International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Public Sentiment Interpretation on Social Web: Twitter

Devaki Ingule[1], Gyankamal Chhajed[2]

[1]Student of ME-II, Department of Computer Engineering, VPCOE, Baramati, Savitribai Phule Pune University, Baramati, Maharashtra, India

[2]Assistant Professor, Department of Computer Engineering, VPCOE, Baramati, Savitribai Phule Pune University, Baramati, Maharashtra, India

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 microblogging services. The two proposed models are general they can also be applied to find topic differences between two or more sets of documents.

2. Literature Survey

2.1. Sentiment Analysis

Sentiment analysis is the process of analyzing the opinions which are extracted from different sources like the comments given on forums, reviews about products, various policies and the topics mostly associated with social networking sites and tweets. Pang et al. [2] work on the supervised machine learning methods existing for analyzing sentiments.

Advantages: Different machine learning methods reduce the structural risks.

Disadvantages: This method is not able to analyze possible reasons occurred behind the public sentiments. The Supervised machine learning methods demands for large amount of labeled training data which is expensive. It may not work when training data are insufficient. M.Hu and B. Liu [3] works on novel techniques which are useful to collecting and summarizing customer reviews. It identifies customer opinions and after that decides whether it is categorize into positive or negative.

Advantages: It predicts movie sales as well as elections which help for making decisions. It also provides a feature-based summary for large amount of customer opinions.

Disadvantages: It will not calculate the opinions strength and also not represent opinions with its verbs, adverbs and nouns. W.zhang et al. [4] studied of opinions retrieval from blogs. In this paper, they have presented a three-component opinion retrieval algorithm.
1. Selects information retrieval module.
2. Classifies document into optionative document.
3. Rank document in certain order.

Advantages: It gives higher performance than state-of-art opinion retrieval methods.

Disadvantages: It is not able to handle more general writings and crossing domains. It also not selects detail features.

2.2. Event Detection and Tracking

A number of events are useful reasons behind the variations of sentiments which are actually related to its target. This task is done by tracking related events to target.

Proposed system for the following memes such as quoted phrases or sentences are done by Leskovec et al. [5]. It detects short distinctive phrases. This work offers some analysis for the global news cycle and also offers dynamics of information propagation between social media.

Advantages: It provides temporal relationships such as the possibility of employing a type of two-species predator prey model with blogs and the news media as the two interacting participants.

Disadvantages: It is useful only for most representative events in whole twitter message stream. It detects fine grained events very hardly.

2.3. Data Visualization

D.Tao et al. [6] works on subspace learning algorithms and ranking technique. Now days, retrieving of images from databases is very active research field which are uses content based image retrieval (CBIR) technique. It is highly connected by semantically to the query of user.SVM classifier behaves like unstable for a smaller size of training set. SVMRF also becomes poor if there is number of samples of positive feedback are small.

Advantages: It helps for increasing the performance of relevance feedback. Whenever SVM classifier behaves like unstable on a smaller size of training set. To address this issue an asymmetric bagging-based SVM (AB-SVM) model is developed. For overfitting problem it used both the random subspace method and SVM together.

Disadvantages: It is not able to use tested tuning method. It is not select the parameters of kernel based algorithms. These works are not useful for noisy text data. Explicit queries are not available in task, so the ranking methods will not solve the reason mining task.

2.4. Correlation between Tweets and Events

Chakrabarti and Punera [7] studied use of sophisticated techniques to summarize the relevant tweets are used for some highly structured and recurring events. Hidden Markov Models gives the hidden events.

Advantages: It provides benefits for existing query matching technologies.

Disadvantages: It does not use the continuous time stamps present in tweets. It is not possible to gets minimal set of tweets which are relevant to an event. In this novel model noises and background topics cannot be eliminated.

T.Sakaki et al. [8] proposed novel models to map tweets into the public sentiments. This model is used to detect real-time events on Twitter such as earthquakes. For monitoring tweets they proposed an algorithm. In this twitter user considered as sensor.Kalman filtering and particle filtering algorithms are used for estimation of location.

