ISSN No. 0976-5697

Volume 5, No. 7, September-October 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Framework for Opinion Access Service To Make Smart Business Decisions - Twitter as a Case Study

Akshata B. Angadi
Dept. of Computer Science and Engineering
K.L.E. Institute of Technology
Hubli, India

Karuna C. Gull
Dept. of Computer Science and Engineering
K.L.E. Institute of Technology
Hubli, India

Abstract: Twitter has grown to be a foremost influence in our daily lives. Twitter has emerged from online short communication network into an increasingly useful platform for the teenagers, marketers, researchers, scholars and many more. It is now on the apex of its colossal spurt. It has become a popular micro blogging site where people can share their experiences, views, thoughts and events. This voluminous data of twitter shared by people helps in making decisions and analyze the opinions of them. Sentiment Analysis, a part of Opinion analysis help in finding the current status of particular product based on which the marketers can improve the business quality i.e. public perception can also be known by tracking their interests thus help in framing a useful pace to predict future of a product. The paper focuses on Analysis of data from the SNS i.e. twitter and also details simple steps to increase the proficiency of the business using Data Mining Techniques. Survey on many research papers is carried out and designed the framework. This paper acts as the stepping stone for the efficient development of twitter mining tools. Further it helps in implementing improved algorithm for the betterment of customers' insight on/towards product.

Keywords- Clustering, Data Mining, NLP (Natural Language Processing), Twitter.

I. INTRODUCTION

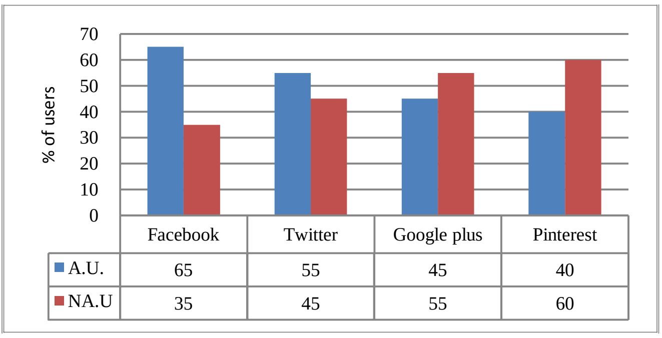
Conventionally, Search Engines enable users to locate the information or to look for a site to fulfill a specific purpose. To meet the intent of people its necessity to analyze the information from different search engines. On second thought, Twitter, a micro blogging social network was originally an internal service for employees of Odeo, gives an opportunity for users to get any current information especially the updates, reviews made by people. I can say Twitter as a kind of discussion forum these days, where people love to tweet their thoughts, update their status or talk about the release of new smart phone, comment on it, and report about the working with comparison. People includes actors, politicians, players etc discuss or share about one or the other stuff. Twitter with its short and hasty messaging is reliant on smart phones usage.

Fig.1 shows the statistics of users using popular social sites like Facebook, Twitter, Google plus and Pinterest as per survey. A.U represents Active users and NA.U represents Non Active Users respectively. By surveying n number of articles I can conclude that the users of twitter not only include teenagers alone but the percentage of adults is about 100% in popular sites like Facebook, Twitter and Google+.

Twitter has attracted many users by its exquisite features in a short span. The benefits include Tweeting and Retweeting facility, Adding Favorites, Following people, Sharing, Can find the news, track the Current Trends, Companies, Contacts, Celebrities and many more to pen. Thus a rich source of data is available on twitter due to the increased usage. This rich data help in knowing users passion and excitement towards particular product. [1] Twitter users generate more than 300M tweets each day, these users are also overwhelmed by the massive amount of information available and the huge number of people they can interact with.

The data is collected from Twitter and is categorized based on its feature specified in methodology and is dumped in particular cluster. The clustered data is further processed to analyze the opinions based on Mining Technique. In our paper a framework has been designed to improve the tactics of business.

The paper is organized as follows: There are many research papers on sentiment analysis. So in Section 2, Survey draws the details of few papers among them and the tools, features they defined can be added on to get the best result and high accuracy. The framework proposed is described in Section 3. Section 4 features the expected output of the methodology defined. The conclusion is in Section 5.



