Market Prediction using Sentiment Analysis on Real Time Data Streams

Prinkle Vilas Shete, Rohan Vijay Gokhale, Sanket Hanmant Masudge, Maya P. Shelke
Department of Computer Engineering,
Zeal College of Engineering and Research, Pune, Maharashtra, India

Abstract: Knowing and understanding the nature of market is difficult task and requires thorough study regarding the context. One of the easiest way is to consider public opinion regarding the market situation to understand the ups and downs in the sector. This paper considers public sentiments along with news related with the market condition to predict the nature of market. Our work consists of data collection from twitter and various news sites to generate overall sentiment regarding the market and with the help of that finding the correlation between market price and sentiments to train our models for prediction. To further improve accuracy of prediction, the proposed approach considered ensemble learning method in which it trained more than one model and generated the poll among them to precise results.

Keywords **- Tweets; Sentiment Analysis; Ensemble Learning; Opinion Mining; Website Scrapping; Machine Learning.**

I. INTRODUCTION

Introduction of internet and easy access to online activities made users to access or share their thoughts, opinions ideas and emotions with others regarding any context. These emotions shared by users can be a valuable asset in understanding the general overview of public. One such platform is twitter which is worldwide used by users to express their opinion or emotions. There are more than 300 million users who actively tweets more than 500 million tweets on daily basis. Twitter provides APIs (Application Programming Interfaces) to access these tweets in real-time. News websites such as Reuters also provides news archives which can be valuable for improving the accuracy of sentiment score calculation.

A classic example of how sentiments can affect market: Fake tweet "Breaking: Two explosions in the White House and Barak Obama injured" from hacked account of *Associate Press* (AP) on April 23, 2013 resulted in immediate drop in Dow Jones Industrial Average (DJIA). DJIA quickly recovered after the tweet was declared fake [4][5]. This shows how news can affect market drastically.

This paper uses one-month twitter data consisting of more than 1 million tweets and news data. Data from twitter can be collected with the twitter API which provided data in JavaScript Object Notation (JSON) format. This JSON format data is cleaned with regular cleaning process with some adjustments to get plain text as much as possible. For collecting news, news websites are scrapped and cleaning process is applied to get relevant context of news for proposed work.

In most of the existing research, sentiments are categorized in many types such as: Positive, negative etc while the proposed work considers only one of the two sentiments and calculates difference between positive (or negative) sentiment score on subsequent days. Thus it reduces task of classifying sentiment into different categories. Finally proposed work analyzes the correlation between sentiment score and market nature and uses it to train different models to implement ensemble learning method.

II. LITERATURE SURVEY

Social media sites can be excellent source of information for sentiment analysis. Further, it can be used for many purposes such as designing business strategies, making business decision, provide users with better knowledge regarding products to assist them with decision making regarding product choice.

Venkata et al [1] considered data from twitter for predicting the stock market using user sentiments. In their paper, authors 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 al [2] used two classifiers: Naïve Bayes and SVM for sentiment analysis. Here data is collected with the help of Twitter Search API which enabled authors to fine tune their queries for data collection. Data is collected in JSON format which included variety of information. In this paper, authors focused on only Time and Tweet text. The extracted tweets are pre-processed to remove noise and then used for feature extraction using Naive Bayes and SVM classifier. Each tweet is processed to create feature matrix by unigram technique.

Bing Yang et al [3] proposed approach for predicting stock market using ensemble method for deep neural network. In this 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, Backpropagation algorithm and Adman algorithm is used to accelerate the training and speed. Finite set of such networks are trained and ensemble of these networks are constructed with the help of bagging technique. Though accuracy prediction of approach in this paper for high and low was up to 75% but prediction on close price was unsatisfactory.

S K Khatri et al [6] extracted sentiments from Twitter and Stock Twist. Authors classified collected data in four categories: Happy, up, down and rejected. In this paper, authors used Feed Forward Neural Network to train and test data. Network is trained using 75% of data, tested using 15% of data remaining 10% was used for validation.

III. PROPOSED METHODOLOGY

Proposed paper uses tweets, historical twitter archive data of one month and news data. Tweets are extracted from twitter using twitter API while news data is collected from news websites by web crawling. News data is scrapped from different news websites with the help of beautiful soup library. Data collected from twitter was in JavaScript Object Notation (JSON) format so required information is extracted by applying regular data processing methods. In case of news websites, html website in html format is directly captured and cleaning process is applied with some adjustment to extract the require information from the webpage.

