ISSN NO. 2456-3129

IJEPAS *International Journal of Engineering, Pure and Applied Sciences, Vol. 2, No. 4, Dec-2017*

An Exploration on the Prediction of Stock Market Movement using Sentiment Analysis Approach

Geetinder Saini[1], Punita[2]

Asst. Prof., Department of CSE, IEC University, Baddi, H.P, India
Research scholar, Dept of CSE, NITTTR, Chandigarh, India
***Email:** contact2geetinder@gmail.com, contact2punita@gmail.com*

Abstract- Stock Market Movement Prediction is one of the latest research topics in both the aspects of research and business. Due to advancements in social media, public opinion and sentiments are very easily available on internet. These opinions and sentiments can be extracted with the API information of that particular dataset and further make in use for the application of stock market prediction. The term stock is generally referred as the partial ownership of any company in terms of shares. Sentiments can be extracted from the websites like twitter, Finance market, debate, and many more dependent upon the kind of product for which prediction need to make. Generally referred site is twitter as it contains public tweets in small and efficient keywords for different kind of products, globe, human interactions etc. In this paper, we are presenting an exploration for the prediction of stock market movement based on online available public sentiments and opinion. Also a comparative analysis for the different concepts with their advantages and drawbacks are presented.

Keywords— Stock Market, Online Dataset, Public Opinion, Sentiment Analysis

1. INTRODUCTION

With the advancement in technology and digitization, people are becoming so active to share their reviews/comments as an opinion/feedback for the products. There is availability of various online social networking websites available to share the reviews regarding anything they take interest in. Generally, used websites are Facebook, Twitter, Tumblr, Weibo etc. With these websites, user can express their sentiments. Further, these sentiments are analyzed for the prediction of any product as their future aspects [1]. This analysis process is known as sentiment analysis.

In this paper, this sentiment analysis approach is considered for the stock market movement prediction. Stocks can be defined as the partial ownership of any company. We can also define it in terms of share of the company. Different companies define different price of their shares. The overall increase or decrease in stocks prices is depends upon the movement of stock market. The term "Stock Market" is commonly used to encompass both the physical location for buying and selling stocks as well as the overall activity of the market within a certain country [2]. Here, the work of different research authors has been presented with their advantages and drawbacks.

Rest of the paper is organized in the following manner. Section II describes the basic concept and levels of sentiment analysis. Section III presents the related work of sentiment analysis for stock market prediction. Section IV discusses about the research gap from the overall presented work and Section V concludes the paper.

2. SENTIMENT ANALYSIS

Sentiment Analysis can be defined as the process to find the public reviews for a specific product. Sometimes it is also referred as the opinion mining. Sentiments can be positive, negative or neutral dependent about the public review for that particular product. As a feedback for the product or stock, company owner can upgrade the quality of products [3]. The basic process of Sentiment analysis is shown in figure 1.

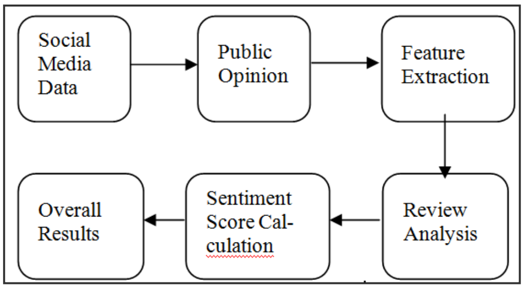

Figure 1: Process of Sentiment Analysis

Sentiment analysis is a two way approach consisting of an offline and an online process. In the offline process, a labelleddataset of public reviews is pre-processed in order to extract useful features, and then it is used to train a classifier. After training, the classification model is stored on secondary storage, in order to be loaded and used in the online phase.In the online process, public reviews from the website stream are received, pre-processed and features are extracted. Then, each review (more accurately its representation using features) is given as input to the classifier, which has already loaded the classification model, and is able to predict the sentiment of the review/comment.

To mine sentiments, the reviews collected can be analysed at three levels.