	Facebook	Twitter	Google plus	Pinterest
■ A.U.	65	55	45	40
■ NA.U	35	45	55	60

Figure 1. Percentage of User statistics as per survey

II. LITERATURE SURVEY

Twitter has become an exclusive SNS that is chosen for every updates over the world. It is a place where people gather and confer their interests. And analyzing the comments, shares, favorites' of the users help to realize the influential users and track their interests.

Survey showed that in sentiment analysis the tools are developed using lexicon or learning based resources. Lexicon methods use a glossary to perform object-level emotion analysis where as Learning resource is creating a model by training the classifier and specifying whether the statement is positive, negative or neutral. Here we discuss few of the tools using different techniques and way to analyze the opinions:

Ilia Chetviorkin et.al (2014) [2] proposed a novel technique for creation of polarity lexicons from the Twitter streams in Russian and English. Advantage of the tool is once system is trained it can be utilized to different domains and languages with minor modifications. To demonstrate the ability of the proposed algorithm to extract sentiment words in various domains they have investigated datasets on different collections for training and testing of tweets.

Robert Remus (2013) [3] described University of Leipzig's contribution to SemEval-2013 task 2B on Sentiment Analysis in Twitter. The approach used one-against-one Support Vector Machines with asymmetric cost factors and linear "kernels" as classifiers, word uni- and bigrams as features and additionally model negation of word uni- and bigrams in word n-gram feature space. He stated proposed system achieves (+,−) averaged F-Scores of 0.6902 and 0.6846 on Twitter and SMS test data, respectively. And concludes polarity classification of tweets and short messages still proves to be a difficult task that is far from being solved.

Prabu Palanisamy et al. (2013) [4] described the system developed by the Serendio team for the SemEval-2013. The experiment investigated on the test dataset yields an F-score of 0.8004. They have presented a lexicon based method for Sentiment Analysis with Twitter data and provided practical approaches to identify and extract sentiments from emoticons and hash tags, also provided a method to convert non-grammatical words to grammatical words and normalize non-root to root words to extract sentiments.

A. Tamilselvi1 et al. (2013) [5] discusses an approach where a stream of tweets from the Twitter micro blogging site are preprocessed and classified based on their feature content as positive, negative and irrelevant; and analyses the performance of various classifying algorithms based on their precision and recall in such cases. The limitations of this work include the time required for queries and, for some applications, the level of accuracy that was achieved.

Ankit Bhakkad et al. **(**2013**)** [6] proposes a fast and efficient algorithm to calculate frequency of all existing bigrams in given text document. Freq count Algorithm exploits the positions of the words stored in E-VSM to find bigram frequency. Analysis shows that it improves the time complexity to O(n) as compared to O(n2) of commonly used approach. Author concludes saying experiment results prove analysis.

Michael Speriosu et al. (2011) [7] experimented on Emoticon-based training set, three annotated datasets i.e. STS, OMD and HCR. Experiment showed that a maximum entropy classifier trained with distant supervision works better than a lexicon-based ratio predictor, improving the accuracy for polarity classification on our held-out test set from 58.1% to 62.9%. By using the predictions of that classifier in combination with a graph that incorporates tweets and lexical features, we obtain even better accuracy of 71.2%.

Efthymios Kouloumpis et al. (2011) [8] investigated the utility of linguistic features for detecting the sentiment of Twitter messages. They used a supervised approach to the problem, but leverage existing hashtags in the Twitter data for building training data. They concluded saying part-of-speech features may not be useful for sentiment analysis in the microblogging domain. And stated research is needed to determine whether the POS features are just of poor quality due to the results of the tagger or whether POS features are just less useful for sentiment analysis in this domain. Experiments showed when microblogging features are included; the benefit of emoticon training data is lessened.

Lei Zhang et al. (2011) [9] combined lexical and learning based methods for twitter analysis. They have compared experimental results with different sentiment analysis methods like learning-based method used by the website Twitter sentiment", which uses Maximum Entropy(ME) as the supervised learning algorithm, FBS-a lexicon-based method, AFBS - the augmented lexicon-based method for Tweets and LLS. They showed the experimental results too.

