Sentiment Analysis and Opinion Mining on Social Media Text Data

[1]Sanjida Hoque Shoshey,[1]Md. Moniruzzamanand [2]Shahida Rafique
[1]Department of Computer Science and Engineering
Institute of Science and Technology, National University, Bangladesh
Email: shoshey.return@gmail.com
[2]Department of Electronics and Communication Engineering
Institute of Science and Technology, National University, Bangladesh

Abstract

Sentiment Analysis has been done on social media text data based on Naive Bayes classifier and Support Vector Machine (SVM). Using these two classifier a new classifier is designed and developed. The new classifier can more accurately identify sentiment then Naive Bayes and SVM. For evaluating the proposed technique efficiency, a comparison between the proposed classifier and two famous classifier has been done. The results have a comparison between the accuracy and performance between the three classifier when applying the techniques on three data sets (training, test and verified). The comparison results illustrate how the proposed technique can increase accuracy and performance with facing many language coverage cases and solving some sentiment analysis challenges. The accuracy results show in Naive Bayes (81.24%), SVM (81.9%) and proposed NBSVM (83.5%) approximately. The technique with the highest F-measure was faced sentiment analysis challenges is NBSVM.

Keywords: Sentiment Analysis, Opinion Mining, Data Mining, Text Mining.

1. INTRODUCTION

Sentiment Analysis and Opinion Mining on Social Media Text Data is an excellent technique for gathering people’s opinions. There are now numerous social media containing such opinions, e.g., product reviews, forums, discussion groups, and blogs. Techniques are now being developed to exploit these sources to help organizations and individuals to gain such important information easily and quickly. The resulting emerging fields are opinion miningand sentiment analysis. It focuses on polarity detection and emotion recognition, respectively. Because the identification of sentiment is often exploited for detecting polarity, however, the two fields are usually combined under the same umbrella or even used as synonyms. Both fields use data mining and natural language processing (NLP) techniques to discover, retrieve, and distill information and opinions from the World Wide Web’s vast textual information.

World Wide Web has become the most popular communication platforms to the public reviews, opinions, comments and sentiments. These sentiments refers to opinions about products, places, books or research papers become daily text reviews. The number of active user bases and the size of their reviews created daily on online websites are massive. There are 2.4 billion active online users, who write and read online around the world [1]. Although the scientific domain is huge as a big world of journals and conferences, there are more than 4000 rated conferences and 5000 ranked journals [2].Notably, a large fragment of WWW researchers make their content public, allowing researchers, societies, universities, and corporations to use and analyze data. According to a new survey conducted by dimensional research, April 2013: 90% of customer's decisions depends on online reviews [3]. According to 2013 Study [4]: 79% of customer's confidence is based on online personal recommendation reviews. As the result, a large number of studies and research have monitored the trending increase of online research resources year by year. Thisnew proposed technique is created for evaluating sentiment analysis and domain parameters. A new proposed technique is presented in this thesis.

2. RELATED WORK

Schukla presented a tool which judges the quality of text based on annotations on scientific papers [6]. Its methodology collects sentiments of annotations in two approaches. It counts all the annotation produces the documents and calculates total sentiment scores. Its problem declares in a relationship between annotations that is complex. The technique needs to have a big query knowledge base containing metadata. Kasper &Vela proposed a "Web Based Opinion Mining system" for hotel reviews [7]. The paper introduced an evaluation system for online user's reviews and comments to support quality controls in hotel management system. It is capable of detecting and retrieving reviews on the web and deals with German reviews. It has multi-topic domain and is based on multi-polarity classification; the system could recognize the neutral e.g., "don't know" to "classify sentiment polarity that as neutral" and the multi-topic cases identified in their corpus. Mobile devices products reviews were analyzed by (Zhang, et-al) in [7]. This research can help in evaluate accuracy. It is useful in a judgment of the product quality and status in the market [7]. This research used three machine learning algorithms (Naive Base Classifier, K-nearest neighbor, and random forest) to calculate the sentiments accuracy. The random forest improves the performance of the classifier. There are some ways in analyzing sentiments and opinions. (Godbole, et-al) analyzed news sentiments and blogs [8]. It splits prior work in the context of their specific task (sentiment analysis for news and blogs) into two categories. First category which - regards with techniques for automatically creating sentiment lexicon and the second one which relates to systems that analyze sentiment for entire documents.

