Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 8, Issue. 5, May 2019, pg.193 – 202

CONTEXT BASED SENTIMENT ANALYSIS OF TWITTER USING HADOOP FRAMEWORK

Tarun R R[1]; Sahana J S[2]; Sadvik B S[3]; Shashank S[4]; Prof. Mahesh T R[5]
[*]Department of CSE, School of Engineering and Technology-Jain University, Bangalore, India
[1] rrtarun04@gmail.com, [2] sahanajs97@gmail.com, [3] sadvik.bs@gmail.com, [4] shashanksrinivas1998@gmail.com, [5] maheshtr.1978@gmail.com

Abstract- People often care for the opinions of others that is where sentiment analysis plays a huge role in analysis how many have positive or good impression on a particular product or object or anything and how many have a negative or bad impression on it. So this sentiment analysis can be made use by those people who give importance to others opinion. As there is a rapid development in technology, social media became a great platform for people to share ideas, views, consult for reviews. This information is used for many purposes, one of them is sentiment analysis. Sentiment Analysis or opinion mining is the process of collecting users' opinion from user generated content. It is used to determine whether a piece of text, word, sentence is positive, negative or neutral. Sentiment analysis provides a very accurate analysis of the overall emotion of the text content incorporated from sources like blogs, articles, forums, consumer reviews, surveys, twitter etc. The opinion is used as data in sentiment analysis. Sentiment Analysis can be widely applied to reviews and social media for a variety of applications ranging from marketing to customer service. It has various applications such as cricket score or win predictions, stock market prediction, product review collection, political predictions based on the sentiments of the people. A number of methods are available for analysis and classification of data.

Keywords- Hadoop, HDFS, Sentiment Analysis, Twitter, Apache

I. INTRODUCTION

People across the world have made themselves obsessed or totally addictive to the social media platforms for expressing every emotion in their day to day life. This rise of internet and social media had made people change their view on expressing their feelings. These emotions are expressed in the form of tweets on Twitter, feeds on Facebook or stories on Instagram and many more platforms. People let others know what they feel or let others know their opinion on various aspects of their life and their surroundings. Sentiments of a huge range of people across the world can be diversified upon different instances, products, people or political parties and these opinions can be well used for better understanding of how much percentage of people feel positive about a thing or political party or any other issue and how many feel negative and the rest who belong to neither side of the opinions. This sentiment analysis can be done with the data that is collected for analysis which is possible by the availability of Application Programming Interfaces that are released by each social media platform so these Application Programming Interfaces(APIs) can help us in analyzing the data and doing the sentiment analysis or also known as the opinion mining.

On Twitter, each year the number of tweets flowing in is increasing rapidly. The data that is tremendous which is Big data is very hard to process due to its enormity so we use Hadoop technology. Hadoop is a distributed framework which is scalable and helps us in processing any amount of data in an efficient way. Hadoop has a scripting language known as pig which we are using for our processing of sentiments. Many researchers have used Tweepy which is a Python library for obtaining the Twitter APIs or Twitter4J which is a java library for obtaining the Twitter APIs.

II. RELATED WORK

In the recent years, there is a lot of work that has been done on Sentiment Analysis by many researchers in different levels as in document level, sentence level and phrase level. Now even Twitter tweets which is real time data which keeps receiving data in a bulk manner has some challenges but still many have done the sentiment analysis to an appreciative level and still many doing to get a better accuracy to make the best model for performing sentiment analysis. This Sentiment Analysis is mainly done or performed in order to get a bipolar result, i.e, positive or negative. Sentiment analysis through machine learning techniques [1] using the unigram feature extraction technique where they have applied preprocessor to raw sentences for better understanding. The classifier they have used here is Naive-Bayes classifier which is used to learn the pattern of testing a set of documents that has been classified. Twitter Sentiment Analysis is a bit strenuous because of the shorter length of the tweets, the misspelled words which need to be processed, emoticons and the various abbreviations used with some kind of sentiment attached to it.

Public opinion, sentiment, view and belief can be used interchangeably though there are subtle differences among them [10] [11]. The tweets or the data under review can be real post of the author or reaction by others on that particular text. The classification of text is a quest

in text mining used in applications which requires strategic decision making [12]. Sentiment analysis (SA) is the technique of classifying text and determining the subjective value of a text document that is its positive or negative orientation.