Rudy Prabowo et al. (2009) [10] combined rule-based classification, supervised learning and machine learning into a new combined method. The method is tested on movie reviews, product reviews and MySpace comments. The results showed that a hybrid classification can improve the classification effectiveness. By using a Sentiment Analysis Tool (SAT), we can apply a semi-automatic, complementary approach, i.e., each classifier contributes to other classifiers to achieve a good level of effectiveness.

Bo Pang et al. **(**2004) [11] examined the relation between subjectivity detection and polarity classification. And said for the Naive Bayes polarity classifier, the subjectivity extracts are shown to be more effective. And showed another interesting point i.e. minimum-cut framework results in the development of efficient algorithms for sentiment analysis. Utilizing contextual information via this framework can lead to statistically significant improvement in polarity-classification accuracy.

Soo-Min Kim et al. (2004) [12] automatically finds the people who hold opinions about that topic and the sentiment of each opinion. The system contains a module for determining word sentiment and another for combining sentiments within a sentence. Experiments are conducted using word and sentence sentiment classifiers.

The above articles have investigated different techniques and showed the results. Few worked on lexicon, graph based. All have given their ideas based on the results drawn.

Twitter has open API that helps the developers to build different tools and applications as per the requirement. The initial process of extracting the tweets with the authentication processes is specified in this paper clearly. This helps the programmers and the researchers to use this to and come up with new ideas.

The two API's that the twitter provides: REST API and STREAM API. [13] REST i.e. *Representational State Transfer* enables developers to access information and resources using a simple HTTP invocation. This API helps in obtaining domain-specific data simply by pointing a URL to a specific location. Using the Twitter REST API, you can automate just about everything you can do with Twitter manually. The RESTful API is useful for getting things like lists of followers and those who follow a particular user, and is what most Twitter clients are built off of. The Streaming APIs provide push deliveries of Tweets and other events, for real-time or low-latency applications.

The rich source/data present in media motivates to come up with idea that helps in predicting the future. [14] Few challenges thrown by tweet sentiment classification include:
Sentence level rather than document level, Short and incomplete sentences, Ambiguous and Abbreviated words, Informal and unedited texts, special cases lik Sony is soo classicccc! Its Clarity- superrrrrr☺!".

Our proposed system reveals the process of Extraction of tweets & influential users of the products. The paper emphasis on analysis of tweets that helps in predicting the future. This also helps retailers to know the targeted users. When we come across the reviews and opinions of people, targeted users can be found easily. This benefits the commodity to promote and advance their brand. The layout of the proposed model reveals the ordered steps to meet the objective of the problem definition.

III. Methodology

The process consists of mainly four phases:

1. Extraction
2. Clustering
3. Filtering
4. Polarity Mining

1. Extraction Process

The first phase Extraction cannot be done without having a proper authentication. So the sub process Authentication of an application is necessary to be dealt before the extraction.

a) Authentication process

Application needs to be registered to get consumer key and secret key which are used to authenticate itself (application) to the twitter.