2. BACKGROUND STUDY

Sentiment analysis is called: also opinion mining which is a computational study of reviews, sentiments, opinions, evaluations, attitudes, subjective, views, emotions, etc., expressed in the text. In the following sections, we will discuss sentiment analysis area and the factors affecting them.

What is Sentiment Analysis?

We present the sentiment and sentiment analysis definitions and the differences between them. Sentimentscan be recognized as emotions, or as judgments, opinions or ideas prompted or colored by emotions or susceptibility or feelings [9]. In Computational Linguistics, the focus is on opinions and sentimentsrather than on feelings or emotions, and the words 'sentiment' and 'opinion' are often used alternately, also in this paper. There are two types of textual information: facts and opinions information. While the facts are objective expressions about objects, features, entities, events and their characteristics, opinions are ordinarily subjective expressions that identify people's sentiments, views or feelings toward objects, entities, events and their characteristics [10].

How does Sentiment Analysis works

Sentiment analysis is a complex process that involves 5 different steps to analyze sentiment data. These steps are:

1. Data collection: the first step of sentiment analysis consists of collecting data from use generated content contained in blogs, forums, and social networks. These data are disorganized, expressed in different ways by using different vocabularies, slangs, context of writing etc. Manual analysis is almost impossible. Therefore, text analytics and natural language processing are used to extract and classify;
2. Text preparation: consists in cleaning the extracted data before analysis. Non-textual contents and contents that are irrelevant for the analysis are identified and eliminated;
3. Sentiment detection: the extracted sentences of the reviews and opinions are examined. Sentences with subjective expressions (opinions, beliefs and views) are retained and sentences with objective communication (facts, factual information) are discarded;
4. Sentiment classification: in this step, subjective sentences are classified in positive, negative, good, bad; like, dislike, but classification can be made by using multiple points;
5. Presentation of output: the main objective of sentiment analysis is to convert unstructured text into meaningful information. When the analysis is finished, the text results are displayed on graphs like pie chart, bar chart and line graphs. Also time can be analyzed and can be graphically displayed constructing a sentiment time line with the chosen value (frequency, percentages, and averages) over time.

