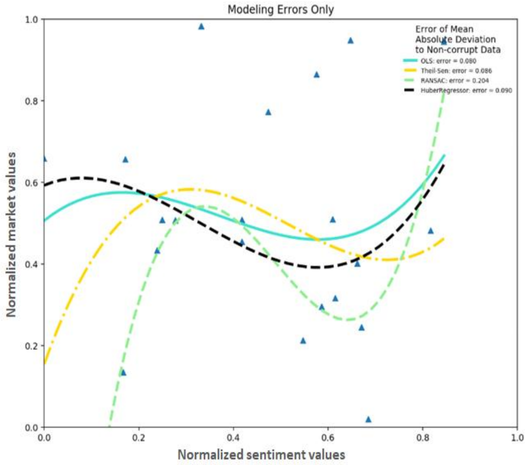

Fig. 13 Positive-Negative diff. vs. normalized market price

Basically to improve accuracy of the results, we carried out the following novel approaches:

- We removed any of the outliers which might affect the results of the models
- We combined the results of all the models in an ensemble model with majority voting to get a minimum error rate of 0.09. By combining the results of all the models for each of the attributes we achieved accuracy in the range of 65-91% depending on the news feed dataset used.

The accuracy is higher than the research cited and negates the hypothesis put forward by authors that "stock market prices generally follow a random walk pattern and cannot be predicted with more than 50% accuracy" [1].

VII. CONCLUSION

Considering the significance of public opinion about a company via tweets, it has been proved beyond doubt that a deep relation exists between twitter sentiments and fluctuations in stock prices of the company. It is found that Ensemble models are promising techniques for predicting the stock market to achieve higher accuracy and can be explored further with increasing the size of ensemble. Also our assumption that positive public sentiment in twitter and news feed about a company reflects in its stock price is reinforced by the results achieved. In this work, we have considered only twitter data for analysing people's sentiment which may be prejudiced because not all the people who trade in stocks share their sentiments on twitter. People associated with stock market normally share their views on Stocktwits [13]. Current research can be extended further by using this stocktwits data. Due to restriction of twitter API, the gathering of real time data is not possible for academic developers as we need to purchase the enterprise edition of API. Also we used only one month's data to train the model, which is very less to train a sentiment analyser. We can better the model performance by increasing the size of training datasets. Since more than one model is trained in ensemble learning, it requires high computation power and time for training the models. In future research, the computational performance can be improved by using GPU parallel or a distributed architecture.

REFERENCES

1. Qian, Bo, Rasheed, Khaled, Stock market prediction with multiple classifiers, Applied Intelligence 26 (February (1)) (2007) 2533, http://dx.doi.org/10.1007/s10489-006-0001-7.
2. E.F. Fama, The behavior of stock-market prices, The Journal of Business 38 (1) (1965) 34105, http://dx.doi.org/10.2307/2350752
3. Pagolu, V.S., Reddy, K.N., Panda, G. and Majhi, B., 2016, October. Sentiment analysis of Twitter data for predicting stock market movements. In Signal Processing, Communication, Power and Embedded System (SCOPES), 2016 International Conference on (pp. 1345-1350).IEEE.
4. Mankar, T., Hotchandani, T., Madhwani, M., Chidrawar, A. and Lifna, C.S., 2018, January. Stock Market Prediction based on Social Sentiments using Machine Learning. In 2018
5. International Conference on Smart City and Emerging Technology (ICSCET) (pp. 1-3).IEEE.
6. Yang, B., Gong, Z.J. and Yang, W., 2017, July. Stock market index prediction using deep neural network ensemble. In Control Conference (CCC), 2017 36th Chinese (pp. 3882-3887).IEEE.
7. Batra, R. and Daudpota, S.M., 2018, March. Integrating StockTwits with sentiment analysis for better prediction of stock price movement. In Computing, Mathematics and Engineering Technologies (iCoMET), 2018 International Conference on (pp. 1-5).IEEE.
8. Megahed, F.M. and Jones-Farmer, L.A., 2015. Statistical perspectives on "big data". In Frontiers in Statistical Quality Control 11 (pp. 29-47). Springer, Cham.
9. Khatri, S.K. and Srivastava, A., 2016, September. Using sentimental analysis in prediction of stock market investment. In Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2016 5th International Conference on (pp. 566-569). IEEE.
10. Chakraborty, P., Pria, U.S., Rony, M.R.A.H. and Majumdar, M.A., 2017, September. Predicting stock movement using sentiment analysis of Twitter feed. In Informatics, Electronics and Vision & 2017 7th International Symposium In Computational Medical and Health Technology (ICIEV-ISCMHT), 2017 6th International Conference on (pp. 1 6).IEEE.

11. BucheArti, Dr. M. B. Chandak B.M., AnStock Market Prediction using Text Opinion Mining: A Survey, International Journal of Advanced Research in Computer Science and Software Engineering Vol. 6, Issue 6, June 2016.
12. B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology.
13. A. Pak and P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, in Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010, pp. 13201326
14. http://stocktwits.com/home

AUTHORS PROFILE

Anuradha Yenkikar is a Research Scholar in Computer Science at Ramaiah University, Bangalore, Karnataka, India with research interests in Artificial Intelligence, Machine learning and Parallel Computing. With 7 years of academic experience, She holds B.E. and M.E. degrees and is an Assistant Professor in the Department of Computer Engineering at Zeal College of Engineering and Research, Pune.

Manish Bali has 27+ years of experience in Enterprise Computing, IT and technology services domain having held various senior level positions in leading Indian and MNC cos. He holds a B.E. in Electronics from Bangalore University and an MBA. He is currently pursuing M.Tech in Machine Learning & Intelligent Systems from RUAS. He is an Adjunct Professor in Computer Science at MIT-ADT University Pune, Chitkara University Chandigarh and Member-Board of Studies at various Engineering colleges across the country helping to bridge the Industry-Academia divide. He is on Advisory board of start-ups in the field of HPC/AI/ML/DL/IoT. He is an active member of various Industry & Academic bodies like ACM and HiPC

Narendra Babu presently working as Associate Professor at Ramaiah University, Bangalore, Karnataka, India. He has completed his PhD in the Computer Science Engineering from JNT University, Anantapuram. His research interests include time series data analysis and mining, soft computing. He has 19 publications in reputed international journals and conferences..