Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 70 (2015) 85 – 91

Procedia Computer Science

4th International Conference on Eco-friendly Computing and Communication Systems

Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty

Aditya Bhardwaj[a*], Yogendra Narayan[b], Vanraj[c], Pawan[a], Maitreyee Dutta[a]

[a]*Computer Science Engineering Department, National Institute of Technical Teachers Training and Research, Chandigarh, India*
[b]*Electrical Engineering Department, National Institute of Technical Teachers Training and Research, Chandigarh, India*
[c]*Mechanical Engineering Department, National Institute of Technical Teachers Training and Research, Chandigarh, India*

Abstract

From the last twenty years, the application of Internet based technologies had brought a significant impact on the Indian stock market. Use of the Internet has eliminated the barriers of brokers and geographical location because now investors can buy and sell their shares by accessing the stock market status from anywhere at any time. Before investing money, it is very important for investors to predict the stock market. In today's digital world Internet based technologies such as Cloud Computing, Big Data analytics, and Sentiment analysis have changed the way we do business. Sentiment analysis or opinion mining makes use of text mining, natural language processing (NLP), in order to identify and extract the subjective content by analyzing user's opinion, evaluation, sentiments, attitudes and emotions. In this research work importance of sentiment analysis for stock market indicators such as Sensex and Nifty has been done to predict the price of stock. Finally, we draw conclusions and provide suggestions for future work.

Peer-review under responsibility of the Organizing Committee of ICECCS 2015

Keywords: Sentiment analysis; Opinion mining; Indian stock market; Sensex; Nifty

1. Introduction

With the emergence of the Web 2.0 era of the internet, social networks have revolutionized the way in which people communicate.

*Aditya Bhardwaj. Tel.: +91-9041710993.
E-mail address: adityaform@gmail.com

Peer-review under responsibility of the Organizing Committee of ICECCS 2015
doi:10.1016/j.procs.2015.10.043

People use social networking sites, like Facebook, Twitter, etc. to express their opinions and views about a particular topic such as news, movie, event and remarks related to product[1]. This information available from social network is beneficial for business analyst for mining the user opinion about their products and considers these opinions as feedback to improve their policies, planning and process for product development. Sentiment analysis is used to extract such opinion and remarks of users by classifying them as positive, negative and natural sentiment[2]. Although there are a number of definitions about sentiment analysis in the literature, but in simple terms sentiment analysis is a technique used to extract intelligent information based on the person's opinion from raw data available on the internet. In this definition, the term opinion means a person's perspective about an object or issue. There are some challenges related to sentiment analysis, the first challenge is a word that is used to express an opinion; it can be positive as well as negative depending upon the type of sentiment[14]. For example: if a word "large " is used for size of mobile device, then it is considered as negative, whereas if the statement contains "large" word for the height of a person then it is considered as a positive statement. The second challenge is related to the opinion holder as opinion holder is always changing its statement, according to his state of mind, it is very difficult to understand such type of statement by the machine. For example: "I like the picture quality, but the battery life is poor". This statement is a combination of both positive and negative statements. Also, there is a problem when the statement is too short to understand even by human being. Indian stock market has gained the interest of investors investing in two main stock market named as Bombay Stock Exchange (BSE) and National Stock Exchange (NSE). There is high risk involved for investors because of more complexity of the stock market. The Sensex and NIFTY are two such prominent market indices that function within the Indian stock market. These two market indexes represent the stocks for BSE (Bombay Stock Exchange) and NSE (National Stock Exchange) respectively. Specifically, under BSE there are 30 companies for Sensex, while under NSE there are 50 companies for Nifty. So there is need to predict the stock market status for investors by using these two most important indicators i.e Sensex and Nifty. This paper is organized as follows. In section 2 sentiment classification methodologies are discussed. In section 3 and 4 previous studies, literature survey and proposed work with implementation is presented. Finally the paper is concluded in section 5.

2. Sentiment Classification Methodologies

Sentiment classification techniques can be mainly divided into Machine Learning approach and Lexicon based approach. These techniques are explained as follows:-

2.1 Machine Learning Approach

Machine learning techniques that are applied in the field of sentiment analysis can be divided as supervised and unsupervised learning methods.