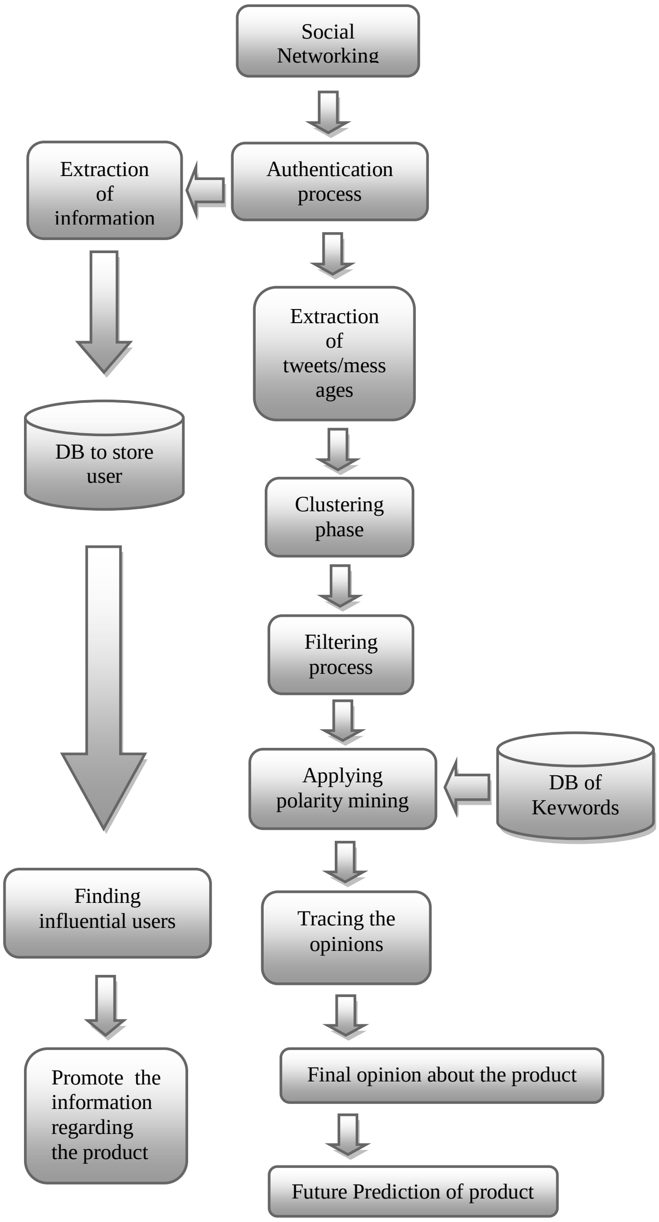

Social Networking
Authentication process
Extraction of information
DB to store user
Extraction of tweets/messages
Clustering phase
Filtering process
Applying polarity mining
DB of Keywords
Finding influential users
Tracing the opinions
Promote the information regarding the product
Final opinion about the product
Future Prediction of product

Figure 2. Architecture of proposed system

The process verifies the application authentication by issuing a token access. Using the token access we can get Access Secret (iPin) which in turn helps the application to authenticate itself to twitter and issues API calls on behalf of user. Responses from the twitter are in JavaScript Object Notation (JSON) format. Use parsing technique/method to convert a JSON format data to string.

2. Clustering Phase

All tweets may not be useful for our analysis. Then pick only those tweets which are related to a particular topic like product name, player name etc. For this process we need to use clustering technique whose task is to group a set of objects in such a way that, objects in the same group/cluster are more similar in some sense or another to each other. Thus Clustering techniques like k-means, Bayes Naïve or other may be used to categorize the tweets based on topic name and restrict the number of tweets.

3. Filtering Phase

Once we get related tweets break them into tokens also called tokenization process. Then remove all the stop words in the set of tokens. And stemming words in the set of tokens are taken care by polarity mining technique. Knowing the number of occurrences i.e. Word Occurrences and calculating total number of valid tokens help us in carrying out the normalization process.

4. Polarity Mining

- Computational study of opinions, sentiments, appraisal, and emotions expressed in text like Reviews, blogs, discussions, microblogs, social networks are known as Opinion Mining Lexicon Based Sentiment Classification: Lexical classes, defined in terms of shared meaning components and similar syntactic behavior of words [15], have attracted a great deal of interest in NLP. A lexical class, or a lexical category) is a linguistic category of words or POS (parts of speech tag), which is generally defined by syntactic or morphological behavior of lexical item in question.
- After extracting the relevant instances (instance may be word, sentence, or tweet), we identify the key words present in them and match them against available sources of positive or negative sentiment words/terms. We collected list of sentiment words from different website or internet sources like http://sentiwordnet.isti.cnr.it/ to create a database of sentiment carrying words. An instance is classified as positive if the count of positive words is greater than the count of negative words. Similarly, an instance is negative if the count of negative words is greater than the count of positive words. Based on the polarity of the word count we can identify the behavior / status of the given topic.
- Thus System analyzes the tweets collected by splitting them into words. Use the dominant polarity of the opinion words in the sentence to determine its polarity. If positive / negative opinion prevails, the opinion of sentence is regarded as positive / negative. Otherwise a threshold is set to have barrier between negative and positive to regard as neutral.
- Summarisation process does Granularity level analysis to get final summary about the product by picking a tweet from set of tweets collected.
- The influentiality of the user is also known by tracking the user's favourites, analysis of his comments and his/her tweets. Fig. 2 shows the architecture of proposed system in brief.