Subjectivity is the main concept involved with the sentiment analysis which is described in [13] as the expressions of others based on their sentiments, emotions or beliefs. Their main idea was to perceive the meaning of the point of view or perspective of the people with respect to a particular matter along with the intention. The perspective could let us understand people's point of view and assessment, their worthwhile circumstance that is what the emotions of various people at the time of checking the viewpoint of them can be known or the required affectional communication. Far more we can note that in this view the word subjective not only means that something is not correct [8]. In analysis of the emotions, going through the subjective language that is the language used to understand a personal occurrence in the context of a text or conversation.

[2] This paper provides a way of sentimental analysis using Hadoop which will process the huge amount on a Hadoop cluster faster in real time, this project uses the naive-bayes approach and an Hadoop cluster for distributed processing of the textual data. They focused more on the speed of performing analysis than its accuracy. They use various processes like stop word removal, unstructured to structured conversion and emoticons. The accuracy of this project was found to be 72.2%. [3] In this paper, the main approaches are pace implementing analysis and also the validity that is totally implementing the sentiment analysis on big data. This exactness is achieved by dividing many parts and register of the performance of following steps combining with HADOOP for mapping it onto various machines.

[4] In this paper, lexicon-based approach is used (looking at each word and sum them up to corresponding scale). Tools are used for subjective lexicom (words of parts of speech), senti word Net 3.0. The aim of the paper is to classify negative, positive and normal tweets. sarcastic tweets were omitted. [5] In this paper, a survey and comparative study of already available techniques for opinion mining along with machine learning and lexicon-based approaches, together with cross domain and cross-lingual methods and some evaluation metrics are used. More accuracy was seen in machine learning approach than in lexicon based approach. [6] In this paper, an algorithm is developed utilizing the contextual information of the words comprising a document. Using this as a foundation we were able to measure the degree to which the words in a document influence each other and the impact of the dynamics of this relationship on the overall document sentiment. On the basis of the results obtained, we conclude that context plays a very significant role in determining a document's sentiment and should not be ignored.

[7] By this paper we could understand about a small survey that they conducted on context based sentimental investigation on opinions based and written in English language and they wide opened the challenges that stood against the evolution of context based sentimental investigation for Arabic content. The main challenges in Arabic Sentiment Analysis are the language itself. one of the main characteristics of Arabic is that a word with the same spelling could have different meaning due to the Arabic punctuation. Those characteristics comprise

the vital issues when inspecting Arabic language, predominantly dealing with the conditions in the process of analysis of sentiment.

Sentiment analysis has gained a lot of popularity in the recent times for knowing how well the product or anything will be accepted by the general public. Sentiment analysis is mainly used to categorize the tweets from twitter into positive, negative and neutral. This polarity of the tweets that we achieve can have innumerable implementations for practical utilization. This can be known from [9] [10]. We can even see that the better the reviews or ratings to a product be, the better the people tend to show some care of buying it or giving value to it.

