DISCLOSING TWEET POLARITY USING FEATURE REPRESENTATION PRACTICE

Dr. Seema Verma
Associate Professor
Department of Electronics
Banasthali Vidhyapeth, Jaipur, Rajasthan, India

Dr. Manisha Sharma
Associate Professor
Department of Computer Science
Banasthali Vidhyapeth, Jaipur, Rajasthan, India

Mrs.Divya Rajput
Head of Centre for Business Innovation
Indian Institute of Corporate Affairs, Gurgaon, Haryana, India

Mr.Mani Madhukar
Technical lead
IBM, Bangalore, Karnataka, India

Ms.Vani Mittal
M.Tech Intern
Indian Institute of Corporate Affairs, Gurgaon, Haryana India

Ms. Rashika Singh
M.Tech Intern
Indian Institute of Corporate Affairs, Gurgaon, Haryana India

Abstract-With the establishment of Internet and the World Wide Web, one has experienced a substantial escalation of data on the web. But in the past decade, novel variety of communication, such as Twitter, Facebook, Blogs, have emerged and have become a medium through which public not only convey messages but also expresses their sentiments or feelings regarding activities happening all around the globe. The concept of cataloguing these views of people as positive, negative or neutral will provide us with the insights which can prove to be a cutting edge for various organizations. In this paper, we are focusing on performing classification of twitter messages with the help of Feature Selection (Bag of Words) and Feature Weighting (Term Frequency-Inverse Document Frequency) practices. The two different datasets has been fetched which will consist of tweets on two topics i.e. iphone and samsung mobile phones respectively and hence, prediction will be made on the basis of the results of experiments which will be portrayed in the form of histograms and a word cloud.

Keywords: - Sentiment analysis, Bag of Words, TF-IDF, Twitter.

I. INTRODUCTION

Sentiments are feelings not facts and their classification aims at deciphering feelings hidden in words or sentences said by a speaker or written by a writer on a particular topic. For example, a comment such as “The film was neither that funny nor witty” can be marked as negative, positive or neutral. After interpreting previous discussions from the history of the conversation, the said sentence can be labelled as positive, negative or neutral. So, on the basis of emotions which are excavated from our evaluation of an event that cause explicit reactions in different people (appraisal theory), we can identify overall polarity of a document can be identified and predictive analysis can be performed as well.

The essential task in sentiment analysis is of sorting the ‘polarity’ of a given text at the document, sentence, or feature/aspect level i.e. divisions of opinion in three groups that are positive, negative or neutral. Beyond ‘polarity’ division takes place at emotional states such as happy, sad, angry etc. But in this literature, sentiment analysis at feature level has been discussed for which feature selection and feature weighting methods will be described.

A. FEATURE SELECTION

Feature selection is also known as ‘attribute selection’ or ‘variable selection’. As our dataset contains many unwanted and unrelated elements, it can create issues while analyzing the sentiments of any specific dataset. So, feature selection techniques assist in removing those components i.e. it is the mechanism of selecting a subset of significant features which will be used as a part of the data analysis practice, as it reveals which features are essential for prediction, and how these features are related. They also help us in the creation of effective predictive models by reducing training time and improving interpretability. Among so many methods, our focus in this literature will be on BOW (Bag of Words) tagging.

- **Bag of Words (BOW)** – BOW is the most famous method for object classification and is used widely in Information retrieval and natural language processing. In this model, a text (document/ sentence) is characterized as the ‘bag’ (multiset) of its words, ignoring word order but maintaining diversity and sentence structure. Further, the existence of word is utilized as a feature for training a classifier. It can enhance the efficiency of the classifier and better outcomes can be obtained.