2.1.1 Unsupervised Learning

Unsupervised learning has no explicit target output associated with input, and it is learning through observation. The goal is to have the machine learn without giving any explicit instruction. Famous approach in unsupervised learning is clustering, in which similarities of elements in the training data is found out. Cluster similarity parameter is defined upon metrics such as Euclidean distance. K-means, Hierarchical, Gaussian mixture models, Self-organizing maps, and Hidden Markov models are some of the clustering algorithms[18].

2.1.2 Supervised Learning

Supervised learning is one that makes use of known dataset to make the prediction of output result. Supervised learning requires two sets of documents: training set and test set. For learning different properties of documents, training set is used and for evaluating the performance classifier test set is used. Various supervised learning techniques are discussed as follows:

2.1.2.1 Decision tree classifier

Decision tree uses a hierarchical decomposition of training data in which data is divided based on the condition of attribute values. Generally a condition for the division is presence or absence of words. Each non-leaf node is associated with attribute feature, and each leaf node is associated with a classification value positive or negative.

2.1.2.2 Rule based classifier

Rule based classifier is based on the rule on occurrences of emotions in the text. If a word contains positive emotions, then it is considered as positive, and if the word contains negative emotions it is considered as negative. Rule base classifier is similar as the fuzzy logic system that allow intermediate vale to be defined between conventional evaluations like yes/no, true/false, high/low, etc.

2.1.2.3Probabilistic classifiers

Probabilistic classifiers are developed by assuming generative models which are product distributions over the original attribute space or more involved spaces. While this paradigm has been shown experimentally successful on real world applications, despite vastly simplified probabilistic assumptions. Probabilistic classifier is based on the prediction of input given probability distribution. Two most important probabilistic classifiers are discussed as follows

- *Naïve Bayes classifier:* This classifier is based on Bayes theorem of probabilistic model. In this we tried to estimate the probability of a text based on whether it belongs to positive or negative class.
- *Maximum Entropy classifier:*-Maximum Entropy classifier is a probabilistic based classifier which belongs to the exponential model class. Principle of maximum entropy is used in this chapter and distribution having largest entropy is chosen [5].

2.1.2.4 Linear classifier

Linear classifier is one that partition a set of object into their respective domain with a line, and partitioning with a curve is called as hyper plane. Two famous linear classifier approaches are as follows:

- *Support Vector Machine (SVM):* A frequent challenge in the assignment of sentiment analysis of a text is to home in on those aspects of the text which are in some way representative of the tone of the whole text. SVM is a supervised learning classifier widely used for classification and regression analysis. The basic idea of SVM is to determine linear separator in the search space which can separate the different classes.
- *Neural Network:* This method is based on the neuron. Multilayer neural networks are used for non-linear boundaries which are used for enclosed regions of a particular class. In artificial neural networks the input of the neuron are combined in a linear way with different weight. The result of this combination is then fed into a non linear activation unit which can in its simplest form be a threshold unit. Neural network offer nonlinearity, input output mapping, adaptivity and fault tolerance. The high connectivity of the network ensures that the influence of error in a few terms will be miner, which ideally gives a high fault tolerance.

2.1.2.5 Lexicon based approach

For sentiment analysis lexicon based approach is robust that result in good cross-domain performance. This method is based on the assumption that the sum of the sentiment orientation of each word makes contextual sentiment orientation. This method is further divided into two types as discussed below:

- *Dictionary Based approach:* This approach use predefined dictionary of words where each word is associated with a specific sentiment polarity strength. Feeling of people such as happy, sad or depressed can be found out by comparing word against lexicons from dictionaries.
- *Corpus based approach:* Corpus based approach try to find co-occurrence patterns of words to determine their sentiments. This approach is based on seeding list of opinion words and then find another opinion words which have similar context. This method is used to assign happiness factor of words depending on frequency of their occurences in “happy” or “sad” blog post.

