International Journal of Research
ISSN: 2348-6848 Vol-3, Special Issue-6
National Conference on Recent Engineering Trends
NCRET- 2016
Held on 16th March 2016 organized by Institute of Knowledge College of Engineering, Pune, India.

# Opinion Mining and Sentiment Analysis on Twitter

**Tathe Varsha Bhimashankar**
(*BE-Dept. of CSE*) SP"s institute of knowledge college of engineering
Email Id-tathem6@gmail.com
**Shelke Swapnil Shantaram**
(*BE-Dept. of CSE*) SP"s institute of knowledge college of engineering
Email Id-swapnilshelkedada@gmail.com
**Lonari Rutuja Govind**
(*BE-Dept. of CSE*) SP"s institute of knowledge college of engineering
Email Id-rutujalonari11@gmail.com
**S.Pratap Singh**
(M.Tech) SP"s institute of knowledge college of engineering
Email Id-pratap.singh.s@gmail.com

**Abstract-** *Twitter platform is valuable to follow the public sentiments. Knowing users point of views and reasons behind them at various point is an important study to take certain decisions. Categorization of positive and negative opinions is a process of sentiment analysis. It is very useful for people to find sentiment about the person, product etc. before they actually make opinion about them. In this paper Latent Dirichlet Allocation (LDA) based models are defined. Where the first model that is Foreground and Background LDA (FB-LDA) can remove background topics and selects foreground topics from tweets and the second model that is Reason Candidate and Background LDA (RCB-LDA) which extract greatest representative tweets which is obtained from FB-LDA as reason candidates for interpretation of public sentiments.*

**Keywords-** Twitter; Public Sentiments; Sentiment analysis; Event tracking; Latent Dirichlet Allocation (LDA); Foreground and Background LDA; Reason Candidate and Background LDA.

1. **INTRODUCTION:** There are number of users who share their views through twitter which are changes rapidly. Sentiment analysis on twitter data helps to expose opinions of peoples. One important analysis is to find possible reasons behind sentiment variation, which can provide important decision making information. It is generally difficult to find the exact reason of sentiment variations. The emerging topics which are discussed in the different changing periods are connected to the some genuine reasons behind the variations. It will be consider these emerging topics as possible reasons. It defines two Latent Dirichlet Allocation (LDA) based models

to analyze tweets in significant variation periods. Foreground and Background LDA (FB-LDA) filter out background topics and selects foreground topics from tweets in the variation period. Another model called Reason Candidate and Background LDA (RCB-LDA) first take outs representative tweets for the foreground topics obtained from FB-LDA as reason candidates and rank the reason candidates. Twitter data helps to analyze and interpret the public sentiment variations in micro blogging services. The two proposed models are general they can also be applied to find topic differences between two or more sets of documents.

## 2. PROPOSED SYSTEM:

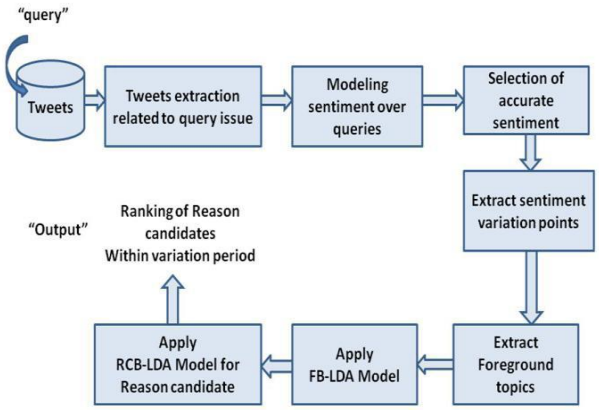


Fig1: Praposed System

### Tweet extraction and Preproccesing:

To extract tweets relatrd to the target ,we go trough the whole database and extract all the tweets which contains the keyword of the target. Compared with regular text document ,tweets are generally less formal. Therefore ,preprocessing techniqueson tweets are necessary for obtaining satisfactory results on sentiment analysis:

1. Slang word translation:
   Tweet often contain a lot of slang word (e.g., omg, hand, bdw,etc). We convert these slang words into theire standard forms and add them to the tweets.
2. Non-english tweets filtering:
   Since the sentiment analysis tools to be used only work for English texts, we remove all non-English tweets.
3. URL removal:
   A lot of users include URL's in theire tweets. These URL's complicate the sentiment analysis process. We decide to remopve them from tweet.

### Sentiment Label Assignment :

It assigns the sentiment labels by using two state-of-the-art sentiment analysis tools. SentiStrength tool which first assign a sentiment score to each word in the text, then choose the maximum positive score and the maximum negative score among those of all individual words in the text, and compute the sum of them to denote Final Score. Then by using the sign of Final Score it will indicate that a tweet is positive, neutral or negative.