Classification of sentiment analysis approaches

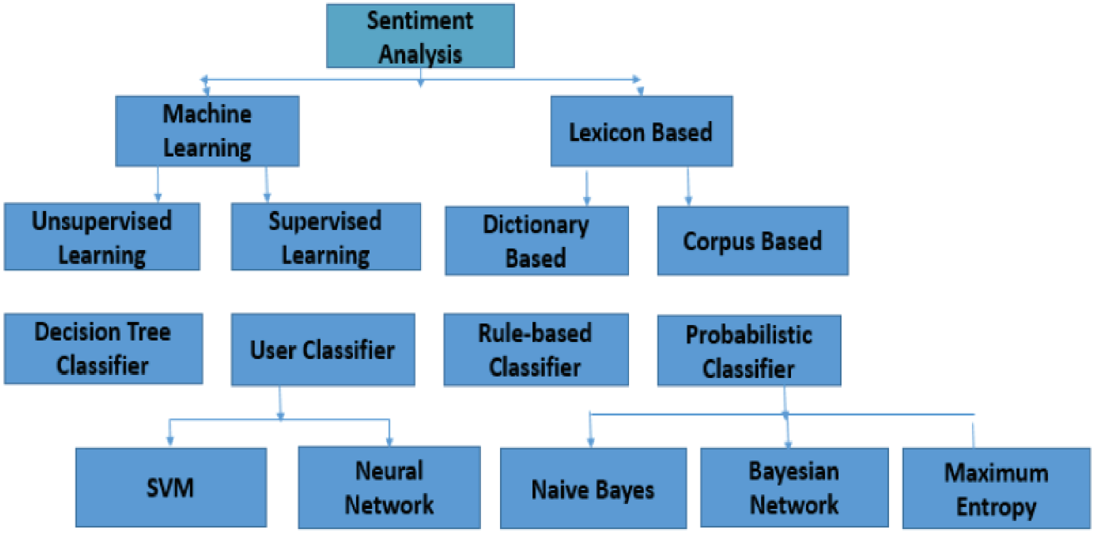


Figure – 1 Classification of Sentiment Analysis

Figure – 1 shows the classification of sentiment analysis approaches from where Naive Bayes classifier and SVM classifier will be used.

3. PROPOSED CLASSIFIER FOR SENTIMENT ANALYSIS

This proposed classifier (Figure – 2) will give more accurate opinion comparing to Naive Bayes classifier and SVM classifier. This classifier will also give positive and negative sentiment.

Overview of Proposed classifier

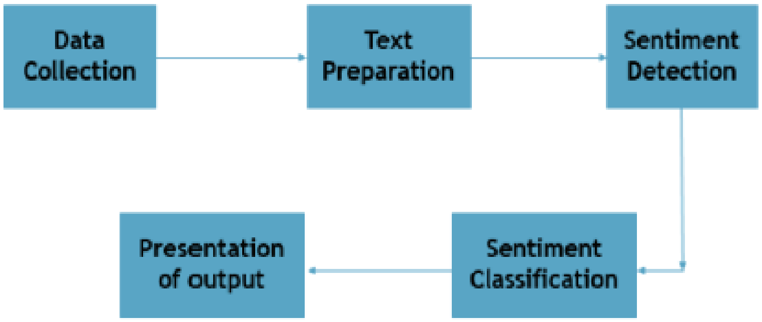


Figure – 2 Proposed Classifier

The sentiment analysis is a complex process that involves 5 different steps to analyze sentiment data. These steps are:

1. Data collection from Tweeter: In figure – 2 the first step of sentiment analysis consists of collecting data from use generated content from tweeter. These data are disorganized, expressed in different ways by using different vocabularies, slangs, context of writing etc. Manual analysis is almost impossible. Therefore, text analytics and natural language processing are used to extract and classify;

2. Text preparation: consists of cleaning the extracted data before analysis. Non-textual contents and contents that are irrelevant for the analysis are identified and eliminated. This step will be done by Eraser Algorithm.

3. Sentiment detection: the extracted sentences of the reviews and opinions are examined. Sentences with subjective expressions (opinions, beliefs and views) are retained and sentences with objective communication (facts, factual information) are discarded. In this step sentence level approachwill be used to identify sentiment keyword.

4. Sentiment classification: in this step, subjective sentences are classified in positive, negative, good, bad; like, dislike, but classification can be made by using multiple points. To classify positive, negative and neutral we will use two classifier named Naive Bayes classifier and Support Vector Machine.

5. Presentation of output: When the analysis is finished, the text results are displayed on graphs. Also time can be analyzed and can be graphically displayed constructing a sentiment time line with the chosen value (frequency, percentages, and averages) over time.

