ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Interpretation of Public Sentiment Variations Using Tweets

Pankaj Bhalerao[1], Trupti Dange[2]

P. G. Student, Dept. of Computer Engineering, RMD Sinhgad College of Engineering, SavitribaiPhule Pune, India[1]

Assistant Professor, Dept. of Computer Engineering, RMD Sinhgad College of Engineering, SavitribaiPhule Pune, India[2]

ABSTRACT: Sentiment analysis or opinion mining is an important type of text analysis that aims to support decision making by extracting and analyzing opinion oriented text, identifying positive, negative, neutral opinions and measuring how positively or negatively an entity is regarded. Peoples express their daily life events on twitter also express their views on product, topics and also express political and religious views on Twitter. So tweets become valuable sources of people's opinions and can be efficiently used to infer people's opinions for marketing or social studies. Previous researches are cantered on classifying and tracking public sentiment but cannot find the exact causes of sentiment variations. It is noticed that emerging topics (named foreground topics) within the sentiment variation periods are highly related to the genuine reasons behind the variations. Latent Dirichlet Allocation (LDA) based model, Foreground and Background LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA) models are used to find the exact cause of sentiment variations.

KEYWORDS: Sentiment analysis, Twitter, LDA, public sentiment, emerging topic mining.

I. INTRODUCTION

Introduction Sentiment analysis is also known as opinion mining refers to the use of natural language processing aims to determine the attitude of a speaker or a writer with respect to some topic. The attitude may be his or her judgment or evaluation [4]. The rise of social media such as blogs and social networks has driven interest in sentiment analysis. Due to the proliferation of reviews, ratings and other forms of online opinion ,online expressions has turned into a kind of platform for businesses looking to market their products, identify new opportunities and manage their reputations[20]. Sentiment analysis is an exciting new research field with the potential for a number of real world applications where discovered opinion information can be used to help people or companies or organizations to make better decisions. For example, if public sentiment changes greatly on some products, the related companies may want to know why their products receive such feedback. If negative sentiment towards Barack Obama increases significantly, the White House Administration Office may be eager to know why people have changed their opinion and then react accordingly to reverse this trend [20].

As Main application of sentiment analysis is to classify a given text to one or more pre-defined sentiment categories and can be used for decision making in various domains. It is generally difficult to find the exact causes of sentiment variations since they may involve complicated internal and external factors. It is observed that the emerging topics discussed in the variation period could be highly related to the genuine reasons behind the variations. When people state their opinions, they often state reasons (e.g. some specific events or topics) that support their current view [20].The Proposed system can analyze public sentiment variations on Twitter and mine possible reasons behind such variations. To track public sentiment, The naïve bayes algorithm is used to obtain sentiment information towards interested targets (e.g. Obama, Apple) in each tweet. After obtaining the sentiment label for each tweet, For tracking the public sentiment regarding the corresponding target some descriptive statistics (e.g., Sentiment Percentage) is used. On the tracking curves significant sentiment variations can be detected with a pre-defined threshold (e.g., the percentage of negative tweets increases for more than 50%).

DOI: 10.15680/ijircce.2015.0306110

The Latent Dirichlet Allocation (LDA) based models are used to analyze tweets in significant variation periods, and infer possible reasons for the variations [2]. The First LDA –Based model, called Foreground and Background LDA (FB-LDA), can filter out background topics and extract foreground topics from tweets in the variation period, with the help of a supplementary set of background tweets generated just before the variation. By taking away the interference of longstanding background topics, FB-LDA can address the first challenge. To handle the next challenges, another generative model called Reason Candidate and Background LDA (RCB-LDA). RCB-LDA first extracts representative tweets for the foreground topics (obtained from FB-LDA) as reason candidates. Then RCB-LDA associate each remaining tweet in the variationPeriod with one reason candidate and rank the reason candidates by the number ofTweets associated with them.