IV. Expected Output

Here we have taken few Sample tweets from my twitter account, to show how the methodology works in a simple way (Fig.3).

Step 1: Initially the tweets are extracted particular to the topic. When we extract the data from the SNS the data will be in the JSON format (Java Script Object Notation). The data should be converted into the format required i.e. string. In Fig.3 the topic Modi was taken and based on the limit given the tweets were extracted. Here we have taken only four tweets as sample from my account.

@NarendraModi-News of a volcanic eruption at Mount Ontake in Japan is quite saddening☹. My prayers with the affected.
@NarendraModi-through their hardwork, actions & strong values, the Indian American community has earned immense respect.We are very proud of them.
@NarendraModi- The programme at Madison Square Garden was overwhelming. It was very special to interact with members of the diaspora.A big Thank you☺
@NarendraModi-There is no reason to be disappointed. India will progress very fast and the skills of our youth will take India ahead.

Figure 3. Sample Tweets Extracted from Twitter Account

Step.2. The formatted data is further proceeded to analyze opinions. In this step the statements are split up into words (Word Extraction) and references, stop words, Uri's are removed. Stop words include is, an, am, a, the, of etc. Below sample stop words are shown [16].

is, of, too, able, about, above, according, after, afterwards, again, against, ain't,.. be, became, because, behind, being, brief, but, by,.. c'mon, c's, came, can, can't,… each, edu, eg, et, etc, even, ever,.. from, further, furthermore,..had, hadn't, happens, hardly, has, hasn't, have, haven't, having, he, he's, hello, help,..

Step.3. Calculation of probabilities of word: The word is checked whether it fall in the range of positive (above +0.5), negative (below -0.5) or neutral (between +0.5 and -0.5). Tokens are matched w.r.t database (Object Structure) and assigned the weights. Even the emoticons are matched. Figure shows the sample database of few words and emoticons and the weights assigned.

Table I. Samples of Emoticons and weights

☺	Happy	Positive
☹	Sad	Negative
:@	Angry	Negative
:-S	Confused	Neutral

Wonderful	0.6
Happy	0.5
Sad	-0.6
Unhappy	-0.5
Awesome	0.8

Handling negation is important for sentiment analysis, as negation words can switch the polarity of a sentiment expression. In the sample tweets, the saddening word is present but in our database the root word sad is present that has negative weight. To resolve these kind of issues, the sad* regular expression can be used. The Pattern Matching Technique helps to solve these problems. If the tweet contains emoticons it will be resolved if the particular emoticon is present in database.

Flow is as follows: After the word extraction is done from tweets, each object from statement is checked in database (object structure). If found, token is matched and is assigned weight. The process repeats for all the tweets or the statements. Thus this helps in knowing the weightage of opinion. This process works if there are max numbers of tweets extracted. As we know Data Mining is achieved if the huge data is mined.

Step.4. The sentences/ tweets are analyzed using clustering technique i.e. bigrams to get better results and to know the accurate meaning between two words. Many recent methodologies in Information Retrieval and Text Mining have used bigram along with unigram since bigram gives more information gain than unigrams [17].Bigram Comparison Function is given by

$$\text{BCF}=\frac{\sum \text{N}}{\text{Average}}$$

Where,
N- Total no. of bigrams that is common between strings
Average- Average no. of bigrams in the strings
Note: Bigram table for English will be different from other, example for German.