B. FEATURE WEIGHTING

The ability of every feature to distinguish document is different, and this ability can be measured by feature weighting **(Zhao, 2014**). In this process, a weight is provided to the relevant features according to their occurrence in the dataset. Although there are many methods like Unigram Features, Feature Presence (FP), Feature Frequency (FF) etc but we’ll be concentrating on TF-IDF (Term frequency- Inverse Document Frequency) technique as it is the best weighting factor in text mining and information retrieval **(Keefe, 2009).**

- **Term Frequency-Inverse Document Frequency (TF-IDF)** – It is an arithmetical statistic that is used for revealing how significant a word is to a document in a dataset. The TF-IDF value increases relatively to the number of times a word occurs in the document, but is offset by the frequency of the word in the dataset, which helps to adapt for the fact that some words are present more commonly, in general. It can be productively used for stop-words filtering in many topics including text summarization and categorization.

a. TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document).
b. IDF (t) = loge (Total number of documents / Number of documents with term t in it).
c. TF.IDF

II. PREVIOUS WORK

The following literature review focuses on various established theories, procedures and techniques of POS tagger (feature selection method) and TF-IDF (feature weighting method) which is able to explain their usage for sentiment analysis on twitter dataset effectively.

In the research work of sentiment analysis many of the application of sentiment analysis has become infeasible due to huge number of features. Feature selection and weighting method by ***Tim O Kfee et.al., (2006)*** has involved a range of feature selector and weighting methods for various classifiers. To define the sentiment of movie review dataset very few features provided the accurate information and hence three feature selection and six feature weighting methods used. Feature selection methods are- Categorical Proportional Difference(CPD) which tell us how close to being two equal numbers are, Sentiword Net Subjective Scores(SWN-SS) defined sentiment of subjective terms and removed unigram having subjectivity score less than threshold. Sentiword Net Propotional Difference (SWN-PD) defined consistent and meaningful sentiment. Further six feature weighting methods were- (i)Unigram feature where each unique word considered as feature, (ii)Feature Frequency defined each unigram feature value,(iii) Feature Presence for unigram that existed in the document ignoring multiple occurrences, Sentiword Net Word Score Groups(SWN-SG), (iv)Sentiword Net Polarity Groups(SWN-PG),(v)Sentiword Net Polarity Sums(SWN-PS), (vi)Term frequency. Proportional difference (PD) and SWN-PD tend to achieve higher accuracy using machine learning techniques with 87.15% accuracy.

A. BAG OF WORDS (BOG)

In the researcher article by **Yessenov & Misailovic (2009),** an analytical study of a movie review comment dataset is performed taken from an accepted social network. The texts are classified with the help of different machine learning techniques like K-mean clustering, Max Entropy Navie Bayes and Decision Tree. The BOW method is used in a feature selection process and Word Net is used for finding relations between the words of a text. As a result, plain bag-of-words model has performed relatively well, and in addition, it can be refined by the selection of features based on linguistic and semantic data from the text.

The research article of **Mukherjee & Bhattacharyya (2012)** has anticipated a lightweight method for the detection of polarity of tweets in which connectives & conditionals have been proposed for integrating discourse data in bag-of-words model, to perk up opinion cataloguing accuracy. As a consequence, bag-of-words model carries out well in a noisy medium as well as accomplishes enhanced outcome than an existing Twitter-support function. Furthermore, this article focuses on their approach which is favourable to prearranged views also. The researchers have compared their proposed practise with other systems and stated that their procedure is less intensive than the present ones.

In the research article by **Jha & Khurana (2013),** emotion labeling of informal text genres like twitter has been undertaken. In the experiment, BOW has been utilized for selecting the features so that frequency can be improved. Moreover, SVM and NB classifier is also employed on the features and as an outcome 68.36% accuracy has been achieved from the combination of SVM and BOW.

The research paper of **Marchand *et. al.,* (2013)** presents the work done by his team at task 2 of SemEval 2013. There are two subtasks- Contextual Polarity Disambiguation and Message Polarity Classification; the researcher applies bag of words with uni, bi, and tri technique in the second subtask with BoosTexter classifier for the categorization of the tweets. As a result, with 4899 training tweets, they have attained fine outcomes and exhibited that words with varying division can manipulate the classification performance as well.