II. Related work

In recent years many researches are carried out on Social Network Analysis and sentiment analysis. B. Pang and L. Lee conducted a detailed survey of the existing methods on sentiment analysis [4]. Previous studies like Oconnor et al. [20] concentrated on tracking public sentiment on Twitter and studying its correlation with consumer confidence and presidential job approval polls. While the results are not satisfactory, the results imply that advanced NLP techniques are needed to get better opinion estimation. Similar researches have been done for investigating the reaction of public sentiment on stock markets [9] and oil price hike [5].

It is observed that events in real life indeed have a significant and immediate effect on the public Sentiment on Twitter [5],[9]. However, not any of these studies performed further analysis to mine useful insights behind important sentiment variation, called public sentiment variation. One valuable analysis is to find possible causes behind sentiment variation, which can give important decision-making information. For example, if negative sentiment towards Barack Obama increases significantly, The White House Administration Offices may be excited to know why people have changed their opinion and then react accordingly to reverse this trend. Another illustration is, if public sentiment changes greatly on some products, the related companies maywant to know why their products receive such feedback [1]. It is generally difficult to find the exact causes of sentiment variations since they may involve complicated internal and external factors.

It is observed that the emerging topics discussed in the variation period could be highly related to the genuine reasons behind the variations. When people state their opinions, they often state reasons (e.g. some specific events or topics) that support their current view. Text clustering and summarization techniques [7], [18] are not appropriate for this task since they will discover all topics in a text collection The events and topics related to opinion variation are difficult to represent. Twitter has become very popular, with hundreds of millions of tweets being posted every day on a wide range of topics. This has helped make real-time search applications possible with prominent search engines routinely displaying relevant tweets in response to user queries. Recent study has revealed that a considerable fraction of these tweets are about events, and the detection of new events in the tweet-stream has attracted a lot of research interest [8][18]. However, very little research has focused on accurately displaying this real-time information about events. For example, the leading search engines only display alltweets matching the queries in reverse chronological order. In this method we argue that for some highly structured and recurring events, such as sports, it is better to use more sophisticated techniques to summarize the relevant tweets. The problem of summarizing event-tweets and given a solution based on learning the underlying hidden state representation of the event via Hidden Markov Models [7], [13]. In addition, through extensive experiments on real-world data. it is observed that this model significantly outperforms some in-built and competitive baselines. Keyword created by topic modelling [2] can describe the underlying events to some extent. But they are not as instinctive as natural language sentences. A probabilistic model for collections of discrete data such as text corpus. LDA is a three-level hierarchical Bayesian model, in which every item of a collection is modelled as a infinite mixture over an underlying set of topics [2]. Each topic is displayed as an infinite mixture over an underlying set of topic probabilities. In the perspective of text modelling, the topic probabilities provide a precise representation of a document.

There are several studies focusing on analyzing the relations between online public sentiment and real-life events (Example-Consumer confidence, Stock market)[9] . It isobserved that events in real life indeed have a significant and immediate effect on the public sentiment in Twitter. Based on such correlations, several other works,

DOI: 10.15680/ijircce.2015.0306110

made use of the sentiment signals in blogs and tweets to predict movie sales and elections. Their results indicate that online public sentiment is indeed a good indicator for elections [20] and movie sales [21]. Unlike from the existing work in this line, the proposed system can analyze possible reasons behind thepublic sentiment variations.

III. PROPOSED METHODOLOGY

Following Fig.1 shows the supportive architectural design of A Novel Approach for Interpreting Public Sentiment Variations on Twitter with its explanation in subsequent section.