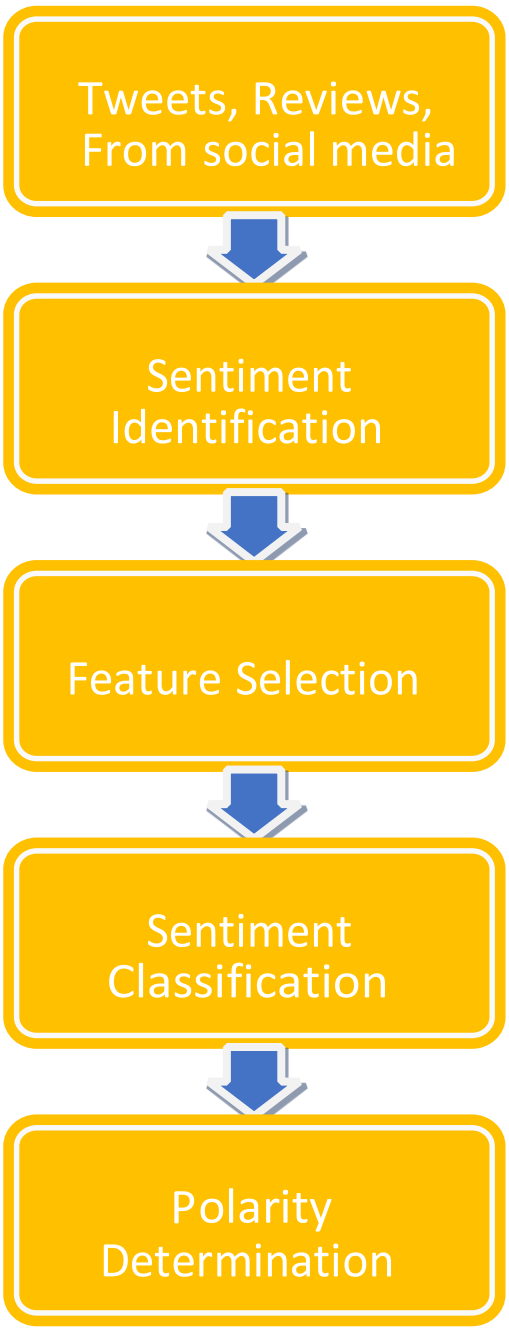

Fig 1: Sentiment Analysis Process

Whereas with negative comments, reviews or ratings the people show less interest towards the product and with that the owners of the product can understand what went wrong with their product and can change upon them and become a better producer by making the changes required by the public. Fig 1. Shows us how the sentiment analysis process is carried on. We can carry out the sentiment analysis by these following steps and achieve the worth of a product or business or political meeting or any other thing.

[14] By this paper we understood that the analysis of sentiments depended mainly two concerns that is sentiment polarity and sentiment score. The polarity of sentiments is a duplex value either positive or negative and the score of the sentiment depends on one of three kinds of models. The kinds of models are Bag with words model (which is also known as bag of words - BOW) [15], parts of speech (POS) [16], and semantic associations [8]. The bag of words model is the most famous among researchers and based on the portrayal of phrases or words, still it abandons language semantics and words ordering. The parts of speech labelling which is a syntactically labelling model especially verbs, adjectives and adverbs [17]. We can take an example "The tap was not nice." mentioning in "The/ET, tap/NO, was/VR not/AV nice/AJ, /.". In this example, ET refers to Determiner, NO refers to Noun, VR refers to Verb, AV refers to Adverb, and AJ refers to Adjective still a semantic associations method is the most tough method, which was regarding the associations between notions or definitions.

III. PROPOSED APPROACH

We have made our approach to perform sentiment analysis through Hadoop framework which is a distributed system for processing of the huge data that we receive. We are basically working on Twitter data, i.e., the tweets that are generated at a bulk each day. These tweets are taken from the twitter and stored on our Hadoop distributed system.

The HDFS (Hadoop Distributed File System) acts as a storage solution for Hadoop framework. For getting the tweets from the twitter we have made use of a tool called flume. Through Flume we have specified a configuration file which has all the properties of how and what kind of tweets need to be loaded into the HDFS and the file path to where the tweets need to be stored in the HDFS.

We have used Apache Flume, Apache Hue, Apache Pig and Pig Script. Apache Flume was used to retrieve the tweets from Twitter using certain configurations on the type of data to receive and the tweets about what we want to retrieve and even the path of HDFS where the retrieved data from twitter has to be stored on Hadoop. We have used Hue file repository as our HDFS for storing our retrieved tweets. These tweets are pre-processed and processed on the pig grunt shell. We have used pig scripting language to script the commands for pre-processing and processing of the tweets accordingly to classify them into positive, negative or neutral.