The routine of a classifier banks on characteristic representation, optimization and regularization. In the research article by **Assefa (2014),** bag-of-words and sentiment lexicon features have been applied. This article presents a conclusion mentioning that the stemming procedure of the terms increases the accurateness of the classifier. The accuracy has shown an increase of 6% from the baseline with 95% confidence interval. A last, the procedure for the methods has been provided by the researcher in detail.

B. Term Frequency- Inverse Document Frequency (TF-IDF)

More than 100 million people around the globe use twitter as a medium for sharing their views with one another. It points out certain well-liked topics from such a wide variety of subjects and they are named as ‘Trending Topics’. In the research article by **Benhardus (2010),** certain methodologies and approaches through which latest issues (Trending Topics) that can be obtained from streaming twitter data has been addressed. Two methods i.e. TF-IDF and relative normalized term frequency, both have been implemented on two respective dataset of Twitter Streaming API and the Edinburgh Twitter corpus, a compilation of nearly 97 million tweets collected between November 2009 and February 2010. The Edinburgh Twitter corpus was used to provide benchmark measurement against the data from the Twitter Streaming API. The procedure is discussed in immense detail and the result shows that recognition of trending topic was more successful with the usage of TF-IDF approach.

In the research article by **Paltoglou & Thelwall (2010),** opinion investigation can be done more accurately by using TF-IDF approach as researchers have tested the same on a wide variety of selected datasets. The researchers also suggest a modified TF-IDF named as ‘Delta TF-IDF’. Experimentation has been performed on three datasets respectively and as a result this approach has provided the increase in classification performance and hence did not require any human annotation or external sources.

The research paper by **Sharifi (2010),** presents algorithms that summarizes the collection of short messages on a subject. The objective is to fabricate summaries that are parallel to what a human would produce for the same compilation of posts on a definite topic. The researchers have evaluated the summaries generated by many summarizing algorithms which also include hybrid TF-IDF method as well. Then, these summaries are compared with human-produced abstracts in which TF-IDF algorithm generates sentences instead of phrases for summaries and threshold of 11 words is being used as a normalization factor. This algorithm produces an average recall of 0.31, an average precision of 0.34, and a combined F1-Measure of 0.33. These values are very close to the performance levels of our manual summaries of 0.34. Therefore, algorithm of tf –idf is very effective in analyzing tweets.

The research literature of **Ostendorf (2010)** pioneers a system designed for automatically creating personalized annotation tags to mark interests and concerns of twitter users. Evaluation of two tagging algorithm is considered which are TFIDF ranking and Text Rank which has been applied to extract keywords from Twitter messages to tag the user. Results shows that Text Rank outperformed TFIDF ranking, but both gave a tagging precision that was comparable to that reported for web page advertizing keyword extraction.

The research article by **Gebremeskel (2011),** interests in what nation are articulating about news in Twitter. This paper has been divided in sections through which researcher describes about methodologies and approaches, collection and pre-processing of data, experiments and results and finally, analysis and conclusion. As a part of experimentation, feature presence and TF-IDF has been used. It is said that the use of presence (absence) or count (frequency) changes the probabilities during learning of the model and the effect on probabilities affects the prediction of the learning algorithm used. Therefore, TF-IDF can be used for sentiment analysis for twitter dataset.

In the research paper of **Aliandu (2013),** purpose is to build a function that can verify public reaction on Indonesian tweet by user query. The method used in this research is Naïve Bayes. Emoticons are employed to make sentiment class annotation easier. Term frequency and TF-IDF are used as weighting factors. All of the data used in this research is Indonesian twitter data. The results obtained an accuracy of 77,45% using term frequency and 75,86% using TF-IDF on test set that annotated by emoticons. The results of manually marked test set are 70.68% for term frequency and 71.26% for TF-IDF. The accuracy measurement also used Support Vector Machine (SVM). The results obtained an accuracy of 77.79% using term frequency and 77.57% using TF-IDF. Hence, accuracy of Support Vector Machine is better than Naive Bayes.