2.1 Document Level Sentiment Analysis

Document level SA is concerned with the overall classification of opinion conveyed by the author in the total document as positive, neutral or negative. The presumption is that the entire document focuses on one particular entity and comprises opinion from one opinion holder. The challenge in the document level classification is that the entire sentence in a document may not be relevant in expressing the opinion about an entity. Hence subjectivity/objectivity classification is significant in this kind of classification. Both supervised as well as unsupervised learning methods may be utilized for the document level classification. The advantage of document level sentiment analysis is that we can get a total polarity of opinion text regarding a specific entity from a document

2.2 Sentence Level Sentiment Analysis

Sentence level sentiment analysis is the most fine-grained analysis of the document. In this, polarity is calculated for each sentence as each sentence is considered as separate unit and each sentence can have different opinions. Sentence level sentiment analysis has two tasks:

- Subjectivity classification of a sentence into one of two classes: objective as well as subjective.
- Sentiment classification of subjective sentences into two classes: positive as well as negative.

An objective sentence presents some factual information, while a subjective sentence expresses personal feelings, views, emotions, or beliefs. Subjective sentence identification can be achieved through different methods. This is an intermediate step that helps filter out sentences with no opinions and helps determine to an extent if sentiments about entities and their aspects are positive or negative.

2.3 Feature Level Sentiment Analysis

Product attributes or components are referred to as product features. Analysis of all such said features in a document or sentence is called feature sentiment analysis. In feature level sentiment classification, from the already extracted features, opinion is determined. This classification is more a specific approach to OM. The phrases which comprise opinions are identified and a phrase level classification is carried out. In certain situations, the exact opinion about an entity can be correctly extracted. But in some cases negation of words can occur locally. In these cases, this level of sentiment analysis suffices. The words that appear very near to each other are considered to be in a phrase.

3. RELATED WORK

In this section, the existing work for the stock market prediction with the availability of user comments is presented here. Also a comparative analysis is presented in table I.

Pagolu et al. (2016) [4] have proposed the concept of stock market movement prediction from the user sentiments extracted from the twitter data in comment format (tweets). For this authors have considered the supervised machine learning approaches of N-grams and Word2vec for the sentiment feature extraction. Authors have defined the relationship between the sentiment analysis and stock market price with the manner to invest in stocks if the user tweets found to be in positive sentiments. Authors have used the feature extraction method with the sentiment analyzer of random forest, SMO and logistic regression. The overall results are evaluated based on the parameters of accuracy, precision, recall and f-measure. Based on the overall results, sentiments based tweets are well analyzed using the Word2vec method as compare to N-gram approach with their analyzer of random forest algorithm.

Karanasou et al. (2016) [5] have used the supervised learning methods for the scalable and real time sentiment analysis based on twitter dataset. The present approach consists of an offline and an online process. In the offline process, a labeled dataset of tweets is preprocessed in order to extract useful features, and then it is used to train a classifier. In the online process, tweets from the Twitter stream are received, preprocessed and features are extracted. Then, each tweet (more accurately its representation using features) is given as input to the classifier, which has already loaded the classification model, and is able to predict the sentiment of the tweet. Authors have used the concepts of Linear SVM, Naive Bayes trained with all features, and the MaxVote technique (also known as Majority vote) for the ensemble classifier that combines the following three classifiers: Linear SVM, Decision Trees, and SGD. The performance is evaluated based on the effect of varying dataset size, different classifiers, different features and feedback effect etc.

Yan et al. (2016) [6] have incorporated the relationship between the Chinese stock market and Chinese local Microblog with the help of public moods extracted from Microblog feeds. C-POMS (Chinese Profile of Mood States) was proposed to analyze sentiment of Microblog feeds. Then Granger causality test confirmed the relation between C-POMS analysis and price series. Authors have used the concepts of SVM and Probabilistic Neural Network to make prediction, and experiments show that SVM is better to predict stock market movements than Probabilistic Neural Network. The main drawback of proposed concept is that it is limited to the prediction of Chinese stock market movement only.

Thirugnanam et al. (2016) [7] have used the dataset of stock market with public sentiments and market for the sentiment analysis. Authors have used the concepts of SVM, Naïve Bayes and Maximum Entropy for the prediction of stock market price with the sentiment. The dataset is extracted from the twitter and yahoo finance market. Granger causality test is used to determine whether sentiment polarity is able to predict the stock price in advance for a company. The use of Naïve Bayes classifier and other had produced better than the baseline methods and had very good accuracy and performed well with the data.

Nguyen et al. (2015) [8] have used the concepts of JST based method and Aspect based sentiment method for the identification of stock market movement by integrating the sentiment analysis in social media. The proposed concepts are compared with the existing approaches of LDA based method, sentiment classification, human sentiment and price only. From the complete investigation, proposed concept shows better results as compare to other existing approaches on the basis of accuracy. The proposed concepts only predicts the price if stock price up or down but people want to forecast drastic movement of stock market. So, considered concept is insufficient. Also authors have used only the historical prices and sentiments derived from social media. So, there is need to extend the model by predicting the degree of change by setting more fine grained classes such as great up, little up, little down, great down etc.