SentiStrength tool calculate the probability of each tweet, if percentage of tweet is more than 50% then according to the percentage it will assign the positive ,negative or neural

### Sentiment Variation Tracking

After assigning the labels of sentiments from all extracted tweets, it will track the sentiment variation using some descriptive statistics. In this

International Journal of Research
ISSN: 2348-6848 Vol-3, Special Issue-6
National Conference on Recent Engineering Trends
NCRET- 2016
Held on 16th March 2016 organized by Institute of Knowledge College of Engineering, Pune, India.

work, it is necessary that analyzing the time period during which the overall positive (negative) sentiment climbs upward while the overall negative (positive) sentiment slides downward.

## 3. SYSTEM REQUIREMENT

**a. Software requirements**

1. Java Development Kit (JDK)7
2. Windows 7/8 (32-bit or 64-bit) operating system
3. Apache Tomcat
4. Eclipse
5. MySQL 6.0

**b. Hardware requirements**

1. Hardware : Pentium
2. Speed : 1.1 GHz
3. RAM : 1GB
4. Hard Disk : 20 GB

## 4. IMPLEMENTATION

**Screenshots**

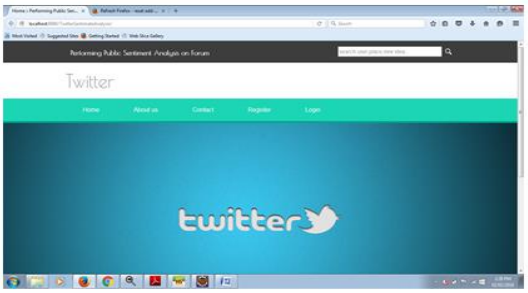


Fig2: Home page

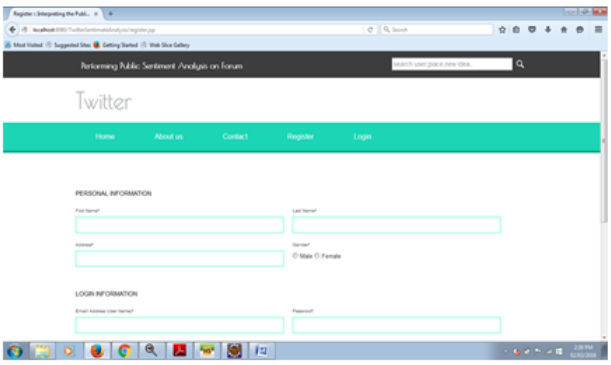


Fig3: Account page

## 5. MODULES FOR SENTIMENT VARIATION ANALYSIS:

To calculate the positive or negative rank of the tweets, we apply the two modules. Modes are listed below:

a. ForeGround and BackGround LDA(FB-LDA)
b. Reason Candidate and BackGround LDA(RCB-LDA)

### a. Fore Ground and Back Ground LDA(FB-LDA)

To obtain foreground topics, we need to filter out all background topics related to the query.
To find the foreground topics we designed the two methods as**: k-mean** and **LDA.**
For k-means, we first run the k-means clustering on the foreground set and the background set respectively. Since clusters from the foreground set contain both foreground and background topics, we design a mechanism to filter out background clusters by comparing clusters

International Journal of Research
ISSN: 2348-6848 Vol-3, Special Issue-6
National Conference on Recent Engineering Trends
NCRET- 2016
Held on 16th March 2016 organized by Institute of Knowledge College of Engineering, Pune, India.

between the foreground set and the background set. If a cluster corresponds to the same topic/event with one background cluster, it will be filtered out. After background topics filtering, the remaining foreground clusters will be ranked by their sizes in descending order. Then for each cluster we find five tweets which are closest the cluster center. The evaluation method for k-means is as same as that of FB-LDA.

For the second baseline LDA, the background topics filtering step is similar to k-means. But instead of comparing cluster centers, here we compare the word distributions of topics. We observed that the most relevant tweets of each topic/cluster are similar with each other and could clearly represent the semantics of the topic/cluster. Moreover, each of the most relevant tweets generally corresponds to a specific event. Therefore, if a representative tweet of foreground topic/cluster appears in the ground truth set, we could reasonably conclude that the foreground topic/cluster corresponds to one ground truth event.

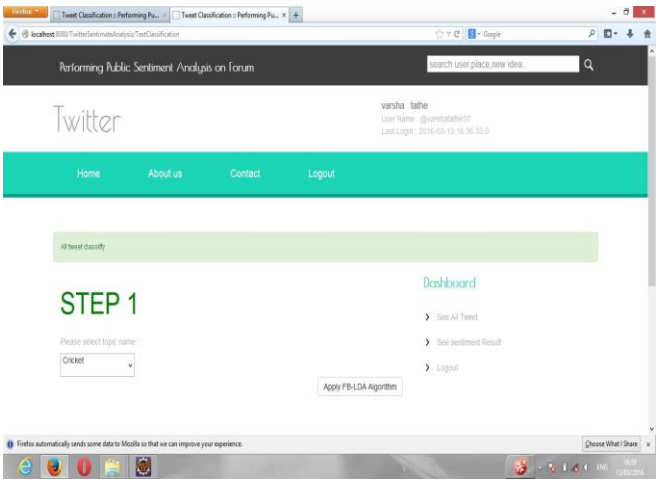


Fig.4:FB-LDA module

**b. Reason Candidate and BackGround LDA(RCB-LDA)**

It will select the most appropriate/representative tweet obtain from FB-LDA to each foreground topic

## 6. NAIVE BAYES:

Naïve Bayes is a simple model which works well on text categorization. We use a multinomial Nsive Bayes model.