In the research paper by **Samoylov (2014),** a feature model is proposed for judging the emotions of the short texts or tweets with the help of Delta TF-IDF approach. The main motto is to improve the quality of prediction by integrating rule based approach and standard bag of words model. Moreover, methodology is provided by the researcher and the experiment has been performed on a tweet dataset. The result shows that delta TF-IDF should be used for analyzing short messages.

C. BOG + TF-IDF

Opinion mining which is the part of Natural language Processing (NLP) incorporates both feature selection and feature weighting factors significantly. So, these should be discussed together for finding best practise among both of them. So, in the research article by **Fiaidhi (2012),** tweets have been sorted with the help of programming R. The BOG has been used and TF-IDF & LDA (Latent Dirichlet Allocation) procedures are used for the identification of the polarity of tweets on six NHL hockey teams. At last, sentiment analysis can be performed efficiently following the procedures mentioned in the paper.

III. METHODOLOGY

Sentiment Analysis of twitter data (tweets) is about analyzing the frame of mind of people belonging to different spheres of life on twitter. One can get a number of tweets, containing keyword one can define, filter out the text of these tweets and then see if there are more positive or negative words. In this paper, we'll be focusing on machine based method i.e. with the help of tool called **R**. It is a free "software environment for statistical computing and graphics" and is available for UNIX platforms, Windows and Mac OSs. It is used because it provides a wide range of options with the help of which one can perform experiments effectively. Now, the procedure consist of the tweets which are fetched using **#search twitter function** and stored separately as two different datasets. The collected tweets are scored on the basis of dictionary of positive and negative words. Thus, the categorization of tweets as positive, negative or neutral is based on the approach of **Bag of Words** (BOW) scheme. To further improve the scoring of tweets, we perform text mining, as all the words in the tweet datasets are not equally important. There are certain features or terms which are more relevant than others. Hence, we utilize **Term Frequency-Inverse Document Frequency** (TF-IDF) weighting technique for the evaluation of mined features and minimize the influence of those terms which are very common and hence carry no meaning. Finally, terms are fetched on basis of defined threshold and it is plotted accordingly.

IV. DATA TYPE

We'll be using tweets which will be directly fetched from the twitter account and will be stored in CSV format. Emoticons are omitted during initial steps (pre-processing of data). Moreover, we are fetching the dataset with the help of a keyword 'hash' (HASH, #) tag. Here, two dataset will be considered and those are **#iphone** and **#Samsung**, respectively. Hence, tweets that are fetched are unbalanced in nature i.e. number of positive tweets and negative tweets are not uniform in the dataset.

V. EXPERIMENT

Our aim for conducting this experiment is to perform sentiment analysis on the twitter messages which are on the topic iphone and samaung respectively. In software R, two datasets which comprise of tweets on both the topics are fetched respectively. After loading all the packages, a new app is created which is named as 'Myappsentitweet'. Moreover, two unique keys are generated i.e. 'consumer key' and 'consumer secret key' through which tweets are accessed in twitter. Now, for beginning the search on #iphone and #Samsung respectively, we have specified the parameter 'n' (length of tweets) which equals to 2000 for both the dataset. For both the dataset, twitter API fetched 2000 tweets. These are stored in CSV file format and evaluation of stored tweets is done using score.sentiment function which will assign a score to each tweet individually and get stored in a new CSV file. Furthermore, a histogram of the scored tweets is created using a function 'hist' for the dataset respectively**.** Finally, for both the dataset, histogram is created respectively as well as a graph of most frequent words occurring in tweets are also shown and at last, comparison cloud is created with the help of frequent words.

VI. RESULTS

In software R, after running text mining algorithm on #iphone and #Samsung datasets respectively, firstly, we get a plot of most recurrent words (Figure 1 & 2) i.e. terms that have frequency equal to 500 or more which can also help us in recognizing response of the public for both the topics.