Advantages: Kalman filtering and particle filtering detects and provides estimation for location who detects the earthquake. Earthquake detection is done by using this system.

Disadvantages: It would not provide advanced algorithms for query expansion. It is difficult to detect multiple event occurrences.

3. Problem Definition

Sentiment analysis on twitter data has provided an effective way to expose public opinion, which is critical for decision making in various domains. The Previous research mainly focused on only modeling and tracking public sentiments. To enhance this proposed system interprets sentiment variations and the possible reasons behind public sentiments. To address this issues a Latent Dirichlet Allocation (LDA) based models such as Foreground and Background LDA (FB-LDA) are proposed. To enhance it further another generative model proposed called Reason Candidate and Background LDA (RCB-LDA) which rank them according to their popularity within the variation period.

4. Methodology

4.1 Tracking Public Sentiments

Tracking Public sentiments includes the following three steps: First, It will take out tweets equivalent to interested targets (e.g., Obama, Apple etc.), then it preprocess that extracted tweets to make them more efficient. Second, by combining two state-of-the-art sentiment analysis tools it will assign a sentiment label to each individual tweet. Finally, based on this sentiment labels, it will track the sentiment variation regarding to the associated target.

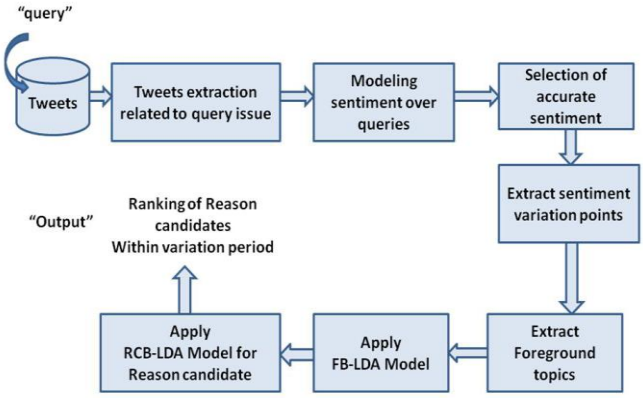


Figure 1: System architecture

4.1.1 Tweets Extraction and Preprocessing

It will take out all tweets associated to the target extracted through Twitter Dataset. It contains the keywords of the target. Sentiment analysis tools are applied on untrained tweets. It will translate Slang words from tweets. These words are usually important for sentiment analysis. It will transfer these slang words into their standard forms by using the Slang Word Dictionary on internet and then add that words into the tweets. It will also filter out Non-English

tweets if more than 20 percent of its words. It will also remove URL from tweets.

4.1.2 Sentiment Label Assignment

It assigns the sentiment labels by using two state-of-the-art sentiment analysis tools [1]. SentiStrength3 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. Another tool is TwitterSentiment4 based on the classifiers outputs; it will allot the sentiment label (positive, neutral or negative) with the maximum probability as the sentiment label of a tweet. By combining these two tools the following strategy is designed for more accuracy: 1) If both tools make the same judgment, adopt this judgment. 2) If the judgment of one tool is neutral while the other is not, trust the non-neutral judgment. 3) In the case where the one judgment is positive and one is negative, it takes sentistrength tools judgment. In this LDA algorithm used which represents documents as mixtures of topics that spit out words with certain probabilities.

4.1.3 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 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.

Algorithm (LDA)

LDA assumes the following generative process for a corpus D, Consisting of M documents each of length N_i.

Where, α is the parameter of the Dirichlet prior on the per-document topic distributions,
β is the parameter of the Dirichlet prior on the per-topic word distribution,
θ_i is the topic distribution for document i,
ϕ_k is the word distribution for topic k,
z_{ij} is the topic for the jth word in document i, and
w_{ij} is the specific word.