Classifier used in proposed technique

Naive Bayes Classifier (NB). Figure – 3 shows the block diagram of Naive Bayes classifier. It is the simplest and most commonly used classifier.Naive Bayes classification model computes the posterior probability of a class, based on the distribution of the words in the document

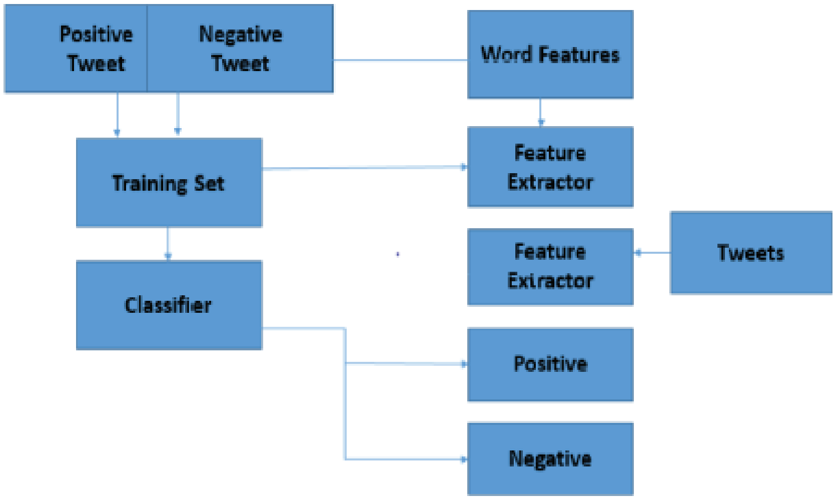


Figure – 3 Naive Bayes Classifier

Support Vector Machines Classifiers (SVM). The main principle of SVMs (Figure - 4) is to determine linear separators in the search space which can best separate the different classes. There are 2 classes x, o and there are 3 hyper planes A, B and C. Hyper plane A provides the best separation between the classes, because the normal distance of any of the data points is the largest, so it represents the maximum margin of separation.