Figure 1: Plot of Frequent Words of #iphone

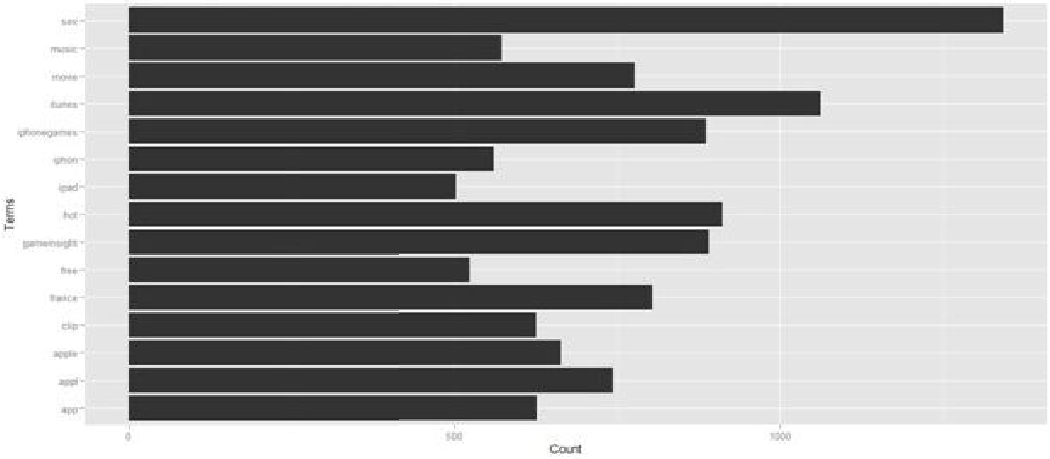
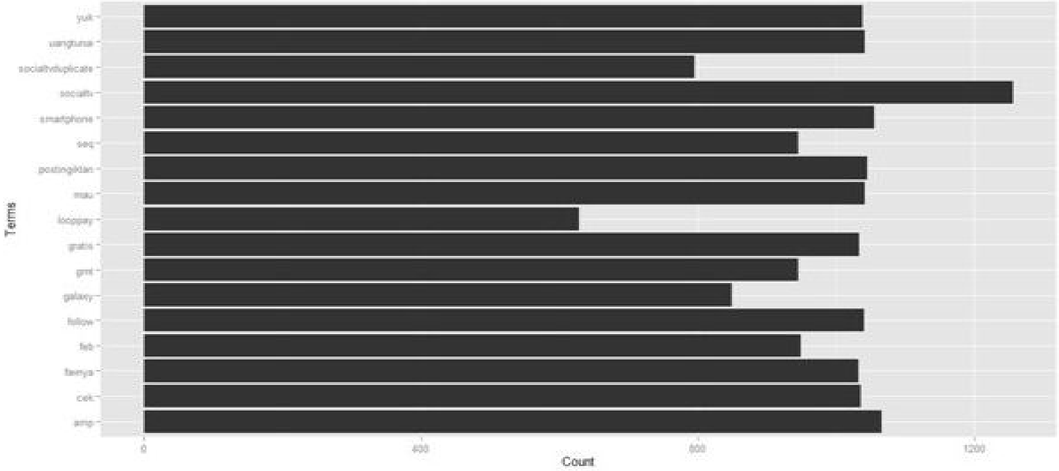


Figure 2: Plot of Frequent Words of #Samsung



Moreover, the histograms created for #iphone and #Samsung dataset respectively illustrates the occurrence of tweets with respect to scores allotted to each tweets. The x-axis shows the total score of tweets as a negative, positive or zero. A positive score represents positive or good sentiments associated with that particular tweet whereas a negative score represents negative or bad sentiments associated with that tweet. A score of zero indicates a neutral sentiment (**Rais, 2014**). The Fig.3 is skewed towards Zero score in which scoring is done from -2 to 5, tells us about sentiments of people which are neutral regarding keyword #iphone, whereas Fig.4 in which scoring has been done from -3 to 3 with the same interpretation of axis, tells us that people are having neutral attitude for it also.