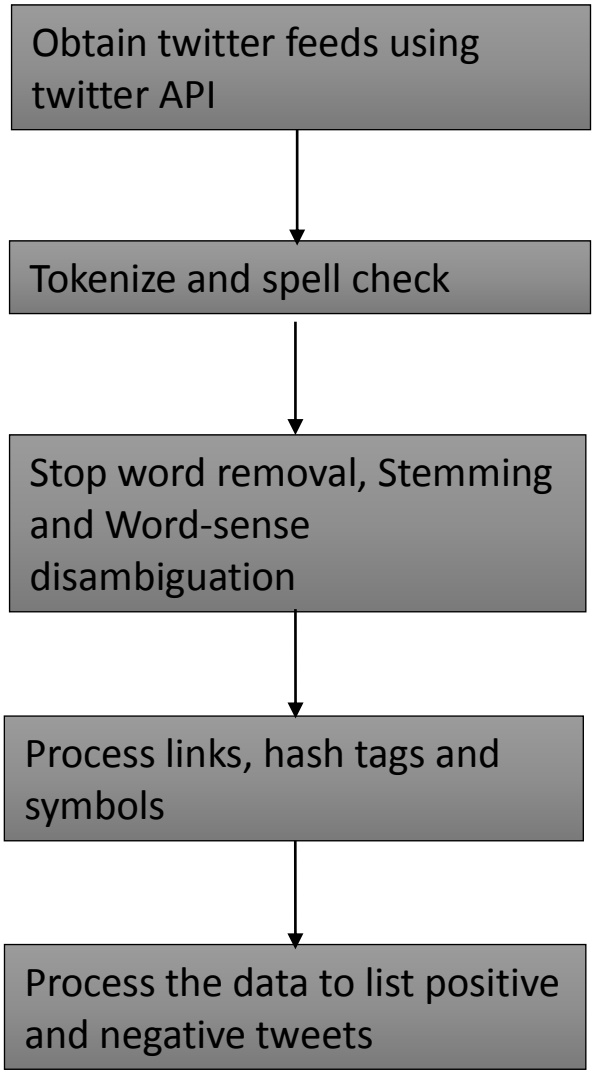

Fig 2: Methodology of Sentiment Analysis Proposed approach for Twitter data

The positive and negative sentiments are calculated and separated to know the number of positive and negative tweets out of the total number of tweets generated from twitter into Hadoop HDFS. We have considered the individual mapping and reduce time along with the number of tweets that are filtered to be positive or negative based on configuration specified. We have used a dictionary of words known as Afinn dictionary which has words that has ratings ranging between -5 and +5 where negative words are given ratings between -5 and 0 (-5 included) and positive words are given ratings between 0 and +5 (+5 included). This dictionary was used to compare the words in each tweet to give rating to tweets individually and based on the average ratings of the words of whole tweet we decided on whether it belonged to positive or negative side.

The total number of words present in this dictionary are 2477 which is of size 28479 bytes.

We have considered two test cases:

i. Data(Tweets) generated from twitter for 15 minutes which is used to calculate its positive and negative tweets constraints.

ii. Data(Tweets) generated from twitter for 30 minutes which is used to calculate its positive and negative tweets constraints.

Test case 1:

Tweets generated for 15 minutes:

Total number of tweets generated = 6067

Size of the total tweets generated = 41287422 bytes

Total number of positive tweets = 1065

Size of the positive tweets = 171108 bytes

Total number of negative tweets = 1184

Size of the negative tweets = 192344 bytes

Average mapping time for positive tweets = 46 seconds

Average mapping time for negative tweets = 44 seconds

Average reduce time for positive tweets = 6 seconds

Average reduce time for negative tweets = 6 seconds

Dictionary mapping time for positive tweets = 7 seconds

Dictionary mapping time for negative tweets = 5 seconds

Dictionary has mapping only.

Test case 2:

Tweets generated for 30 minutes:

Total number of tweets generated = 12766

Size of the total tweets generated = 85369927 bytes

Total number of positive tweets = 2184

Size of the positive tweets = 350270 bytes

Total number of negative tweets = 2210

Size of the negative tweets = 353846 bytes

Average mapping time for positive tweets = 484 seconds

Average mapping time for negative tweets = 561 seconds

Average reduce time for positive tweets = 12 seconds

Average reduce time for negative tweets = 8 seconds

Dictionary mapping time for positive tweets = 6 seconds

Dictionary mapping time for negative tweets = 5 seconds

Dictionary has mapping only.

By the above two test cases we can observe the amount of data that can be processed by Hadoop in certain time depending on the size, map- reduce to be made. We can see in the two test cases that it takes a longer time to process larger data. If we process larger data still Hadoop is efficient enough to process it without any difficulty. Hadoop is scalable. We can process data which is generated for hours and days as well. Hadoop is a distributed system, it can process any amount of data, be it large number of tweets(millions) or large sizes of data (Gigabytes and even more than that). Hadoop (Distributed system) is far more efficient than the stand-alone systems where only samples of data can be taken for processing.