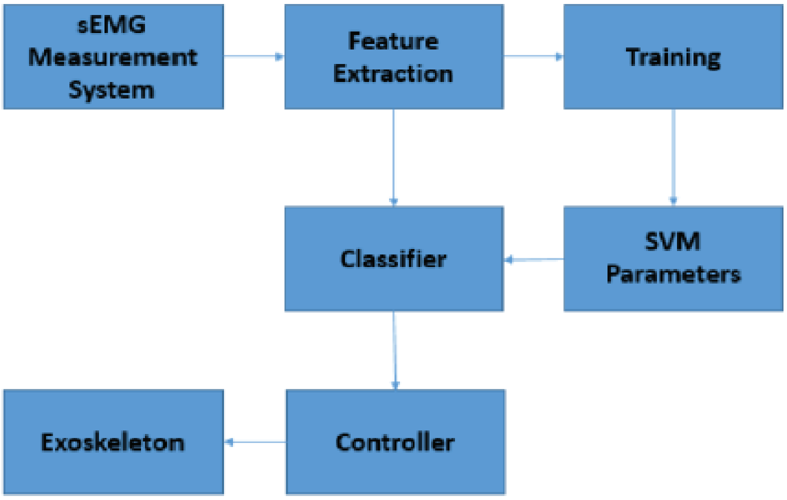


Figure – 4 SVM classifier

Proposed System Design Flow chart

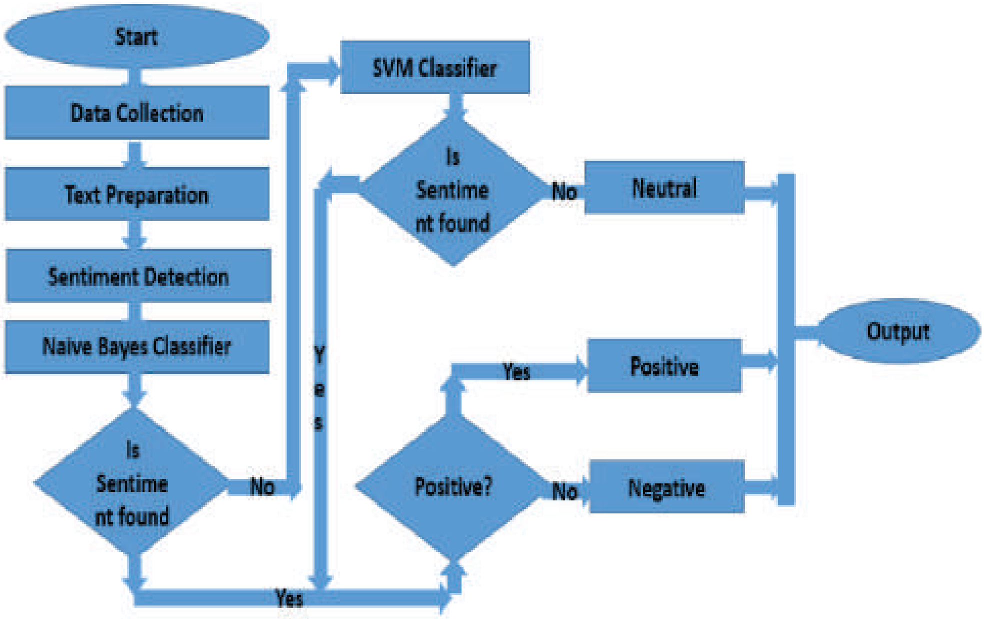
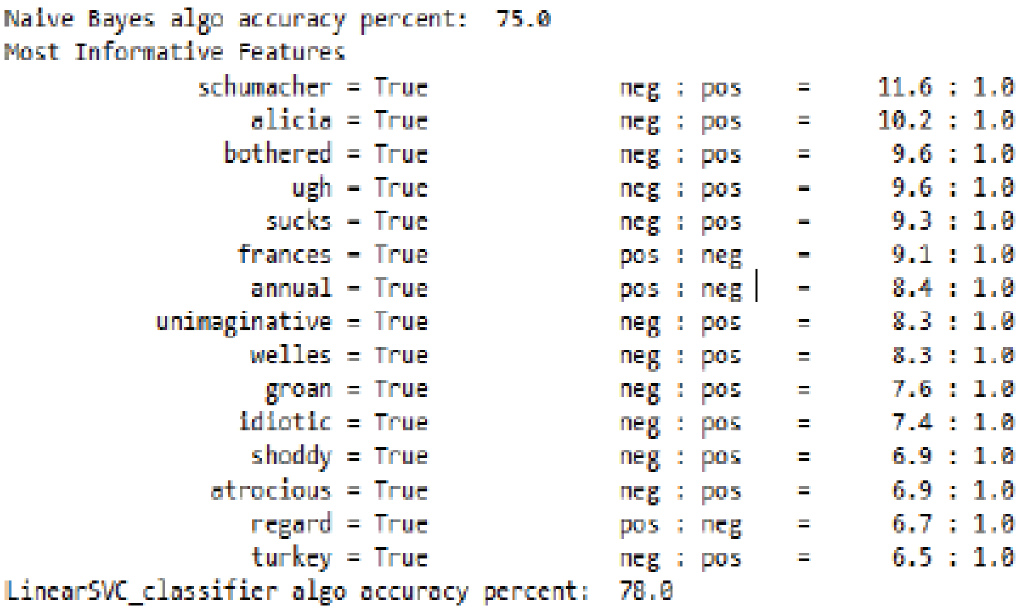