Skuza et al. (2015) [9] have used the Naïve Bayes algorithm for the prediction of stock market movement prediction. Authors have used the twitter data within big data distributed environment. Prediction of future stock prices is performed in this work by combining results of sentiment classification of tweets and stock prices from a past interval. Taking into consideration large volumes of data to be classified and the fact they are textual, Naïve Bayes method was chosen due to its fast training process even with large volumes of training data and the fact that is it is incremental. Considered large volumes of data resulted also in decision to apply a map reduce version of Naïve Bayes algorithm.

Induja K. et al. (2014) [10] have proposed an application oriented fuzzy logic based sentiment analysis for review documents of different products. The author classified the reviews into three parts: negative, positive and neutral. The proposed work shows the 85.58 of accuracy level. In this paper, the corpus used for testing is SFO corpus with text size of 1.45MB containing a collection of 2000 user's reviews from different social networking sites. Empirical results of proposed algorithms show the higher accuracy performance for both binary and fine grained sentiment classification.

Fornacciari et al. (2014) [11] have incorporated the variations of Naive Bayes classifiers for detecting polarity of English tweets. Two different variants of Naive Bayes classifiers were built namely Baseline (trained to classify tweets as positive, negative and neutral), and Binary (makes use of a polarity lexicon and classifies as positive and negative. Authors have neglected the neutral tweets. Authors have included the new features for classifier such as Lemmas (nouns, verbs, adjectives and adverbs), Polarity Lexicons, and Multiword from different sources and Valence Shifters. Authors highlighted the typical problems of Sentiment Analysis (irony, sarcasm, lack of information, etc.). Some peculiar problems of the considered channel were also detected.

Hu et al. (2014) [12] have considered the dataset of tweets of one important Twitter user and the corresponding one stock price behavior. The tweets of Elon Musk, who is the CEO of Tesla, and the change of Tesla stock price are used as data. They tried different sets of features using SVM model. The accuracy and the confusion matrix of this set of features and labeling are reasonable. Around 60% accuracy can be reached if we leave 10% data to be testing set. The major drawback of proposed concept is that it is limited to a single user. Also the method is not able to give up the overall scenario of the market.

Bing et al. (2014) [13] have proposed extracting ambiguous textual tweet data through NLP techniques to define public sentiment, then make use of a data mining technique to discover patterns between public sentiment and real stock price movements. The proposed algorithms have a better prediction performance in some certain industries such as IT and media. On the other hand, authors' study indicates the proposed algorithms have a better performance in using current tweets sentiment to predict the stock price of three days later. The major drawback of this concept considers the daily or weekly closing values of the stock price only. The dataset used is the Twitter data. Most Twitter messages are very short and some of them are actually meaningless so not able to give clear picture.

Xie et al. (2013) [14] have proposed a novel tree representation based on semantic frame parsers. They indicated that this representation performed significantly better than bag-of-words. By using stock prices from Yahoo Finance, they annotated all the news with labels in a transaction date as going up or down categories. However, the weakness of this assumption is that all the news in one day will have the same category. In addition, this becomes a document classification problem, not stock prediction.

Rechenthin et al. (2013) [15] have incorporated Yahoo Finance Message Board into the stock movement prediction. They tried to use various classification models to predict stock. They used the explicit sentiments and predicted sentiments obtained by a classification model with the bag-of-words and meta-features.

Kaur and Gupta (2013) [16] have given a survey on sentiment analysis and opinion mining. Beside English, there is also existence of algorithms that have successfully applied on sentiment analysis to detect the public opinion. In India, scarcity of resources has become the biggest issue for Indian languages. This paper shows that SentiWordNet has successfully implemented for Hindi, Telgu, Bengali and others, a sum of 57 languages for detection of sentiments.