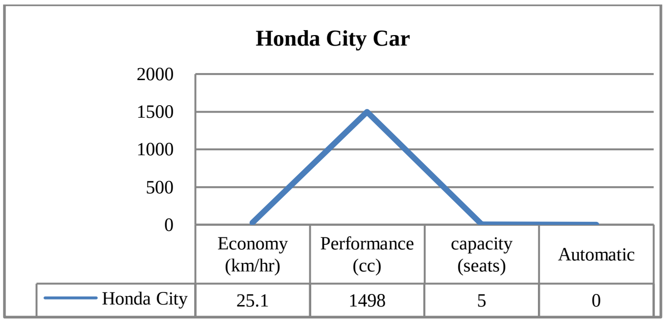


Figure 4. Opinion Graph of a sample tweet

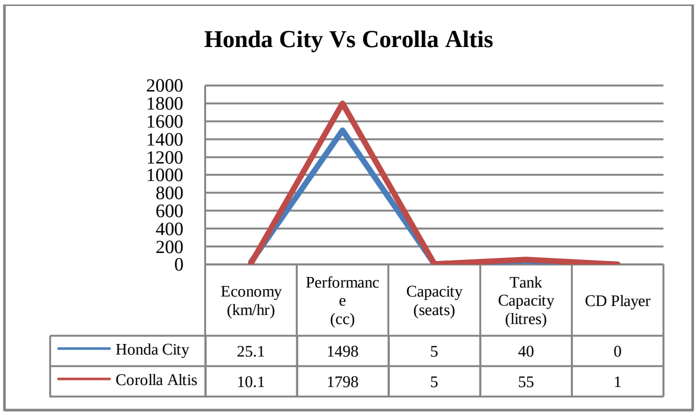


Figure 5. Summary report.

Bigram tables can also be constructed for subsets of languages, such as words relating to a specific subject. Total Weight is evaluated at last. It is the sum of all weights assigned to each sentence based on polarity mining and bigram technique. Calculation of final weight of all the tweets of the product selected will give the final opinion of that product. This helps us to predict the future of that product by collecting different numbers of tweets at different instant of time as time changes opinion of people about the product changes.

Step.5. Summarization Process: The summary of the process can be known as follows: Extraction of the related tweets from twitter based on the topic or product name given.
Eg: "Mileage of Honda City car is better but the option of CD-Player is not present in car"

In the above sample tweet, the user used words Mileage, CD-player which is the most related keywords of Car object. If the tweets of particular topic like car are extracted we will get few tweets specifying few more keywords of Car object. Fig.4 shows a sample graph of a tweet related to Honda city car.

When similar kind of opinions on different car is found the comparison can be done as shown in Fig.5. And find out the positiveness of particular car over the other. This helps user to find which is better in which entity (Economy, Performance).When a person is in ambiguity to decide which car to buy with his requirements this particular analysis will help lot.

Step.6. Finding the influentiality user: The process includes extraction of data from warehouse and applying clustering technique. Further apply Filtering process i.e. removal of stop words. Later evaluating the average of extracted follows, Tweets, Comments, Favorites of particular user. Fig.6 shows the flow diagram of finding targeted users.

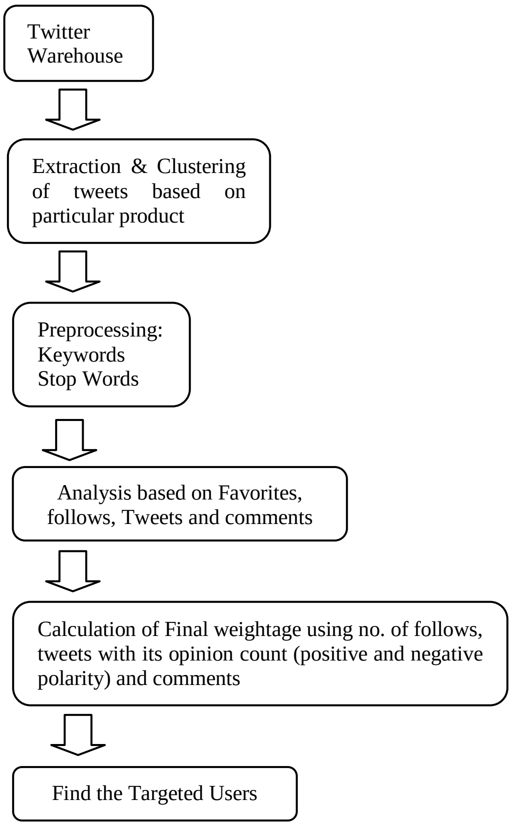

Twitter Warehouse
Extraction & Clustering of tweets based on particular product
Preprocessing: Keywords Stop Words
Analysis based on Favorites, follows, Tweets and comments
Calculation of Final weightage using no. of follows, tweets with its opinion count (positive and negative polarity) and comments
Find the Targeted Users

Figure 6. Flow of Influentiality process

V. CONCLUSION

SNS have become the means of communication or forums to interact and know the feedbacks of the brand. Public's attitude towards the company can be altered by using the social networking sites. Twitter being a popular blogging place gives us a platform to raise the business artistry. Opinion Analysis is a hot area in which data mining tools and techniques can be employed to provide summary information i.e. opinions by extracting words, emoticons, hash tags and phrases from the tweets.