The algorithm that we had used in our project is:

1.Access tweets from twitter using the twitter API and key.

2.Filter and extract the tweets using flume configuration file.

3.Store extracted JSON data in HDFS which acts as a distributed storage solution in Hadoop.

4.Data is processed, queried and analyzed using a data processing tool known as Apache Pig. Links, Hash tags, symbols and many other types of data are processed.

5.Sensitive data is replaced by unique identification symbols which is a process known as tokenization.

6.Spelling are checked to match with related words.

7.Stopwords are removed which can be punctuations, short function words such as the, is, at; and also lexical words such as want.

8.Based on the weightage, the sentiments are decided to be positive, negative or neutral.

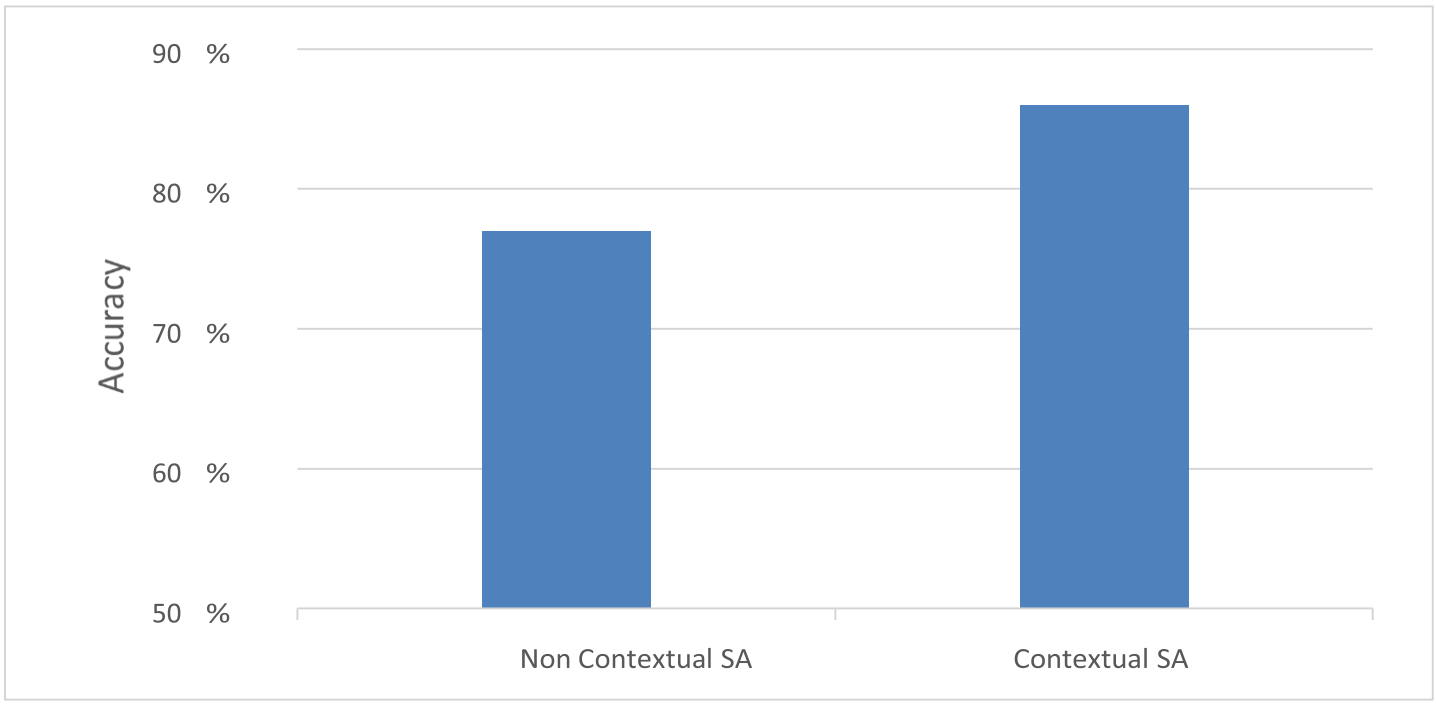

Fig 3: Comparison of accuracy between Contextual and Non Contextual Sentiment Analysis

IV. CONCLUSION

Sentiment Analysis is a very broad area. We have implemented a few tools of Hadoop and extracted, pre-processed and successfully segregated into positive, negative and neutral tweets on the basis of average rating of each tweet by aggregating the rating of each word in the tweet. We have considered the test cases and checked on the reliability and scalability of Hadoop. Hadoop can hold and process any amount of data and has shown great scope of analyzing data for understanding better on the context of people views. We have made better use of Hadoop which is a distributed system where not only based on sample data the sentiment analysis takes place but based on a large stream of data as compared to the standalone system.