Figure – 5 Proposed Flow Chart

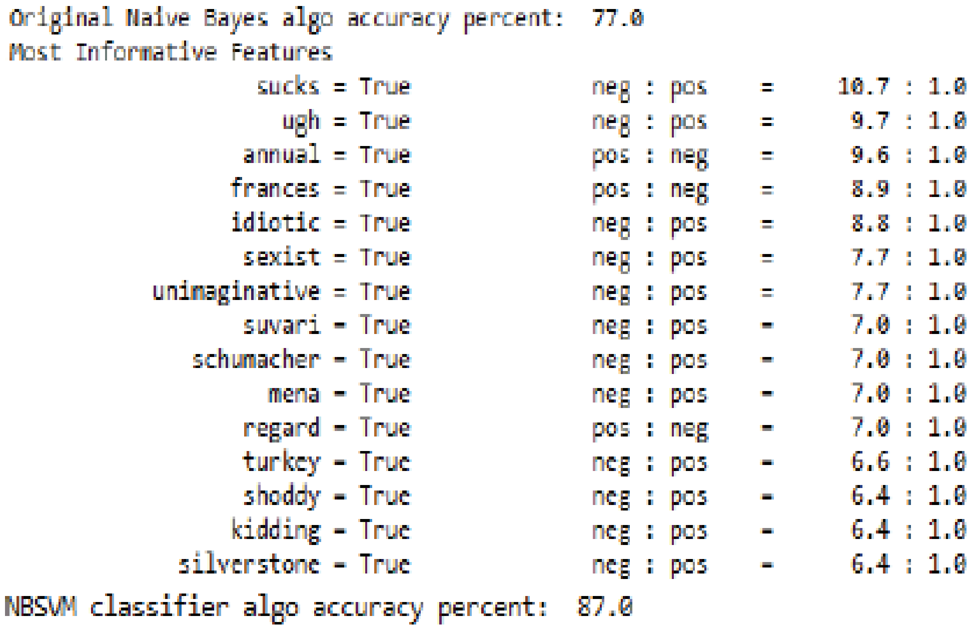
4. IMPLEMENTATION AND RESULT

According to the proposed technique (Figure - 5) the implementation is constructed based on python programming language working on Spyder windows application. Lexicon is based on real time data from Twitter API. Figure – 6 and figure – 7 showed the existing classifier and proposed classifier.



```
Naive Bayes algo accuracy percent:  75.0
Most Informative Features
              schumacher = True              neg : pos    =     11.6 : 1.0
                  alicia = True              neg : pos    =     10.2 : 1.0
                bothered = True              neg : pos    =      9.6 : 1.0
                     ugh = True              neg : pos    =      9.6 : 1.0
                   sucks = True              neg : pos    =      9.3 : 1.0
                 frances = True              pos : neg    =      9.1 : 1.0
                  annual = True              pos : neg    =      8.4 : 1.0
           unimaginative = True              neg : pos    =      8.3 : 1.0
                  welles = True              neg : pos    =      8.3 : 1.0
                   groan = True              neg : pos    =      7.6 : 1.0
                 idiotic = True              neg : pos    =      7.4 : 1.0
                  shoddy = True              neg : pos    =      6.9 : 1.0
               atrocious = True              neg : pos    =      6.9 : 1.0
                  regard = True              pos : neg    =      6.7 : 1.0
                  turkey = True              neg : pos    =      6.5 : 1.0
LinearSVC_classifier algo accuracy percent:  78.0
```

Figure – 6 Comparison between Naive Bayes and SVM



```
Original Naive Bayes algo accuracy percent:  77.0
Most Informative Features
                   sucks = True              neg : pos    =     10.7 : 1.0
                     ugh = True              neg : pos    =      9.7 : 1.0
                  annual = True              pos : neg    =      9.6 : 1.0
                 frances = True              pos : neg    =      8.9 : 1.0
                 idiotic = True              neg : pos    =      8.8 : 1.0
                  sexist = True              neg : pos    =      7.7 : 1.0
           unimaginative = True              neg : pos    =      7.7 : 1.0
                  suvari = True              neg : pos    =      7.0 : 1.0
              schumacher = True              neg : pos    =      7.0 : 1.0
                    mena = True              neg : pos    =      7.0 : 1.0
                  regard = True              pos : neg    =      7.0 : 1.0
                  turkey = True              neg : pos    =      6.6 : 1.0
                  shoddy = True              neg : pos    =      6.4 : 1.0
                 kidding = True              neg : pos    =      6.4 : 1.0
             silverstone = True              neg : pos    =      6.4 : 1.0
NBSVM classifier algo accuracy percent:  87.0
```

