Global topics Sentiment Analysis by using Social Media data mining

Vaishali J. Shimpi
M.E. Student, Department of Computer Engineering
Dr. D.Y. Patil School Of Engineering & Technology,
Savitribai Phule Pune University, Pune, India
vaishalikapure19@gmail.com

Roshani Ade
Assistant Professor, Department of Computer Engineering
Dr. D.Y. Patil School Of Engineering & Technology,
Savitribai Phule Pune University, Pune, India
Rosh513@gmail.com

***Abstract*— In recent years, there has been an emerging interest which shows and support social media analysis for marketing, opinion analysis and understanding community coherence. Understanding gained from this social media analysis is used by many organizations like marketing, advertising, disaster management, etc. to improve their performance by planning future strategies. This paper proposes a workflow to integrate both qualitative analysis and large-scale data mining techniques, which by using parallelism and multi-threading technique provide the fastest way to data analysis.**

Keywords— Real time data; Data analyzer; Social media;

I. INTRODUCTION

Social media data mining is emerging as a very powerful tool in todays subjective analysis world. Its increased use of social media sites such as Twitter, Facebook and private sites provides great venues for many organizations or social system to share their experiences, vent emotion and requirements. On various social media sites, many discuss and share their everyday encounters in an informal and casual manner. Digital footprints provide vast amount of implicit knowledge and a whole new perspective for researchers and practitioners understand experiences. This understanding can help decision-making on interventions for product organization, product quality, and thus enhance better decisions and endorsement regarding the application, Product, market, situation, policies. The abundance of social media data provides opportunities to understand experiences, but also raises methodological difficulties in making sense of social media data for analysis purposes.

The challenges and difficulties in methodology present opportunities for data miners to develop new methods and algorithms for social media data mining. Conditional attribute used for classic data mining cannot be implemented directly on social media data mining. Social media data are huge, noisy, unstructured, distributed, and dynamic. These characteristics pose challenges to data mining tasks to invent new, efficient methodology and algorithms.

Depending on social media platforms, noise of social media data can often be vary. Elimination of the noise from the data is required before performing effective mining. Researchers found that spammers generate more data than true users[1,2]. There is no central control authority to maintain data for all social media platforms. A distributed data pose a challenging task for researchers to understand the information flows on the social media. As data on social media data is always unstructured. To extract meaningful observations on unstructured data from various data sources is a big challenge. For example, social media sites like Twitter, Facebook, LinkedIn and Flickr serve different purposes and meet different needs of users..

Social media platforms are continually evolving and dynamic. For example, recently Facebook introduced many features, including timeline for each user, in-groups creation of a user, and numerous user policy of privacy changes. The dynamic nature of social media data is one of the big challenges for evolving social media sites. Social media mining can help to find answers to many interesting questions related to human behavior. Advertisers can find the influential people to increase the reach of their products within a targeted budget.

Social media data analysis can be of different type based on here targeted outcome and data. One of the major part of the social media data analysis is Sentiment analysis and opinion mining. In this type of analysis or mining aim is to extract opinions expressed in the user-generated content automatically. Understanding is gained from by Sentiment analysis and opinion mining tools assist businesses to understand brand perception, new product perception, product sentiments and reputation Management. These tools help to understand product opinions or sentiments on a global scale for product organization. Each social media site report user opinions in different format. So, segregating and analyzing opinions related to a particular organization or product on social media sites which is one more new challenge.

Ambiguous content in user language for creating opinions produces new challenges to sentiment analysis. Below are the steps of usually used sentiment analysis are (1) Searching relevant documents, (2) Searching relevant sections, (3) Searching the overall sentiment, (4) quantifying the sentiment, and (5) Creating overview from aggregating all sentiments. Basic components of an opinion are (1) an object on which opinion is expressed, (2) an opinion expressed in an object, and (3) the opinion holder. Generally, objects are represented by a finite set of attributes, a finite set of phrases and synonyms for each attribute is represented. Opinion mining can be implemented at the document level, sentence level, or attribute level [3,4,5,6,7,8,9]. Segregation of opinions expressed in comparative sentences is a challenge. Performance evaluation of sentiment analysis is another challenge.