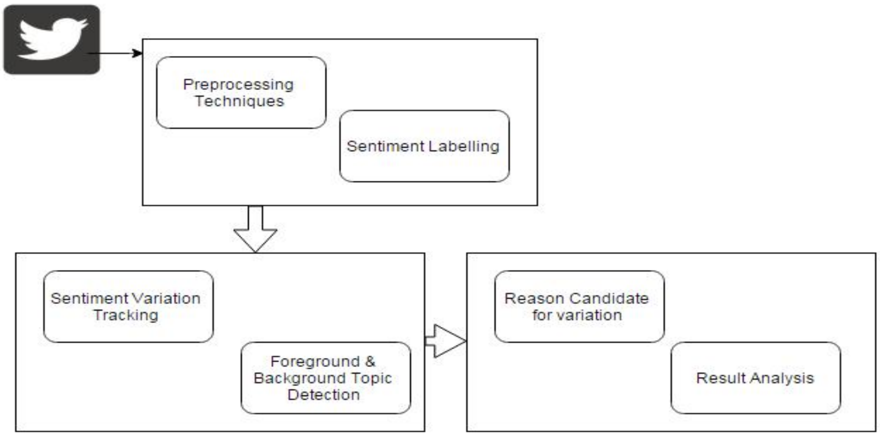

Preprocessing Techniques
Sentiment Labelling
Sentiment Variation Tracking
Foreground & Background Topic Detection
Reason Candidate for variation
Result Analysis

Fig. 1. System Architecture

Twitter Data as Input

Twitter messages are taken as input, as twitter messages are very informal , these massages are filter out messages using techniques like, url filtering, Slang words translation, Non-English tweets filtering, stopwords removal. To extract tweets related to the target, we can go through the whole dataset and extract all the tweets which contain the keywords of the target. Compared with usual text documents, tweets are generally less formal and repeatedly written in an adhoc manner. Sentiment analysis tools applied on raw tweets repeatedly achieve very poor performance in most cases. Therefore, pre-processing techniques on tweets are needed for obtaining satisfactory outcomes on sentiment analysis. The naïvebayes algorithm is used for this purpose and the messages are labelled as positive or negative or neutral sentiment.

Sentiment Variation Tracking

Once obtaining the sentiment labels of all extracted tweets regarding a target, The Proposed system can track the sentiment variation using various descriptive statistics. Here the percentage of positive or negative tweets among all the extracted tweets is adopted as an indicator for tracking sentiment variation over time [1]. Based on these descriptive statistics, sentiment variations can be found using various heuristics (Example- the percentage of positive/ negative tweets increases for more than 50%) [1].

FB- LDA

Foreground and Background LDA (FB-LDA), can filter out background topics and extract foreground topics from tweets in the variation period, with the help of supplementary set of background tweets generated just before the variation. The FB-LDA algorithm is used for this purpose.

Reason Ranking of RCB-LDA

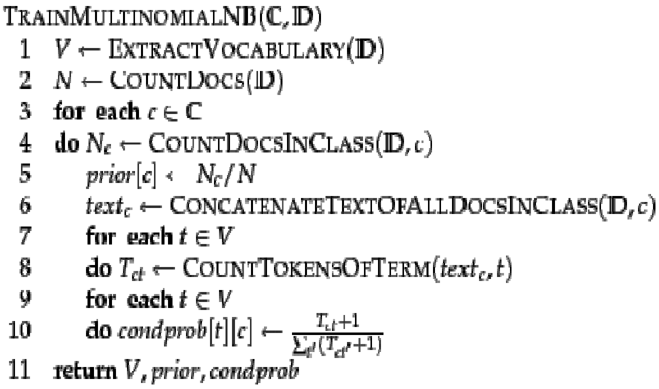
RCB-LDA first extracts representative tweets for the foreground topics (obtained from FB-LDA) as reason

DOI: 10.15680/ijircce.2015.0306110

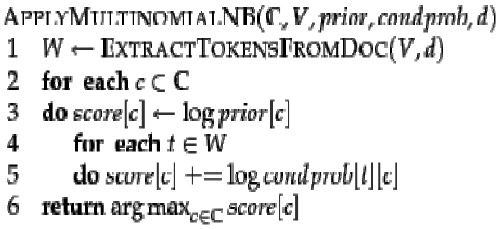
Candidates. Then it will associate each remaining tweet in the variation period with one reason candidate and rank the reason candidates by the number of tweets associated with them [1]. The RCB-LDA algorithm is used for this purpose.