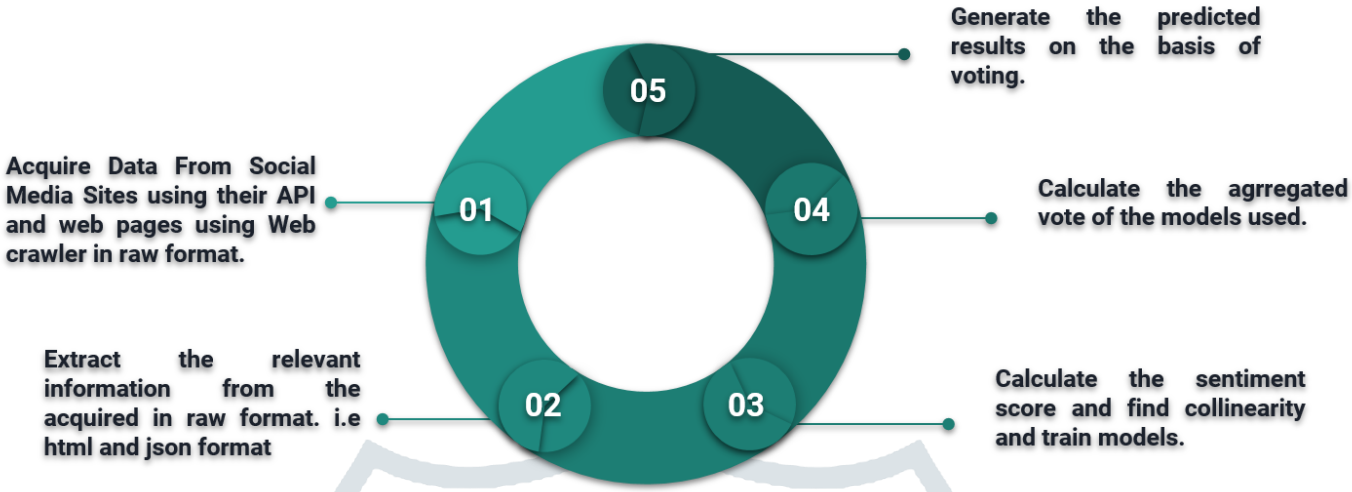


Figure 1: Process flow model of proposed methodology

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 shars their views, opinion in 256 characters on the platform. These opinions can be used as a valuable information for sentiments analysis. Twitter provides API for 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. Proposed paper focused on three parameters: time, location and tweet posted by users. Even though Twitter provides API, it has many restrictions regarding fetching of tweets or request limits so we used archive of one month historical tweets for 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 as they provide generic public view about market. To get news from websites we created script to scrap websites using beautifulsoup library. To scarp website, we fetched the news website in html format and cleaned whole html document and focused our goal towards extracting content from paragraph tags( ) as they contain most of the essential information required.

2. Extracting relevant information: Data Preprocessing

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 required to remove noise in order to keep sentiment score as accurate as possible. Along with this, proposed approach also creates regular expression to remove such type of anomalies from text data.

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 analyzing sentiment of text data and hence need to be removed. So 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 which is later used to calculate sentiment score using NLTK.

2.4 Storing in comma separated 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 made task easier for further process flow.

3. Analyzing sentiment and finding correlation

Natural Language Tool Kit (NLTK) is used for calculating the sentiment from word list. The calculated sentiment is compared with the stock market price to find the correlation between them. We concluded that when the sentiments (positive/negative) are high then market is most likely to rise/collapse. In other words the sentiments of each day is compared with the market price of each individual day to find the correlation between them to find how prices are fluctuated with respect to sentiments of that day.

We used this relationship between market price and sentiment score to train our models. And then aggregated result is generated by conducting poll between the trained models.

4. Calculating aggregated result

The result generated by all the trained models are combined together or in other words poll is conducted between them to generate more precise result. It means for a particular day if sentiments are positive and out of 5 models 3 are predicting market will rise and remaining 2 are predicting market will collapse that time it means the market is more likely to rise as per the aggregated result generated.

5. Calculating aggregated result

The result generated by all the trained models are combined together or in other words poll is conducted between them to generate more precise result. It means for a particular day if sentiments are positive and out of 5 models 3 are predicting market will rise and remaining 2 are predicting market will collapse that time it means the market is more likely to rise as per the aggregated result generated.

IV. RESULTS

Initially in our work we analyzed the correlation between the four sentiment attributes: Positive, Negative, Compound and Positive negative difference and Market Price the below table shows the correlation matrix:

Table 2: Correlation Matrix

	Price(DJI)	Compound	Negative	Positive	Pos_Neg_Diff
Price(DJI)	1.000000	0.150638	-0.076796	0.027312	0.116545
Compound	0.150638	1.000000	0.550880	0.873587	0.949286
Negative	-0.076796	0.550880	1.000000	0.855912	0.504821
Positive	0.027312	0.873587	0.855912	1.000000	0.878474
Pos_Neg_Diff	0.116545	0.949286	0.504821	0.878474	1.000000

Instead of selecting most correlating features to train our models, we trained our models with consideration of four different features or attributes: Positive, Negative, Compound and difference of positive and negative. The table below illustrates the results we got after training different models on those 5 attributes.

Table 1: Mean squared error of models

Model	Positive	Negative	Compound	Pos_Neg Difference
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
HuberRegressor	0.090	0.117	0.102	0.127

Mean squared error for each model is calculated using the below 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 1 depicts the error rate of each model corresponding to each attributes that we have considered for ensemble learning. The graphs below show comparison of models with each other.