2.1.3 Hybrid Fuzzy Neural network based learning

Hybrid system is those for which more than one technology is employed to solve the analysis problem for Indian stock market prediction .The hybrid system are classified as (i) sequential hybrids (ii) auxiliary hybrids (iii) embedded hybrids. Sequential hybrid systems make use of technologies in a pipeline like structure. In auxiliary hybrid systems subroutine is used to process or manipulate information provided to it where as in embedded hybrid systems, the technologies participating are integrated in such a manner that they appear intertwined. The fusion is so complete that it would appear that no technology could be used without the others for solving the problem.

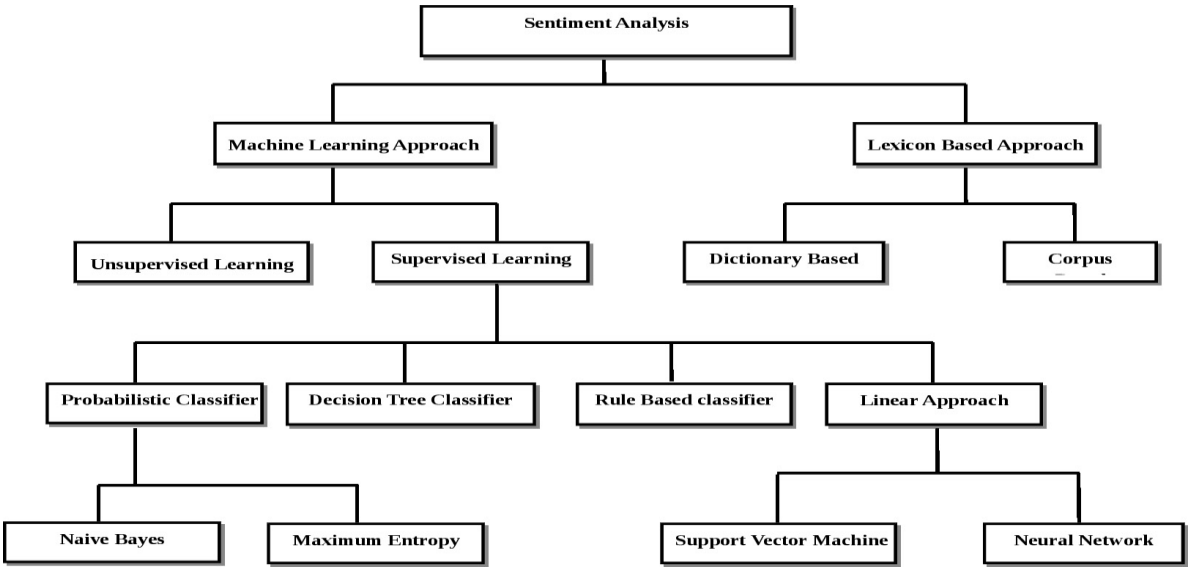

Figure1: Sentiment Classification Techniques

3. Literature Survey

In a recent study by Singh, P. K et al.[7] sentiment analysis was done across Flipkart E-commerce websites for filtering of irrelevant reviews and MongoDB database technology was used at backend for this research work.

In another study by Gunduz et al.[9] sentiment analysis between sentiments of people on social network and academic success of Turkish universities was done to find out that is there any relation between university's academic success and sentiment about those university in the social media based on the Naive Bayes classifier. For this purpose top 10 most successful Turkish Universities ranked by URAP were selected for analysing sentiment about them on the social media. Twitter, which allows users to share tweets with social friends or followers was chosen as a specific social media for this study. Firstly tweets were collected via Twitter REST API, after which tweets were labelled as positive, negative or neutral. Pre-processing of feature extraction was done by extracting meaningful special character from tweets. Tweets then classified into word list based on the two approaches: one was Time Frequency and another one was Inverse Document Frequency. From the results evaluated, the success rate of the system was found to be 72.33%.

Molla et al.[10] made sentiment analysis for user opinions about different Samsung products using different twitter official accounts of Samsung Company. For visualizing the result of user's opinion data visualization tool such as NodeXL was used for the social network graph. Future work was proposed that it may be focus on the location management of each tweet and inclusion of emotions.