TABLE I

COMPARATIVE ANALYSIS OF SENTIMENT ANALYSIS FOR STOCK MARKET PREDICTION

Author and Year	Paper Title	Dataset Used	Advantages	Drawbacks
Pagolu VS, Challa KN, Panda G, Majhi B. (2016)	Sentiment Analysis of Twitter Data for Predicting Stock Market Movements	Twitter Dataset for stock market	Concludes sentiments based tweets are well analyzed using the Word2vec method as compare to N-gram approach with their analyzer of random forest algorithm	In case of stock market, there are very less people who trades shares their information on twitter.
Karanasou M, Ampla A, Doulkeridis C, Halkidi M. (2016)	Scalable and Real-time Sentiment Analysis of Twitter Data	Twitter Dataset	The performance is evaluated based on the effect of varying dataset size, different classifiers, different features and feedback effect	Results are based on common twitter dataset. Twitter dataset is not domain based.
Yan, Danfeng, Guang Zhou, Xuan Zhao, Yuan Tian, and Fangchun Yang (2016)	Predicting stock using microblog moods	Chinese Microblog	Concludes with SVM is better to predict stock market movements than Probabilistic Neural Network.	The main drawback of proposed concept is that it is limited to the prediction of Chinese stock market movement only
Thirugnanam, Mythili, Smit Patel, Prakhar Vyas, and Tamizharasi Thirugnanam (2016)	Analyzing Company's Stock Price Movement Using Public Sentiment In Twitter Data	Twitter and Yahoo finance market	Granger causality test is used to determine whether sentiment polarity is able to predict the stock price in advance for a company	Stock prediction is with lesser future time duration aspects.
Nguyen, Thien Hai, Kiyoaki Shirai, and Julien Velcin (2015)	Sentiment analysis on social media for stock movement prediction	Social media dataset	JST based method and Aspect based sentiment method shows better results as compare to other existing approaches on the basis of accuracy	authors have used only the historical prices and sentiments derived from social media. Not able to forecast drastic movement of stock market
Skuza, Michał, and Andrzej Romanowski (2015)	Sentiment analysis of Twitter data within big data distributed environment for stock prediction	Twitter dataset as Big data	Naïve Bayes method was chosen due to its fast training process even with large volumes of training data	Lesser accuracy due to large volume of Data as BigData
Indhuja, K., and Raj PC Reghu (2014)	Fuzzy logic based sentiment analysis of product review documents	Social Networking Sites	Empirical results of proposed algorithms show the higher accuracy performance for both binary and fine grained sentiment classification	Only a review of concepts is presented without any particular conclusion

Fornacciari, Paolo, Monica Mordonini, and Michele Tomaiuolo (2014)	A Case-Study for Sentiment Analysis on Twitter	English Tweets	Inclusion of new features for classifier such as Lemmas, Polarity Lexicons, and Multiword from different sources and Valence Shifters	Typical problems of Sentiment Analysis (irony, sarcasm, lack of information, etc.). some peculiar problems of the considered channel were also detected
Hu, Zhiang, Jian Jiao, and Jialu Zhu (2014)	Using Tweets to Predict the Stock Market	Tweets of Elon Musk (who is the CEO of Tesla)	Achieve 60% accuracy if 10% testing data is used.	Limited to a single user. Also the method is not able to give up the overall scenario of the market.
Bing, Li, Keith CC Chan, and Carol Ou (2014)	Public sentiment analysis in twitter data for prediction of a company's stock price movements	Textual tweets	Proposed algorithms have a better performance in using current tweets sentiment to predict the stock price of three days later.	Most Twitter messages are very short and some of them are actually meaningless so not able to give clear picture.
Xie, Boyi, Rebecca J. Passonneau, Leon Wu, and Germán G. Creamer (2013)	Semantic frames to predict stock price movement	Yahoo Finance dataset	Novel tree representation based on semantic frame parsers performed significantly better than bag-of-words	Annotated all the news with labels in a transaction date as going up or down categories. This becomes a document classification problem, not stock prediction.
Rechenthin, Michael, W. Nick Street, and Padmini Srinivasan (2013)	Stock chatter: Using stock sentiment to predict price direction.	Yahoo Finance Message dataset	used the explicit sentiments and predicted sentiments obtained by a classification model with the bag-of-words and meta-features	Not well performed for the stock market prediction
Kaur, Amandeep, and Vishal Gupta (2013)	A survey on sentiment analysis and opinion mining techniques	Social media dataset for various Indian languages	SentiWordNet has successfully implemented for Hindi, Telgu, Bengali and others, a sum of 57 languages for detection of sentiments	Only review of concept is presented. Not comes up with any conclusion

4. RESEARCH GAP

There are several challenges in Sentiment analysis for stock market movement prediction. As work presented in literature survey, most of the authors have used the dataset of twitter for the sentiment analysis. But the major problem with twitter dataset is that usually tweets are limited to some words that sometimes not expressed the exact meaning. Another one is that all the peoples who trade in stock market are not available on twitter. Also sometimes, there may be the challenge that people don't always express opinions in a same way. All the times different people give different opinion for the same sentences. But the problem here raised the understandability of people’s comments that what a person thought based on the short piece of text because it lacks context. Mostly user’s comments are dependent on other things. To improve this there is the need of a technique that can perform classification as well as analysis of the sentiments. Also there is the need of some optimization technique to improve the accuracy evaluation parameters.