The rest of this paper is organized as follows. Section 2 gives brief idea about related work and problem analysis. Section 3 includes proposed work with the system architecture. Section 4 deals with result comparison. Section 5 includes the conclusion of overall work with future scope.

II. Related Work

There has been lots of work done on sentiment analysis, especially in the area of product, book and movie reviews and blogs. Sentiment analysis determines whether each opinion is expressed in a positive, negative, or neutral way. This is also related to subjective/objective polarity [10,11]. Sentiment classification using machine learning is a well-studied field [12]. A method that automatically distinguishes each word as either positive, neutral, or negative using semi-supervised learning found by Esuli and Sebastiani [11]. Pang and Lee applied various machine learning method to sentiment classification problems[11,4].

Micro-blogging has nowadays become one of the major types of the communication. A recent research has identified it as online word-of mouth branding [13]. The large amount of information contained in a micro-blogging website makes them an attractive source of data for opinion mining and sentiment analysis.

In the case of the Mumbai terrorist attack, performed a qualitative analysis to argue that the terrorists monitored tweets posted by networked citizens to their advantage as they utilized situational information to mount attacks against civilians [16].

Using tweets extracted from Twitter during the Australian 2010-2011 floods, social network analysis techniques were used to generate and analyze the online networks that emerged at that time. The aim was to develop an understanding of the online communities for the Queensland, New South Wales and Victorian floods in order to identify active players and their effectiveness in disseminating critical information[15].

Different algorithm and classification technique is used for finding exact meaning and draw conclusion from large data. Naive Bayes, Support vector machine, Max margin algorithm sprovide a classification of the data [14,15].

Social media data is growing continuously author describe it a set of event, it is possible to classify Social media data by the events using using scoring and ranking [18]. One of the algorithm with support for multi-label. When Multiple label is used in the system, it reduces performance, author describe such algorithm which improves the speed [19].

III. Propose Work

In this paper a system workflow for sentiment analysis of social media data is proposed as shown figure-1 which is composed of three major stages 1) Data collection, 2) Attribute extraction and 3) Data analysis. And new algorithm based on string base kernel (SSK) and the support vector machine (SVM) algorithm.

A. Data Collection

This is the first step in the database for analysis is collected from social media. There aremany techniques used to collect data such as NodeXL, Radian6, Twitter API and Facebook API [21,22,23]. In proposing work API for different platform is used for data collection. In this work data will be always updated with new data available which make it real-timedata analysis system.

B. Attribute Extraction

In this step based on subject knowledge keyword or attribute are generated which used to classify database collected. There can be automatic or manual technique used to do this. In proposing work attributes extraction is done manually and addictive set or each attribute set will be generated by using google data base.

C. Data Analysis

Data analysis is carried out with different classifier such as SVM, M3L, Nave Bayes, etc. [22]. The result is populated in the form of visual graph or reports. Through this paper proposes a new modified algorithm which is based on string base kernel (SSK) and the support vector machine (SVM) algorithm. Using parallelism and multi-threading technique this algorithm able to provide faster categorization and better accuracy.