Step 1: Choose $\theta_i \sim \mathrm{Dir}(\alpha)$, where $i \in \{1,\ldots,M\}$ and $\mathrm{Dir}(\alpha)$ is the Dirichlet distribution for parameter α

Step 2: Choose $\phi_k \sim \mathrm{Dir}(\beta)$, where $k \in \{1,\ldots,K\}$

Step 3: For each of the word positions i, j,

where $j \in \{1,\ldots,N_i\}$, and $i \in \{1,\ldots,M\}$

(a) Choose a topic
$z_{i,j} \sim \mathrm{Multinomial}(\theta_i)$.

(b) Choose a word
$w_{i,j} \sim \mathrm{Multinomial}(\phi_{z_{i,j}})$

[Reference: Wikipedia (LDA algorithm)]

5. Mathematical Model

Sentiment interpretation on Twitter:

Let, consider Dataset contains number of tweets.
D=Dataset of tweets.
N=Number of sentiments with reasons.

Set Theory Model
Consider,
T=Number of tweets.
S= Set of input, output, function.
S = {I, F, O} I: Input, F: Functions, O: Output.

Set of function F = {F1; F2; F3; F4; F5}
F1 = Preprocess.
F2 = variation points.
F3 = set of foreground topics.
F4 = set of background topics
F5 = set of reason candidate
F1: Preprocess
F1 = {I1; f1; O1}

Consider,
1. Slang words translation (S1)
2: Stop words Removal (S2)
f1 = {S1 + S2}
Now,
I1 = {Tweets, f1}
O1 = {Tweets1}
Where,
Tweets1 =Tweets Extraction.
F1 = {I1; f1; Tweets1}

F2: Variation points
F2 = {Output of F1; LB}
Where,
LB is a Label assign to tweets that is Positive or Negative.
Output of F2 = {F1; LB}

F3: Set of foreground topics
F3 = {Output of F2; FB}
Where,
FB is a set of Foreground topics.
Output of F3 = {F2; FB}

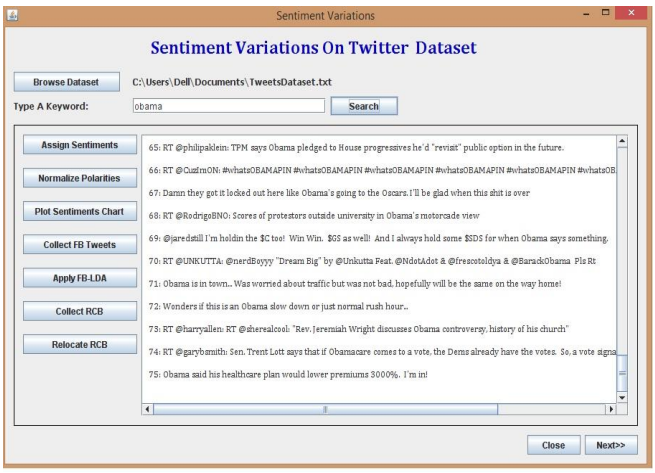
F4: Set of Background topics
F4 = {BB}
Where,
BB is a set Background topic.

F5: Set of Reason candidates
F5 = {Output of F3; FB; F4; BB}
Where,
FB is a set of Foreground topics.
BB is a set of Background topics.
Output of F5 = {F3; FB; F4; BB}