5. CONCLUSION

In this paper, the work of different authors in domain of stock market prediction using sentiment analysis has been presented. Different authors have used different datasets like Twitter dataset, Chinese Microblogs, Yahoo & Twitter finance market dataset and some other social media dataset in languages other than English also. As per their used dataset and methods, their advantages and drawbacks has also been discussed in table I. As per the existing work, there is the need of user’s comments is dependent on other things and there may be the challenge that people don't always express opinions in a same way. So, there is the need of some other efficient approach with that we can optimize the results using optimization concepts like swarm intelligence.

REFERENCES

[1]. Sang, Jitao, Yue Gao, Bing-kun Bao, Cees Snoek, and Qionghai Dai. "Recent advances in social multimedia big data mining and applications." *Multimedia Systems* 22, no. 1 (2016): 1-3.

[2]. Levine, Ross, and Sara Zervos. "Stock markets, banks, and economic growth." *American economic review* (1998): 537-558.

[3]. Cambria, Erik. "Affective computing and sentiment analysis." *IEEE Intelligent Systems* 31, no. 2 (2016): 102-107.

[4]. Pagolu VS, Challa KN, Panda G, Majhi B. “Sentiment Analysis of Twitter Data for Predicting Stock Market Movements”. arXiv preprint arXiv:1610.09225. 2016 Oct 28.

[5]. Karanasou M, Ampla A, Doulkeridis C, Halkidi M. “Scalable and Real-time Sentiment Analysis of Twitter Data”, 2016.

[6]. Yan, Danfeng, Guang Zhou, Xuan Zhao, Yuan Tian, and Fangchun Yang. "Predicting stock using microblog moods." *China Communications* 13, no. 8 (2016): 244-257.

[7]. Thirugnanam, Mythili, Smit Patel, Prakhar Vyas, and Tamizharasi Thirugnanam. "Analyzing Company’s Stock Price Movement Using Public Sentiment In Twitter Data.", *The IIOEB Journal,* vol. 7, Suppl 1, pp. 127–136, 2016.

[8]. Nguyen, Thien Hai, Kiyoaki Shirai, and Julien Velcin. "Sentiment analysis on social media for stock movement prediction." *Expert Systems with Applications* 42, no. 24 (2015): 9603-9611.

[9]. Skuza, Michał, and Andrzej Romanowski. "Sentiment analysis of Twitter data within big data distributed environment for stock prediction." In *Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on*, pp. 1349-1354. IEEE, 2015.

[10]. Indhuja, K., and Raj PC Reghu. "Fuzzy logic based sentiment analysis of product review documents." In *Computational Systems and Communications (ICCSC), 2014 First International Conference on*, pp. 18-22. IEEE, 2014.

[11]. Fornacciari, Paolo, Monica Mordonini, and Michele Tomaiuolo. "A Case-Study for Sentiment Analysis on Twitter." In *WOA*, pp. 53-58. 2014.

[12]. Hu, Zhiang, Jian Jiao, and Jialu Zhu. "Using Tweets to Predict the Stock Market." (2014).

[13]. Bing, Li, Keith CC Chan, and Carol Ou. "Public sentiment analysis in twitter data for prediction of a company's stock price movements." In *e-Business Engineering (ICEBE), 2014 IEEE 11th International Conference on*, pp. 232-239. IEEE, 2014.

[14]. Xie, Boyi, Rebecca J. Passonneau, Leon Wu, and Germán G. Creamer. "Semantic frames to predict stock price movement.", *In Proceedings of the 51st annual meeting of the association for computational linguistics*, pp. 873-883, (2013).

[15]. Rechenthin, Michael, W. Nick Street, and Padmini Srinivasan. "Stock chatter: Using stock sentiment to predict price direction." *Algorithmic Finance* 2, no. 3-4 (2013): 169-196.

[16]. Kaur, Amandeep, and Vishal Gupta. "A survey on sentiment analysis and opinion mining techniques." *Journal of Emerging Technologies in Web Intelligence* 5, no. 4 (2013): 367-371.