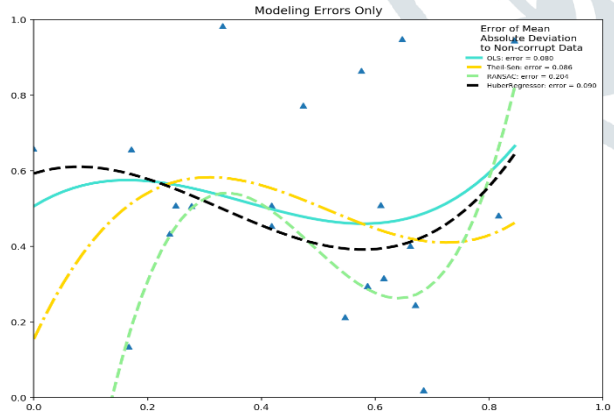


Fig 2: Positive sentiment Vs Price

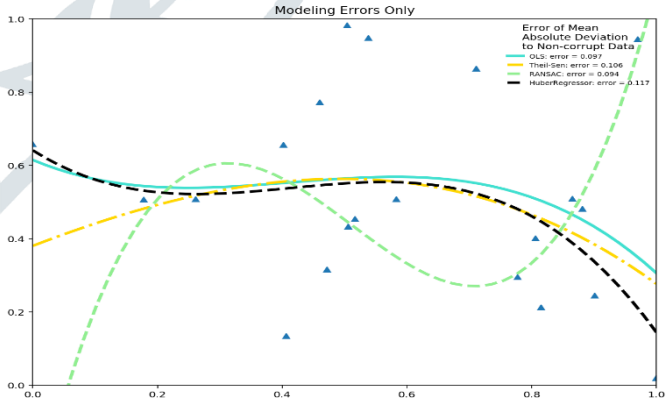


Fig 3: Negative sentiment Vs Price

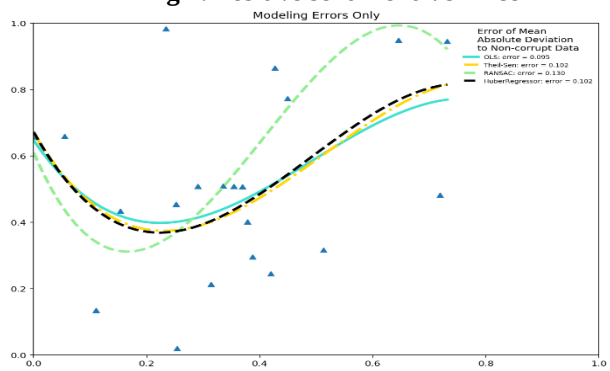
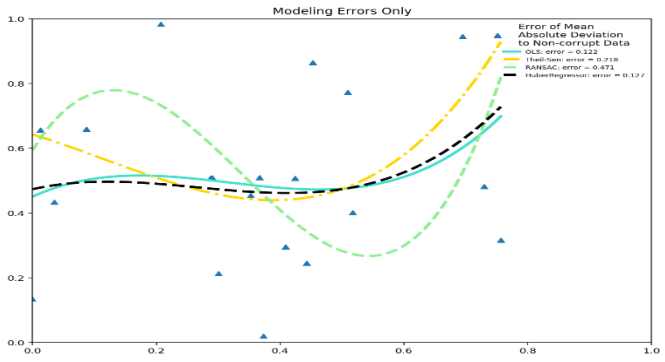


Fig 4: Compound Vs Price



Fid 5: Positive Negative Difference Vs Price

To furthermore improve accuracy of the result we combined the results of all the models to get minimum error rate of 0.09. By combining the results of all the models for each of the attributes we achieved accuracy in between of 65% to 91% depending on the dataset used. To furthermore improve the accuracy, we removed any of the outliers which might affect the results of the models.

V. Conclusion

In early research predictions are made using raw numerical data or by finding patterns of fluctuation in prices. The user sentiments were not considered for prediction of market nature. In our work we combined both numerical data and public opinion to predict the behavior of market. And we come to the conclusion that public opinion dose affects the market behavior.

We gathered one-month twitter data along with news data we helped us to improve overall sentiment calculation and finding the correlation with price fluctuation. To further improve the accuracy of our result we used ensemble learning method.

VI. Future Scope And Limitations

Limitation:

Due to restriction of twitter api the gathering of real time data is not possible for normal developers as we need to make purchase for enterprise edition of api. Since more than one models are trained in ensemble learning it required high computation power and time for training the models.

Future Scope:

This work can be further improved from business perspective as well as from customer's perspective. If certain business decision. The performance of computation can be improved by using parallel or distributed architecture.

The accuracy can be improved if more detailed public emotions are extracted in our research paper we extracted emotions as positive or negative if it is categorized furthermore in attributes as anger, fear, happy it can contribute towards the accuracy of the model. Industrial quarterly report are published it can be considered to further improve results.

Deep learning ensemble techniques can be used to improve results.

.

References

[1] 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.

[2] 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 International Conference on Smart City and Emerging Technology (ICSCET) (pp. 1-3). IEEE.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.