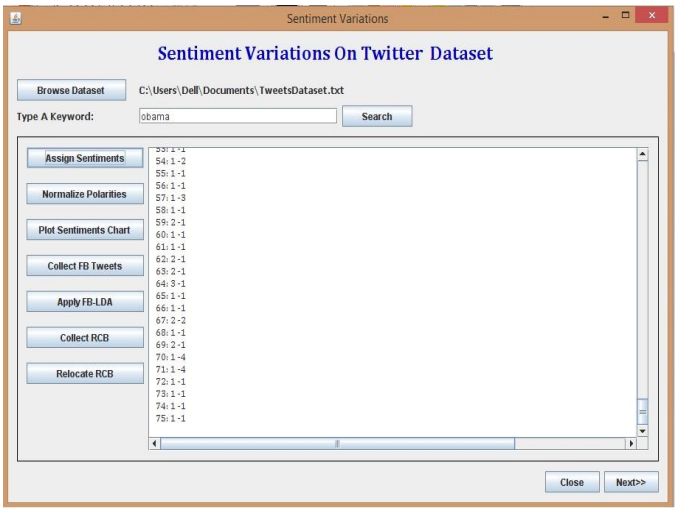
Where,
Output O gives the public sentiment interpretation With its ranked reasons occurred behind them.

6. Results

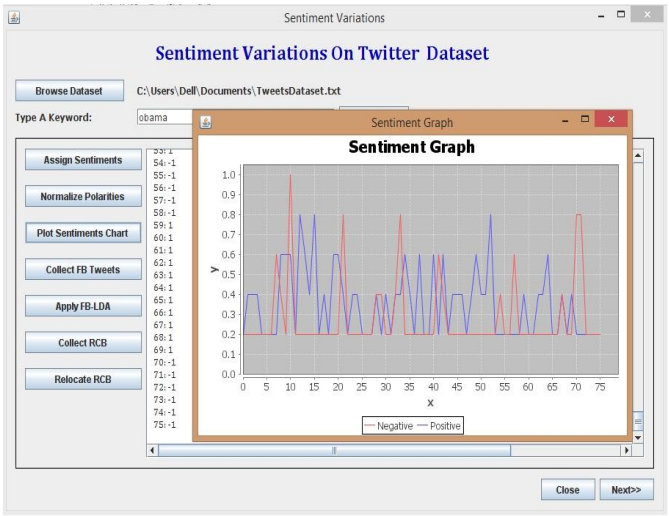
In this it shows the all set of tweets which are extracted from twitter dataset related with target query that is "Obama".



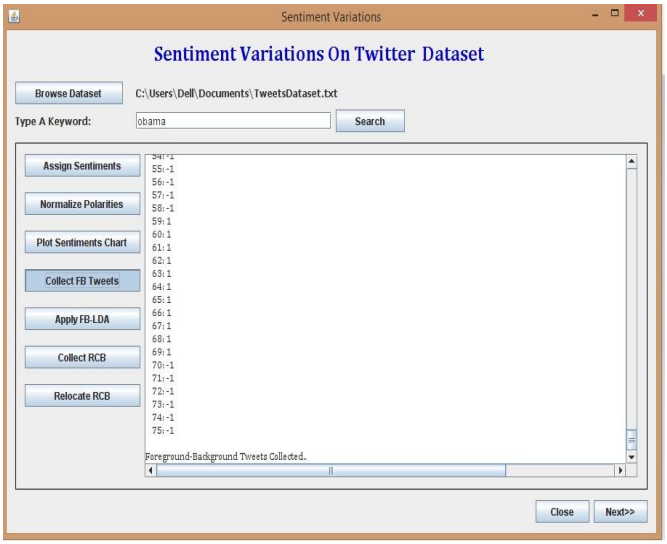
This shows the sentiment score for each tweet according to sentistrength tool.



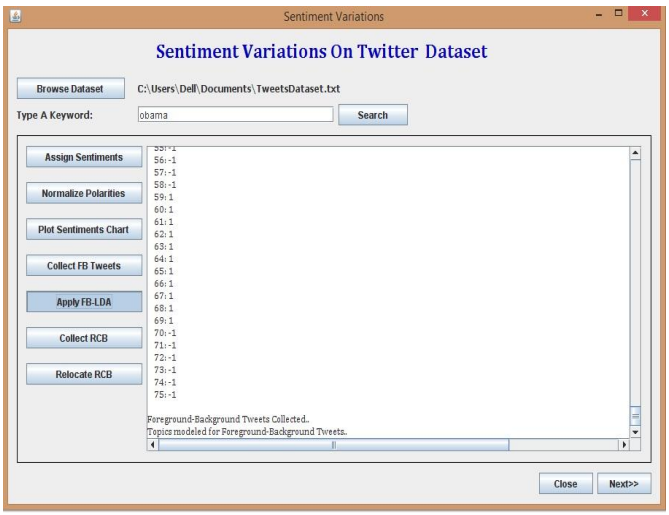
This is the variation graph of positive and negative sentiments.