Lu, Y., & Chen, J[12] presented a study for the opinion analysis of micro blog content. The public opinion model was divided into four modules: data collection module, corpus processing module, sentiment analysis module and data management module. For retrieving online microblog content crawler was used, and for classifying mircoblog, text classification method called support vector machine was used. The result shows that precision classification exceeded 90% by the use of classifier support vector machine. It was proposed that more work could be done for improving the performance of support vector machine.

Batool, R et.al[19] analysed 4000 tweets to classify data and sentiment more precisely from twitter containing information such as food, diet, diabetes, education, and movies. First knowledge generator was used to classify tweets into different categories, and then knowledge enhancer with synonym binder was applied to increase the information gain. Knowledge enhancer module adds additional knowledge that was not extracted by Alchemy API used in knowledge generator phase. Synonym binder was used to the bind synonyms with entity and keyword

extracted by knowledge generator and knowledge enhancer. Results showed that overall significant improvement from 0.1% to 55% had been achieved using the said approach.

M. Meral & B. Diri[25] performed sentiment analysis of Turkish tweets on nine different domains such as insurance, sport, finance, food, automotive, politics, real estate, Telecommunication, and health. Collection of Turkish tweets was done by using Naïve Bayes, Support Vector Machines and Random Forest. Classification of tweets was done as neutral, positive, and negative. Tweets were then divided as- health, politics, finance, and telecommunication in negative sentiment category while food, real estate, sport, and automotive tweets in neutral category, and rest tweets as positive. From the results obtained it was concluded that Support vector machine give best results as compare to other classifiers.

Li, S., Wang et al.[27] applied sentiment analysis by using twitter data for predicting success rate of movie. For this purpose movies were classified as Flop, Hit, and Average. The tweets from 2009 to 2013 duration were extracted, and each tweet was classified as positive, negative, neutral and irrelevant. Lingpipe sentiment analyzer was used to test the sentiments, and result showed that movie prediction accuracy of the developed system was 64.4 % better than conventional system.

In another study conducted by Wang, X., & Luo[28], for predicting the movie performance based on social networking sites data using sentiment analysis technique. Sentiments from various social media such as Twitter and YouTube were collected. Prediction of movie was done by using K-means clustering algorithm.

4. Proposed System and Implementation

The prime aim of the proposed system is to fetch live server data by using Python programming language, which can be used for performing sentiment analysis on the extracted datasets from online news portal. In this context, first python is installed on the Ubuntu 14.04 LTS host machine, after that required software such as Beautiful Soup is installed using Ubuntu terminal command prompt environment. For Debian or Ubuntu platform Beautiful Soup software can be installed with system package manager. Beautiful Soup use Python library for fetching live data, and this tool helps to pull contents from desired webpage then save the required information. Beautiful Soup supports HTML parser which is included in Python's standard library. For illustration purpose we fetch the live data of Sensex and Nifty that can be used further for sentiment analysis.
The steps needed for sentiment analysis using python scripting language are given in the flowchart as follows:

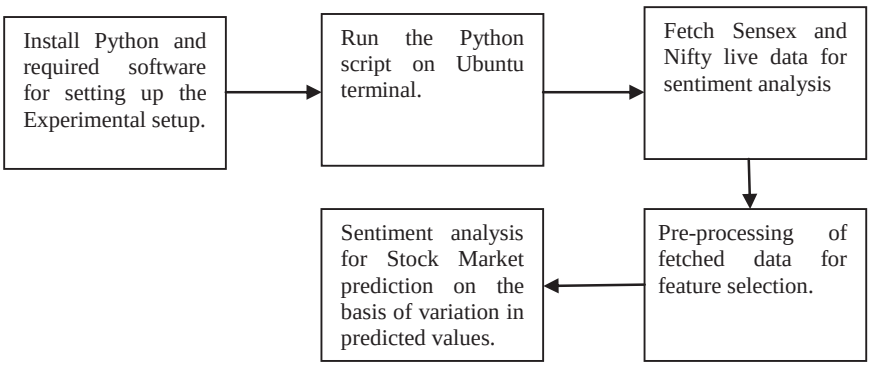

Figure2: Flowchart for the proposed system

For implementation purpose the proposed system fetched the live Sensex and Nifty data values from Timesofindia.com. We have run python script with sleep count time interval of one second for fetching the data, and values were calculated for different time interval. After that result is drawn which shows that for a particular time interval the fetched values of Sensex and Nifty remains constant.

```
from bs4 import BeautifulSoup
import urllib.request
from time import sleep
from datetime import datetime

def get news();
url = http://timesofindia.indiatimes.com/business
req = urllib.request.urlopen(url)
page = req.read()
scraping = BeautifulSoup (page)
price = scraping.findAll ("dd", alttrs = {"class : "value"}) [0].text
return price
with open("TimesofIndia.out", "w") as f:
      for x in range (2, 100):
            sNow = datetime.now().strftime("%I:%M:%S%p")
            f.write("{0}, {1} \n ". format (sNow, (get news()))
            sleep (1)
```

Figure 3: Python script code for fetching live server data.

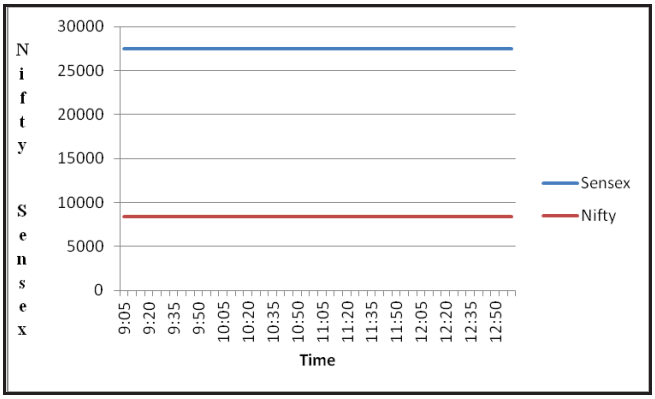

Figure 4: Sensex vs Nifty values fetched on different time interval

5. Conclusion

There are various ups and downs in Indian stock market. In order to invest money in stock market for purchasing the shares it is very essential for the investors to predict the stock market condition. In India scenario Sensex and Nifty are two major indicator for prediction of stock market condition. For BSE (Bombay Stock Exchange) companies Sensex and for NSE (National Stock Exchange) companies Nifty is used as an indicator of stock market prediction. But the major problem for the investors are to predict the stock market condition which depends upon regular checking and testing of Sensex and Nifty prediction values. In order to allow this, in this paper we have demonstrated sentiment analysis for stock market by fetching Sensex and Nifty live server data values on different interval of time that can be used for predicting the stock market status. For this purpose we have used python scripting language which have fast execution environment and this will help out the investors in order to make a prediction of on what shares money should be invested, it will also help in maintaining the economical balance of share market. In Future work can be done by running these python script code with more advanced functions.