Figure – 7 Comparison between Naive Bayes and NBSVM

Table – 1 Percentage of accuracy between classifier

	Data Set	NBSVM	Naive Bayes	SVM
Test 1	Training Set-1900 Testing Set – 1900 Real Set - 3000	87.0%	82.0%	85.0%
Test 2	Training Set-1000 Testing Set – 1000 Real Set - 3000	81.2%	79.7%	80.0%
Test 3	Training Set-1800 Testing Set – 1800 Real Set - 3000	80.5%	79.5%	80.0%
Test 4	Training Set-1900 Testing Set – 1900 Real Set - 4000	86.0%	85.0%	85.5%
Test 5	Training Set-1900 Testing Set – 1900 Real Set - 4200	83.0%	80.0%	79.0%
Average		83.5%	81.24%	81.9%

Implementation Requirements

Programming Language – Python 3.6

Programming Environment – Anaconda Navigator

Programming Editor – Spyder 3.0

NLP tool kit – Python NLTK

Social Media Data Sets – Twitter API

5. DISCUSSION AND CONCLUSION

Due to the sheer volume of opinion rich web resources such as discussion forum, review sites, blogs and news corpora available in digital form, much of the current research is focusing on the area of sentiment analysis. People are intended to develop a system that can identify and classify opinion or sentiment as represented in an electronic.

In this paper, a new technique is present for analyzing online sentiments. A proposed technique targets performing statistical and numerical analysis. It is a hybrid model (the enhancement Bag-of-words (BOW) model combing with Part-of-Speech POS model) for English sentiments. The proposed technique can improve accuracy and support understanding implicit and explicit meaning. The evaluation of these papers based on online social media sentiments and parameters and features of the scientific domain. Then we measure the newly proposed technique efficiency by making a comparison among it and two techniques based on the accuracy and performance.

REFERENCES

[1] Louis, F.& Wojciech, C., Book "Advances in The Human Side of Service Engineering", 5th International Conference on Applied Human Factors and Ergonomics, Volume Set, Proceedings of the 5th AHEE Conference 19-23 July 2014.

[2] Larsen, Peder Olesen, and Markus von Ins. “The Rate of Growth in Scientific Publication and the Decline in Coverage Provided by Science Citation Index.”, Scientometrics 84.3 (2010): 575–603. PMC. Web. 25 Sept. 2015.

[3] Peng, L., Cui, G., Zhuang, M., and Li, C., “What do seller manipulations of online product reviews mean to consumers?” (HKIBS Working Paper Series 070-1314). Hong Kong: Hong Kong Institute of Business Studies, Lingnan University, 2014.

[4] Thomas, B., Keep Social Honest, “What Consumers Think about brands on social media, and what businesses need to do about it” Report, 2013.

[5] Mitchell, P.M., Mary A.M., and Beatrice, S., “Building a Large Annotated Corpus of English: The Penn Treebank”, Computational Linguistics Journal, Vol. 19, Number 21993.

[6] Schukla, A., “Sentiment analysis of document based on annotation”, CORR Journal, Vol. abs/1111.1648, 2011.

[7] Kasper, W. & Vela, M., “Sentiment analysis for hotel reviews”, proceedings of the computational linguistics-applications, Jacharanka Conference, 2011.

[8] Godbole, N., Srinivasaiah, M., and Skiena, S., “Large-Scale Sentiment Analysis for News and Blogs”, ICWSM’2007 Boulder, Colorado, USA, 2007.

[9] Banea, C., Mihalcea, R., and Wiebe, J., “Multilingual Sentiment and Subjectivity Analysis”, In Multilingual Natural Language Processing”, Editors Imed Zitouni and Dan Bikel, Prentice Hall, 2011.

[10] Hassena, R.P., “Challenges and Applications”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 5, 2014.