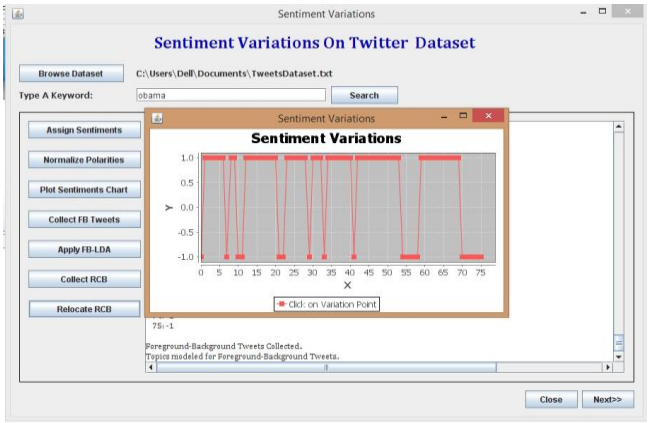
This shows the collection of Foreground and Background tweets according to variation points.



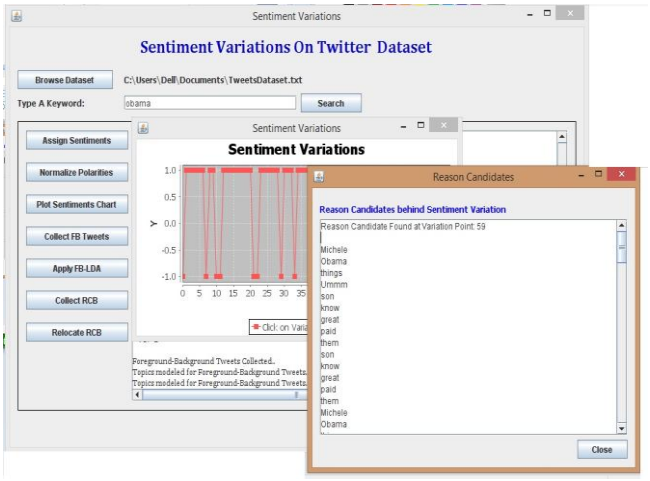
This is the topic modeling for foreground and background tweets FB-LDA.



This shows the graph for RCB with its variation points.



This shows the final output that is variation points with its reason candidate behind the variation of sentiments.



7. 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.

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. Intell., 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.

Author Profile

Ms.Devaki V. Ingule received the Bachelor degree (B.E.) in Computer engineering in 2012 from Satara college of Engineering and Management, SATARA, Shivaji University. Currently, She is pursuing Master's degree in Computer Engineering at Vidya Pratishthan's College of Engineering, BARAMATI, Pune University. Her current research interests include Data Mining and Information Retrieval.

Prof. Mrs. Gyankamal J. Chhajed obtained Engineering degree (B.E.) in Computer Science and Engineering in the year 1991-95 from S.G.G.S.I.E.T, Nanded and Postgraduate degree (M.Tech.) in Computer Engineering from College of Engineering, Pune (COEP) in the year 2005-2007 both with distinction. She is approved Undergraduate and postgraduate teacher of Pune university and having about 17 yrs. of experience. Gyankamal authored a book and has 21 publications at the national, international level for Conferences and Journal. She is life member of the ISTE & International Association IACSIT. Her research interests include Steganography and Watermarking, Image processing, Data mining and Information Retrieval, Biomedical Engineering.