Figure 3: Histogram of #iphone

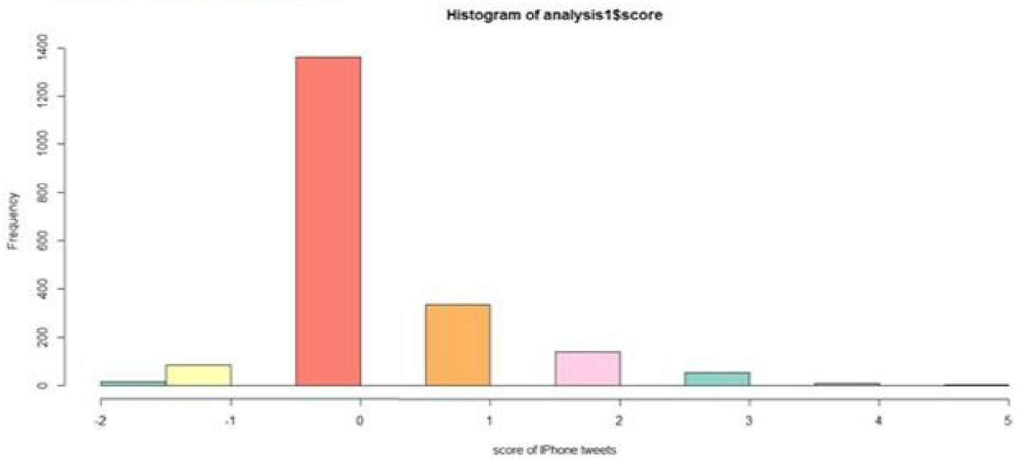
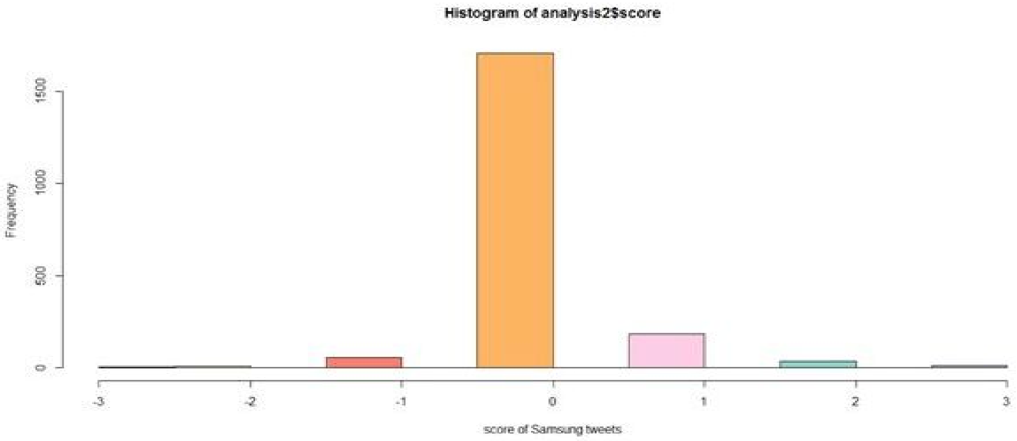