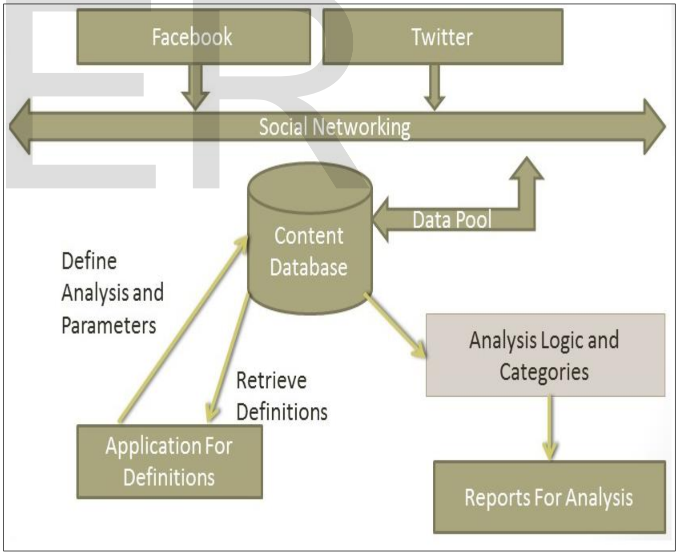


Fig. 1. Fig 1: System workflow

IV. Algorithm

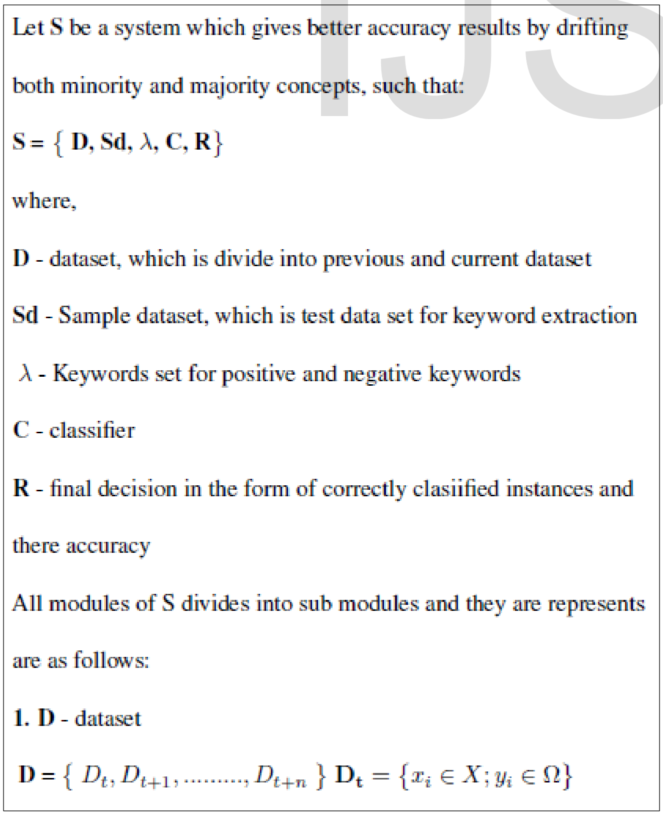
Using parallelism and multi-threading technique this algorithm able to provide faster categorization and better accuracy. This algorithm work with parallelism and multi-threading, which provide benefits over other algorithm like Naïve Bayes multi-leble algorithm in which accuracy is more but speed is less due to sequential execution.

```
Input:  Ts= {"Yes it was good"}, λ()=  {"good, bad, sweet, spicy, like, not liked, more"}

Categories Identifier P: Positive, N: Negative, X: Neutral

Step 1:    λ="good"
           If Ts= λ then set P go to step 4
           Else   for each item in TS as Item <> λ  go to
           step 2

Step 2:  Compare with all words
             i.e yes{it,was,good} or  it{yes,was,good} or
             was{yes,it,good}…..
          Now calculate every High value with Low value.
          This will return new value {p} =P|N and go to
          step 3 else set X and go to step 4

Step 3: complete set will be compare with word adjective
            For e.g good { good, goodest,..}  if  Old > New
                         then mark P|N

Step 4 :  Mark complete line as P|N|X based on last
           Probability And   λ="bad"
```

Fig. 2. Algorithm System workflow

V. Mathematical Model

In fig 3 mathmatical model for above mention system is briefly explained.



Let S be a system which gives better accuracy results by drifting both minority and majority concepts, such that:

S = { D, Sd, λ, C, R}

where,

D - dataset, which is divide into previous and current dataset

Sd - Sample dataset, which is test data set for keyword extraction

λ - Keywords set for positive and negative keywords

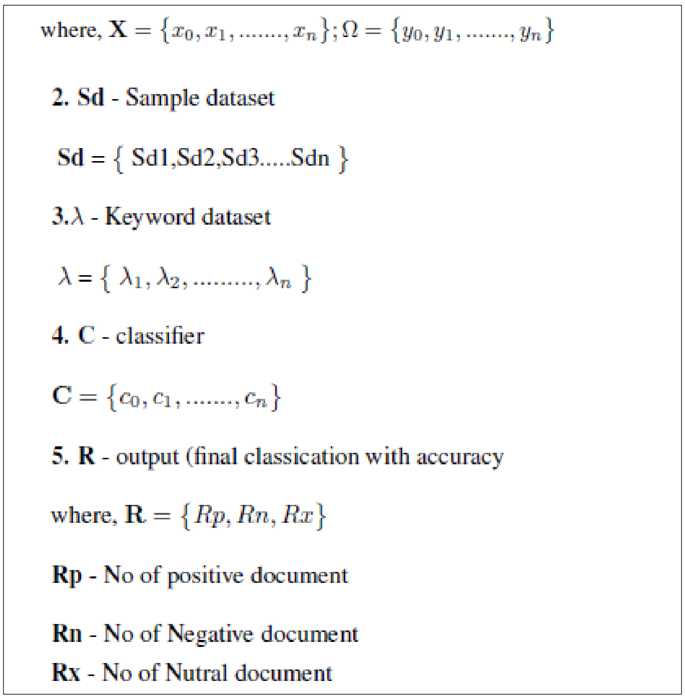
C - classifier

R - final decision in the form of correctly clasiified instances and there accuracy

All modules of S divides into sub modules and they are represents are as follows:

1. D - dataset

D = { $D_t, D_{t+1}, ..., D_{t+n}$ } $\mathbf{D_t} = \{x_i \in X; y_i \in \Omega\}$



where, $\mathbf{X} = \{x_0, x_1, ..., x_n\}; \Omega = \{y_0, y_1, ..., y_n\}$

2. Sd - Sample dataset

Sd = { Sd1,Sd2,Sd3.....Sdn }

3.λ - Keyword dataset

λ = { $\lambda_1, \lambda_2, ..., \lambda_n$ }

4. C - classifier

$\mathbf{C} = \{c_0, c_1, ..., c_n\}$

5. R - output (final classication with accuracy

where, $\mathbf{R} = \{Rp, Rn, Rx\}$

Rp - No of positive document

Rn - No of Negative document

Rx - No of Nutral document

Fig. 3. Mathematical model of system

VI. Experimental Results

In this section we explain the experimental setup the dataset used and the approach used for this research. As per methodology discussed, we collected Reviews of students Learning Experience from Twitter & College social sites. We have created dataset of 1000 comments/ tweets and categorized it based on 6 different keywords based on analysis of the collected data. Performed analysis is based on Text based SVM method with our modifications and previously proposed Nave Bayes algorithm.

As per the tweets collected to the system, we created analysis for the same on accuracy. We have defined accuracy based on below measures.

Accuracy (%) =   ("No. Of Correct Classification" / "No. Of the samples in dataset") * 100

Also, Precision and Recall were used as the metrics for evaluating the performance of each text classification approach. There are two measures considered in the results
1) Performance based on dataset size,
2) Performance based on keywords sized,
3) Speed for classification.

A. Effect of dataset on performance

An experiment measuring the performance against the size of dataset with 6 keywords was conducted using dataset of different sizes listed in Fig.1. The experiment was performed with seven different categories. For example, in case of 1000 data sets, Accuracy was 95.80% using NB classifier.

TABLE I. COMPARISON IN TECHNIQUES (THE ABOVE MENTIONED ACCURACY BASED ON DATASET PERFORMANCE MEASURES)

Dataset Size	Naïve Bayesian	Text Based SVM	Text Based SVM - Improved
1000	98.20%	92.70%	97.00%
2000	98.15%	95.50%	98.20%
3000	97.83%	92.40%	97.27%
4000	97.75%	93.40%	97.63%
4500	96.47%	91.10%	97.56%