$P(C_k|x1,\ldots\ldots,xn)$

$$P(C_k|x) = \frac{P(C_k)\,P(x|C_k)}{P(x)}$$

Where ,

X = represents a features, P(x)= total no of feature database

C= class of related tweets

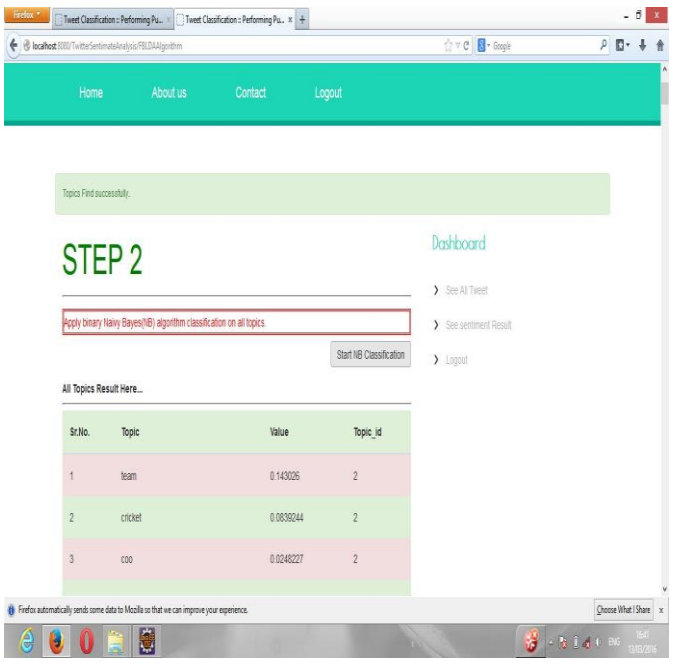


Fig5: Naive Bayes

## 7. EXPERIMENTAL RESULTS

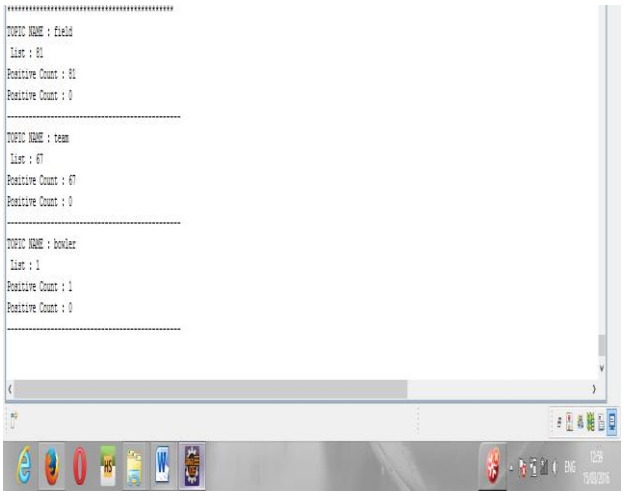


Fig6: RCB-LDA

## 8. CONCLUSION:

In this paper, the problem of analyzing public sentiment variations and finding the possible reasons behind it are solved by using two Latent Dirichlet Allocation (LDA) based models such as Foreground and Background LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA). This system can mine possible reasons behind sentiment variations which provide the sentence level reasons. This are the actual causes for sentiment variations .This system is general so it can also be used to discover special topics or aspects in one text collection comparison with another background text collection.

## 9. REFERENCES

[1] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, "Interpreting the Public Sentiment Variations on Twitter," IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO. 5, MAY 2014.

[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inform. Retrieval, vol. 2, no. (12), pp. 1135, 2008.

[3] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. 10th ACM SIGKDD, Washington, DC, USA, 2004.

[4] W. Zhang, C. Yu, and W. Meng, "Opinion retrieval from blogs," in *Proc. 16th ACM CIKM*, Lisbon, Portugal, 2007.

[5] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in Proc. 15th ACM SIGKDD, Paris, France, 2009.

[6] D. Tao, X. Tang, X. Li, and X. Wu,"Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval," IEEE Trans. Patt. Anal. Mach. Intel., vol. 28, no. 7, pp.10881099, Jul. 2006.

[7] D. Chakrabarti and K. Punera,"Event summarization using tweets," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.

[8] T. Sakaki, M. Okazaki, and Y. Matsuo,"Earthquake shakes twitter users: Real-time event detection by social sensors," in Proc. 19th Int. Conf. WWW, Raleigh, NC, USA, 2010.

[9] V. Hatzivassiloglou and J. M.Wiebe."*Effects of adjective orientation and gradability on sentence subjectivity*".pp. 299-305, 2000.

[10] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. "*Twitter power: Tweets as electronic word of mouth*". J. Am. Soc. Inf. Sci., 60(11):2169-2188,2009.