The study gives a brief survey and provides a framework to predict the user's sentiment on products. In this paper, we have used bigram and clustering technique to cluster based on specific topic. The accuracy can be improved by considering specific terms into account and using higher n-gram techniques and Mining techniques.

VI. REFERENCES

[1] Su Mon Kywe, Ee-Peng Lim and Feida Zhu, "A Survey of Recommender Systems in Twitter" supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office,2011.

[2] Ilia Chetviorkin and Natalia Loukachevitch ,"Two-Step Model for Sentiment Lexicon Extraction from Twitter Streams", Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Baltimore, Maryland,USA. Association for Computational Linguistics, June 27, 2014, pp. 67–72.

[3] Robert Romus, "ASVUniOfLeipzig: Sentiment Analysis in Twitter using Data-driven Machine Learning Techniques", Second Joint Conference on Lexical and Computational Semantics, Seventh International Workshop on Semantic Evaluation, Atlanta, Georgia, Association for Computational Linguistics, SemEval 2013, vol 2 ,June 14-15, 2013,pp. 450–454.

[4] Prabu Palanisamy, Vineet Yadav and Harsha Elchuri , "Serendio: Simple and Practical lexicon based approach to Sentiment Analysis", SemEval 2013, June 14-15, 2013, pp. 543-548.

[5] A. Tamilselvi and ParveenTaj, "Sentiment Analysis of Microblogs using Opinion Mining Classification Algorithm",International Journal of Science and Research, ISSN:2319-7064,vol 2 issue 10, October 2013, pp. 196-202.

[6] Ankit Bhakkad , S. C. Dharamadhikari and Parag Kulkarni, "Efficient Approach to find Bigram Frequency in Text Document using E-VSM", International Journal of Computer Applications, April 2013,ISSN:0975-8887, vol 68-No.19,pp. 9-11.

[7] Michael Speriosu, Nikita Sudan , Sid Upadhyay and Jason Baldridge ,"Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph", Proceedings of EMNLP,Conference on Empirical Methods in Natural Language Processing,2011, pp. 53-63.

[8] Efthymios Kouloumpis , TheresaWilson and Johanna Moore , "Twitter Sentiment Analysis:The Good the Bad and the OMG!", Proc. of the Fifth International AAAI Conference on Weblogs and Social Media,2011.

[9] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil and Meichun Hsu, Bing Liu "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", Hewlett-Packard Development Company, L.P, June 2011.

[10] Rudy Prabowo1 and Mike Thelwall, "Sentiment Analysis: A Combined Approach", Journal of Informetrics,Elseveir, vol 3,issue 2, 2009, pp.143-157.

[11] Bo Pang and Lillian Lee, "Opinion mining and sentiment analysis",Foundations and Trends in Information Retrieval, vol 2, No 1-2 ,2008, pp.1–135.

[12] Soo-Min Kim and Eduard Hovy, "Determining the Sentiment of Opinions" Proceedings of the COLING conference, Geneva, 2004.

[13] http://www.ibm.com/developerworks/library/x-twitter REST/

[14] Furu Wei " Sentiment Analysis and Opinion Mining", Natural Language Computing Group, Microsoft Research Asia fuwei@microsoft.com.

[15] Levin, B. "English verb classes and alternations: a preliminary investigation",Chicago,IL:University of Chicago Press, 1993.

[16] Karuna C. Gull, Akshata B. Angadi, Seema C. G and Suvarna Kanakaraddi "A Clustering Technique To Rise Up The Marketing Tactics By Looking Out The Key Users", Conference, Advance Computing Conference (IACC), IEEE International, 21-22 Feb. 2014,pp. 579-585.

[17] Ankit Bhakkad, S. C. Dharamadhikari and Parag Kulkarni, PhD "Efficient Approach to find Bigram Frequency in Text Document using E-VSM", International Journal of Computer Applications, ISSN:0975-8887,vol 68- No.19, April 2013.