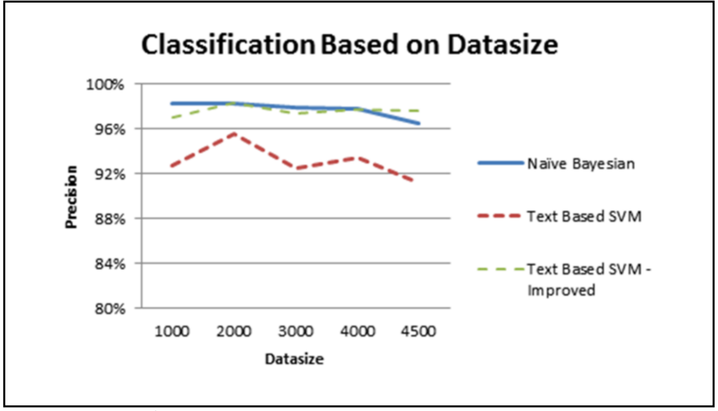


Fig. 4. Classifcation Based on Datasize

A few observations can be made from this experiment. As shown in Fig. 1, the average of correct classification rate for both Improved SVM and NB was over 95%. Dataset size was not an important factor in measuring precision and recall. The results show that the performance of classification was not 100% stable. The above results may vary based on 'Keywords- Sized of features to be extracted'

2) *Effect of Keywords size on performance:*

The other experiment measuring the performance against the size of dataset was conducted using different keywords listed in Fig. 2. Let say size of 1000 dataset was used for the experiment.

TABLE II. COMPARISON IN TECHNIQUES (THE ABOVE MENTIONED ACCURACY BASED ON KEYWORDS PERFORMANCE MEASURES)

Keywords Size	Naïve Bayesian	Text Based SVM	Text Based SVM - Improved
7	94.42%	91.70%	95.00%
20	95.60%	85.73%	96.50%
30	95.64%	88.87%	96.27%
40	97.13%	89.93%	97.00%
50	97.69%	92.10%	97.69%

As shown in Fig.5 good classification result order in the experiment were NB, Improved SVM and SVM for all cases of classification. The overall precision and recall for classification increase and become stable, according to the increase of the number of Keywords/feature. Gradually, the accuracy increase and finally saturated with the increased keywords/feature size.

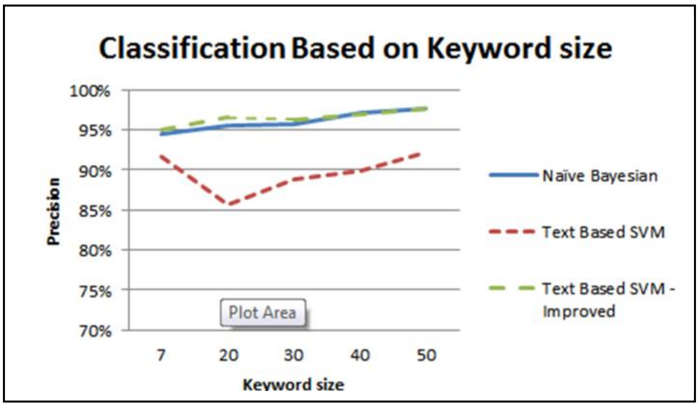


Fig. 5. Classification Based on Keyword

As per above to results Naive Bayesian & Improved Text based SVM results almost same. But, considering our third components speed measures between these two techniques.

3) *Speed for classification*

The third experiment was based on Speed of classification based on above performance measures. Let's consider size of 1000 dataset with size on 7 keywords used under the classification. Based on an experiment we found the following results.

TABLE III. APPROX. TIMING FOR CLASSIFICATION ACCURACY BASED ON SPEED OF CLASSIFICATION

Naïve Bayesian	Text Based SVM	Text Based SVM-Improved
94.42% - 3.5Mins	91.70% - 2.4Mins	95.00% - 1.5Mins

As shown in Table III. The Naïve Bayesian technique was taken large time because it is leaning process which takes the time. Where Text Based SVM method is resulting in less time, but accuracy was less than others. Improved algorithm achieved accuracy as well as speed because it supports parallelism which improves speed unit for classification.

Thus, improved algorithm will cover all aspects required for classifying and retrieval process.

CONCLUSION

The paper presents a system that is able to classify and show data from social media on the basis of analysis. Applied method to a data set derived from twitter, Facebook using keyword classification, though the third party API.