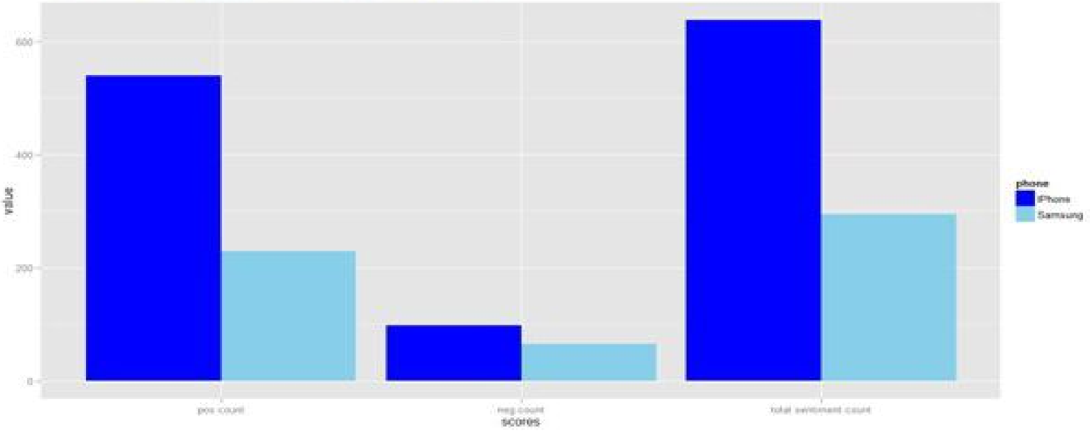
Figure 4: Histogram of #Samsung



Furthermore, a comparison plot is also created among #iphone and #Samsung which briefly shows the difference between positive and negative tweets of both the dataset and variation between total numbers of tweets (positive + negative) as well.

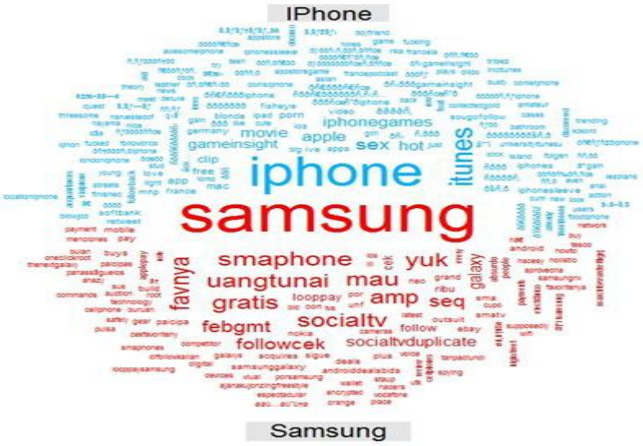
It tells us that people are giving more affirmative reaction to iphnone as compared to Samsung mobiles.

Figure 5: Comparison between #iphone and #Samsung



Finally, a word cloud (Comparison Cloud) is created with the help of text mining algorithm for the dataset of #iphone and #Samsung together which shows that words like itunes, smartphone, apple, samsung, etc which are more used in the tweets posted by the people.

Figure 6: Word Cloud of #iphone and #Samsung



VII. CONCLUSION

We carried out an experiment using software R on a well accepted microblogging site, Twitter. The tweets are directly obtained from the twitter account so, two different datasets has been retrieved but fetching was made with the help of different keywords. We used ‘hash’ (HASH, #) tags for it and we found that firstly, while requesting tweets for a particular keyword, it sometime happens that the number of retrieved tweets is less than the number of tweets that has been requested. Secondly, while requesting tweets for a particular keyword, the older tweets cannot be retrieved. By following various steps mention in this article, one can easily obtain a histogram of the dataset as well as achieve most frequent words and association among them. Moreover, tweets that have been obtained are more neutral in nature, but natives are more inclined towards iphone optimistically than samsung because it got more positive tweets than latter. Hence, iphone are more preferred than Samsung relatively.

REFERENCES

[1]Y. Zhao, S. Dong, L. Li, “Sentiment Analysis on News Comments based on Supervised Learning Method”, *International Journal of Multimedia and Ubiquitous Engineering*, Vol. 9, pp 333-346, 2014.

[2]J. Benhardus, “Streaming Trend Detection in Twitter”, *UCCS REU for artificial intelligence, natural language processing and information retrieval final report*, 2010.

[3]G. Paltoglou, M. Thelwall, “A study of Information Retrieval weight schemes for sentiment analysis”, 2010.
[4]T. O'Keefe, I. Koprinska, "Feature selection and weighting method in sentiment analysis". *In Proc of the Australasian document computing symposium*, pages 67-74, 2009.
[5]B. Sharifi, M. A. Hutton, J. K. Kalita, “Experiments in Microblog Summarization”, 2010

[6]A. Bakliwal, “Mining sentiments from tweets”, *3rd Workshop on Sentiment and Subjectivity Analysis (WASSA) in Conjunction with 50th annual meeting of Association for Computational Linguistics (ACL)*, 2012.