V. FUTURE WORK

Sarcasm handling has been a problem. Ontology can be built based on the words and using the concept of semantic technology we can process the sarcasms better and the perfect sentiments of the people can be known.

REFERENCES:

[1] Geetika Gautam, Divakar Yadav. Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis. In *Seventh International Conference on Contemporary Computing* (IC3), 2014.
[2] Sunil b. mane, Yashwant sawant, Saif kazi, Vaibhav shinde. Real time sentimental analysis of twitter data using Hadoop. In *International journal of Computer Science and Information Technologies*, Vol. 5(3), College of Engineering, Pune, India, 2014.
[3] Divya Sehgal and Dr. Ambuj Kumar Agarwal. Sentiment Analysis of Big Data Applications using Twitter Data with the Help of HADOOP Framework. In Proceedings of the SMART -2016, 5th *International Conference on System Modeling & Advancement in Research Trends*, 25th-27th November, 2016, College of Computing Sciences & Information Technology, Teerthanker Mahaveer University, Moradabad, India, 2016.
[4] Bakliwal, Akshat, Foster, Jennifer, van der Puil, Jennifer, O'Brien, Ron, Tounsi, Lamia and Hughes, Mark. Sentiment analysis of political tweets: towards an accurate classifier. In: NAACL Workshop on Language Analysis in Social Media, 13 June 2013, Atlanta, GA.
[5] Vishal A. Kharde and S.S. Sonawane. Sentiment Analysis of Twitter Data: A Survey of Techniques. In *International Journal of Computer Applications* (0975 – 8887) Volume 139 – No.11, April 2016, Pune Institute of Computer Technology, Pune, University of Pune (India), 2016.
[6] Srishti Sharma, Shampa Chakraverty and Akhil Sharma. A context-based algorithm for sentiment analysis. In International Journal of Computational Vision and Robotics, Vol. 7, No. 5, 2017, Netaji Subhas Institute of Technology, Dwarka, New Delhi, India, 2017.
[7] Oumayma El Ansari, Jihad Zahir, and Hajar Mousannif. Context Based Sentiment Analysis: A Survey. LISI Laboratory, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech, Morocco.
[8] Richard, S., Alex, P., Jean, Y.W., Jason, C., Christopher, D.M., Andrew, Y.N. and Christopher, P., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", Conference on Empirical Methods in Natural Language Processing, 2013.
[9] Kumar A, Sebastian T M. Sentiment analysis on twitter. IJCSI Int J Comput Sci Issues 9(4):372, 2012.
[10] Kumar A, Teeja M S. Sentiment analysis: a perspective on its past, present and future. Int J Intell Syst Appl 4(10):1, 2012.
[11] Pang B and Lee L . Opinion mining and sentiment analysis. Found Trends Inf Retr (1–2), 1–135, 2008.
[12] Kumar A, Dabas V, Hooda P. Text classification algorithms for mining unstructured data: a SWOT analysis. Int J Inf Technol, 2018.
[13] Wiebe J, Wilson T, Bruce R, Bell M, Martin M. Learning subjective language. Comput Linguist 30(3):277–308, 2004.
[14] Liu, B. Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.
[15] Yin, Z., Rong, J and Zhi-Hua, Z. Understanding Bag-of-Words Model: A Statistical Framework. In *International Journal of Machine Learning and Cybernetics*, 2010.
[16] Nisha, J., & Dr.E. Kirubakaran. M-Learning sentiment analysis with Data Mining Techniques. In *International Journal of Computer Science and Telecommunications*, Volume 3, Issue 8, 2012.
[17] Farah, B., Carmine, C., & Diego R. Sentiment analysis: Adjectives and Adverbs are better than Adjectives Alone. In *International Conference on Weblogs and Social Media -ICWSM* , Boulder, CO USA, 2007.