Proposed methodology addresses successfully two key problems: i) manual qualitative analysis, and ii) large scale computational analysis of user-generated textual content, i.e. the problem of determining whether an incoming data item belongs to the categories. These problems addressed by i) searching through naive Bayes algorithm and text based SVM method and ii) by using multi-threading in .net will be achieving goal for fast scanning data. This way this paper provide new approach to data classification which one of accurate and faster way.

REFERENCES

[1] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd. Detecting spam in a Twitter network. First Monday 15(1),1-4, 2009.

[2] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on Twitter: Human, bot,or cyborg?, Proceedings of the 26th Annual Computer Security Applications Conference, Association for Computing Machinery, New York, 21-30, 2010

[3] P. D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classfication of reviews., Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, PA, 417-424, 2002..

[4] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques., Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Vol. 10. Association for Computational Linguistics, Stroudsburg, PA,79-86, 2002

[5] E. Rilo and J. Wiebe. Learning extraction patterns for subjective expressions., Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA, 105-112, 2003..

[6] H. Yu and V. Hatzivassiloglou, Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences., Proceedings of the 2003, Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA, 129-136, 2003

[7] M. Hu and B. Liu, Mining and summarizing customer reviews. Proceedings of the Tenth ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, 168-177, 2004.

[8] A. M. Popescu and O. Etzioni, Extracting product features and opinions from reviews. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA, 339-346,2005.

[9] H. Liu and P. Maes, InterestMap: Harvesting social network proles for recommendations., Workshop: Beyond Personalization, San Diego, 2005

[10] B. Pang and L. Lee, Opinion Mining and Sentiment Analysis, in Foundations and Trends in Information Retrieval, 2 (1-2), pp. 1135, 2008.

[11] A. Esuli and F. Sebastiani, SE ENTIWORDNET: A Publicly Available Lexical Resource for Opini ion Mining, Proceedings of the 5th Conference on Language Resources and Evaluation, Genoa, 24-26 May 2006.

[12] C. D. Manning and H. Schu utze, Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[13] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. Micro-blogging as online word of mouth branding. In CHI EA 09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems, pages 38593864, New York, NY, USA.ACM. 2009.

[14] Oh, O., Agrawal, M., & Rao, H. R., Information control and terrorism: tracking the Mumbai terrorist attack hrough Twitter., Information Systems Frontiers, 1-11. (2010).

[15] Cheong, France and Cheong, Christopher, "Social Media Data Mining: A Social Network Analysis Of Tweets During The 2010-2011Australian Floods" . PACIS 2011 Proceedings. Paper 46, 2011

[16] Songtao Zheng, Nave Bayes Classifier: A MapReduce Approach, Fargo, North Dakota, October 2014

[17] Huma Lodhi,Craig Saunders,John Shawe-Taylor,Nello Cristianini,Chris Watkins: Text Classification using String Kernels Journal of Machine Learning Research 2, 2002

[18] Reuter, Timo and Philipp Cimiano. Event-based Classification of Social Media Streams. Proceedings of the 2nd ACM International Conference Multimedia Retrieval, article 22.23, 2012.

[19] Muazzam-Khan, Fareed-ud-din, Data Classification and Text Mining in CRIMS in the Perspective of Pakistan, Proceedings of the International Conference on Engineering and Emerging Technologies, 2014.

[20] Luciano Barbosa and Junlan Feng, Robust sentiment detection on twitter from biased and noisy data., COLING'10 Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 3644, 2010.

[21] Alemu Molla, Yenewondim Biadgie and Kyung-Ah Sohn, Network-based Visualization of Opinion Mining and Sentiment Analysis on Twitter the National Research Foundation of Korea, pp. 443-749,2012.

[22] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan, Mining Social Media Data for Understanding Students Learning Experiences, IEEE transaction on learning technology, Vol 7,No. 3, July-Sept 2014.

[23] Mohit Tare1, Indrajit Gohokar2, Jayant Sable3, Devendra Paratwar4, RakhiWajgi5, Multi-Class Tweet Categorization Using Map Reduce Paradigm, International Journal of Computer Trends and Technology (IJCTT) , vol 9, number 2, ISSN: 2231-2803, Mar 2014.