[7]L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, B. Liu, “Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis”, June 21, 2011.

[8]N. S. Joshi, S. A. Itkat, “A Survey on Feature Level Sentiment Analysis”, *International Journal of Computer Science and Information Technologies*, Vol. 5 (4), pp 5422-5425, 2014.

[9]L. Zhang, “Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation, 2013.

[10] G. Angulakshmi, R. ManickaChezian, “An Analysis on Opinion Mining: Techniques and
Tools”, *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 3, Issue 7, July 2014.

[11]F. Liu, D. Pennell, F. Liu, Y. Liu, “Unsupervised Approaches for Automatic Keyword Extraction Using Meeting Transcripts, Human Language Technologies”, *Annual Conference of the North American Chapter of the ACL*, pages 620–628, 2009.

[12]P. Aliandu, “Sentiment Analysis on Indonesian tweet” *Proceedings of the 7th ICTS, Bali*, May 15th-16th, 2013.

[13]S.H Kumar, “Twitter Sentiment Analysis”, *CMPS 242 Project Report,* 2013.

[14]G. Gebremeskel, “Sentiment Analysis of Twitter Posts About News”, 2011.

[15]W. Wu, B. Zhang, M. Ostendorf, “Automatic Generation of Personalized Annotation Tags for Twitter Users”, “*The Annual Conference of the North American Chapter of the ACL*”, pages 689–692, June 2010.

[16]A. B. Samoylov, “Evaluation of Delta TF-IDF Features for Sentiment Analysis”, *Springer International Publishing*, 2014.

[17] J. Fiaidhi, O. Mohammed, S. Mohammed, “Opinion Mining over Twitterspace: Classifying Tweets Programmatically using the R Approach”, *Institute of Electrical and Electronic Engineers*, 2012

[18] R. K. Jha, S. Khurana, “Sentiment Analysis in Twitter”, *CS 671: Natural Language Processing*, 2013

[19] B. G. Assefa “KUNLPLab:Sentiment Analysis on Twitter Data”, *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 391–394, Dublin, Ireland, 2014.

[20] M. Marchand, A. L. Ginsca1, Romaric Besanc, O. Mesnard, “Using Syntactic Features and Multi-polarityWords for Sentiment Analysis in Twitter*”, French National Research Agency(ANR),* 2013

[21] M. A. Hameed, A. R. Hussain, S. F. Sayeedunnissa, “Sentiment Analysis using Naïve Bayes with Bigrams*” Proc. of the Intl. Conf. on Advances in Computing and Information Technology*,2014.

[22] S. Mukherjee, P. Bhattacharyya, “Sentiment Analysis in Twitter with Lightweight
Discourse Analysis” December 2012.

[23] Online Source: *https://www.cs.tau.ac.il/~nin/Courses/Workshop13a/TF-IDF.pdf*

[24] Online Source: *file:///F:/attachments%20(2)/New%20folder/What%20is%20TFIDF%20%20%20Moz%20Q&A.html*

[25]Online Source: *cs.wellesley.edu/~cs315/315_PPTs/L10-VectorSpace/CS349-TF-IDF.ppt*
[26]Online Source: www.umiacs.umd.edu/~hwa/POS.ppt
[27]Online Source:
www.coli.uni-saarland.de/courses/korbay/.../TokenizationPOS-Tagging
[28]Online Source: https://www.kaggle.com/c/word2vec-nlp-tutorial/details/part-1-for-beginners-bag-of-words
[29]Online Source: - Kafy Rais, “Twitter Analysis-(In Rstudio using R programming language)”, in.linkedin.com/pub/kaify-rais/31/346/886/, Oct 6. 2014.