References

1. Sun, B., & Ng, V. T, Analyzing sentimental influence of posts on social networks IEEE International Conference on Computer Supported Work in Design, 2014, pp.546-551.
2. Zhang, J., & Huang, M. L., 5Ws model for bigdata analysis and visualization, IEEE 16th International Conference on Computational Science and Engineering, 2013, pp.1021-1028.
3. Usha, M. S., & Indra Devi, M, Analysis of sentiments using unsupervised learning techniques, IEEE International Conference on Information Communication and Embedded System, 2013, pp.241-245.
4. Ameur, H., & Jamoussi, S, Dynamic construction of dictionaries for sentiment classification, IEEE 13th International Conference on Data Mining Workshops, 2013, pp.896-903.
5. Yi, L., Miao, F., & Xiaoxia, Z, The design and implementation of Feature-Grading recommendation system for e-commerce, IEEE International Conference on Information and Automation, 2011, pp.236-241.
6. Wen, B., Fan, P., Dai, W., & Ding, L, Research on analyzing sentiment of texts based on semantic comprehension, IEEE 3rd International Conference on Consumer Electronics Communications and Networks, 2013, pp.529-532.
7. Singh, P. K., Sachdeva, A., Mahajan, D., Pande, N., & Sharma, A, An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites, IEEE International Conference on the next Generation Information Technology, 2014, pp.329-335.
8. Zhang, X., Fuehres, H., & Gloor, P. A, Predicting stock market indicators through twitter, ELSEVIER Procedia-Social and Behavioral Sciences, 2011, pp. 55-62.
9. Gunduz, S., Demirhan, F., & Sagiroglu, S, Investigating sentimental relation between social media presence and academic success of Turkish universities, IEEE 13th International conference on Machine Learning and Applications, 2014, pp.574-579.
10. Molla, A., Biadgie, Y., & Sohn, K. A. Network based visualization of opinion mining and sentiment analysis on twitter, IEEE International conference on It Convergence and Security, 2014, pp.1-4.
11. Kherwa, P., Sachdeva, A., Mahajan, D., Pande, N., & Singh, P. K, An approach towards comprehensive sentimental data analysis and opinion mining, IEEE International conference on Advance Computing Conference, 2014, pp.606-612
12. Lu, Y., & Chen, J, Public opinion analysis of microblog content, IEEE International conference on Information Science and Applications, 2014, pp.1-5.
13. Shidaganti, G., & Prakash, S, Feedback analysis of unstructured data from collabrative networking a BigData analytics approach, IEEE International Conference on Circuits, Communication, Control and Computing, 2014, pp.343-347.
14. Kechaou, Z., Ben Ammar, M., & Alimi, A. M., Improving e-learning with sentiment analysis of users' opinions, IEEE International Conference on Global Engineering Education, 2011, pp.1032-1038.
15. Wu, Y., & Ren, F, Learning sentimental influence in twitter, IEEE International Conference on Future Computer Sciences and Application, 2011, pp.119-122.
16. Lin, E., Fang, S., & Wang, J, Mining Online Book Reviews for Sentimental Clustering, IEEE 27th International Conference on Advanced Information Networking and Applications Workshops, 2013 , pp.179-184.
17. Thanangthanakij, S., Pacharawongsakda, E., Tongtep, N, An empirical study on multi-dimensional sentiment analysis from user service reviews," IEEE 7th International Conference on Knowledge, Information and Creativity Support Systems, 2012, pp.58-65.
18. Batool, R., Khattak, A. M., Maqbool, J., & Lee, S, Precise tweet classification and sentiment analysis, IEEE 12th International Conference on Computer and Information Science, 2013,pp.461-466.
19. Mao, X., Rao, Y., & Li, Q, Recipe popularity prediction based on the analysis of social reviews, IEEE International Conference on Awareness Science and Technology and Ubi-Media Computing, 2013, pp.568-573.
20. Sanchez, F. S., & Vazquez, A. M, Sentiment Analysis for e-Services, IEEE International Conference on Advanced Applied Informatics,2014, pp.42-47.
21. Liu, D., Quan, C., Ren, F., & Chen, P, Sentiment and sentimental agent identification based on sentimental sentence dictionary, IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2008, pp.1-5.
22. Wen, B., Dai, W., & Zhao, J, Sentence Sentimental Classification Based on Semantic Comprehension., IEEE 5th International Symposium on Computational Intelligence and Design, 2012, pp.458-461.
23. Fong, S., Zhuang, Y., Li, J., & Khoury, R, Sentiment Analysis of Online News Using MALLET, IEEE International Symposium on Sentiment Analysis of Online News Using MALLET, August 2013, pp. 301-304.
24. Meral, M., & Diri, B "Sentiment analysis on Twitter, IEEE 22nd International Conference on signal Processing and Communications Applications Conference, 2014, pp. 690-693.
25. Celikyilmaz, A., Hakkani-Tur, D., & Feng, J, Probabilistic model-based sentiment analysis of twitter messages, IEEE Spoken Language Technology Workshop, 2010, pp. 79-84.
26. Li, S., Wang, Z., Lee, S. Y. M., & Huang, C. R, Sentiment Classification with Polarity Shifting Detection, IEEE International Conference on Asian Language Processing, 2013, pp.129-132.
27. Wang, X., & Luo, X, Sentimental Space Based Analysis of User Personalized Sentiments, IEEE 9th International Conference on Semantics, Knowledge and Grids, October 2013, pp. 151-156.