IV. Algorithms Used

A. Naive Bayes Algorithm for Proposed System



```
TrainMultinomialNB($\mathbb{C}, \mathbb{D}$)
$V \leftarrow$ ExtractVocabulary($\mathbb{D}$)
$N \leftarrow$ CountDocs($\mathbb{D}$)
for each $c \in \mathbb{C}$
do $N_c \leftarrow$ CountDocsInClass($\mathbb{D}, c$)
    $prior[c] \leftarrow N_c/N$
    $text_c \leftarrow$ ConcatenateTextOfAllDocsInClass($\mathbb{D}, c$)
    for each $t \in V$
    do $T_{ct} \leftarrow$ CountTokensOfTerm($text_c, t$)
    for each $t \in V$
    do $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
return $V, prior, condprob$
```



```
ApplyMultinomialNB($\mathbb{C}, V, prior, condprob, d$)
$W \leftarrow$ ExtractTokensFromDoc($V, d$)
for each $c \subset \mathbb{C}$
do $score[c] \leftarrow \log prior[c]$
    for each $t \in W$
    do $score[c] \mathrel{+}= \log condprob[t][c]$
return $\arg\max_{c \in \mathbb{C}} score[c]$
```

B. Algorithmic steps for generative process of FB-LDA

1. Go through each tweet, and randomly assign each word in the tweet to one of the K topics.
2. Notice that this random assignment already gives you both topic representations of all the tweets and word distributions of all the topics.
3. So to improve on them, for each tweet d...
4. foreach word w in d...
And for each topic t, compute two things:
    1) p(topic t | tweet d) = the proportion of words in tweet d that are currently assigned to topic t,
    2) p(word w | topic t) = the proportion of assignments to topic t over all tweets that come from this word w.
5. Reassign w a new topic, where you choose topic t with probability p(topic t | tweet d) * p(word w | topic t.
6. In other words, in this step, we're assuming that all topic assignments except for the current word in question are correct, and then updating the assignment of the current word using our model of how tweets are generated.
7. After repeating the previous step a large number of times, we will eventually reach a roughly steady state where your assignments are pretty good. So use these assignments to estimate the topic mixtures of each tweet (by counting the proportion of words assigned to each topic within that tweet) and the words associated to each topic (by counting the proportion of words assigned to each topic overall).

C. Algorithm steps for generative process of RCB-LDA

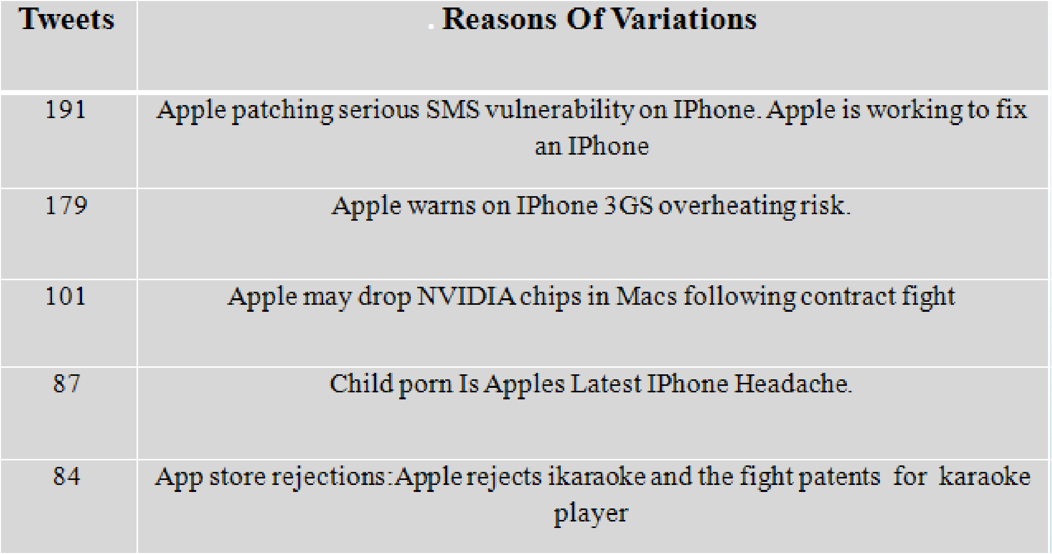
We automatically select reason candidates by finding the most relevant tweets for each foreground topic learnt from FB-LDA, using the following measure:

1. Get Foreground topics, and find relevance of it tweets using below formula:

$$Relevance(t, k_f) = \sum_{i \in t} \phi_f^{k_f, i},$$

where φ kf f is the word distribution for the foreground topic kf and i is the index of each non-repetitive word in tweet t.
2. For each tweet find word distribution and Find word relevance.
3. Extract tweets which have more relevance.
4. Display its count and tweet.

V. EXPECTED RESULTS



Tweets	Reasons Of Variations
191	Apple patching serious SMS vulnerability on IPhone. Apple is working to fix an IPhone
179	Apple warns on IPhone 3GS overheating risk.
101	Apple may drop NVIDIA chips in Macs following contract fight
87	Child porn Is Apples Latest IPhone Headache.
84	App store rejections: Apple rejects ikaraoke and the fight patents for karaoke player

Fig. 2.Ranking results of reason candidates by RCB-LDA.

The proposed system finds the sentiment variations and mines the possible reasons behind these variations as shown in above fig. 2. The above fig. 2 shows the expected results i.e. reasons for of a negative sentiment variation towards “Apple “and number of tweets associated with them.

VI. CONCLUSION AND FUTURE WORK

The Proposed system investigates the problem of analyzing public sentiment variations and finding the possible reasons causing these variations. To solve the problem, two Latent Dirichlet Allocation (LDA) based models; Foreground and Back ground LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA) are used. The Naive bayes algorithm is used for sentiment Classification which gives the better accuracy than existing systems. The proposed models finds the sentiment variation and mine possible reasons behind these sentiment variations. In Future the real time data associated with the twitter account can be given to the system by using a twitter plug-ins and sentiment variations and reasons behind the sentiment variations can be find out on real time data.

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] D.M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res., vol. 3, pp. 9931022, Jan. 2003.
[3] H. Becker, M. Naaman, and L. Gravano, Learning similarity metrics for event identi_cation in social media, in Proc. 3rd ACM WSDM, Macau,

China, 2010.
[4] B. Pang and L. Lee, Opinion mining and sentiment analysis, Found. Trends Inform. Retrieval,vol. 2, no. (12), pp 1135, 2008.
[5] J. Bollen, H. Mao, and A. Pepe, Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.
[6] D. Hall, D. Jurafsky, and C. D. Manning, Studying the history of ideas using topic models, in Proc. Conf. EMNLP, Stroudsburg, PA, USA, 2008, pp. 363371.
[7] D. Chakrabarti and K. Punera, Event summarization using tweets, in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.
[8] T. L. Gri_ths and M. Steyvers, Finding scienti_c topics, in Proc. Nat. Acad. Sci. USA, vol. 101, (Suppl. 1), pp. 52285235, Apr. 2004.
[9] J. Bollen, H. Mao, and X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci., vol. 2, no. 1, pp. 18, Mar. 2011.
[10] G. Heinrich, Parameter estimation for text analysis, Fraunhofer IGD, Darmstadt, Germany, Univ. Leipzig, Leipzig, Germany, Tech. Rep., 2009.
[11] Z. Hong, X. Mei, and D. Tao, Dual-force metric learning for robust disractor-resistanttracker, in Proc. ECCV, Florence, Italy, 2012.
[12] A. Go, R. Bhayani, and L. Huang, Twitter sentiment classification using Distant supervision, CS224N Project Rep., Stanford: 112, 2009. 40.
[13] M. Hu and B. Liu, Mining and summarizing customer reviews, in Proc. 10th ACM SIGKDD, Washington, DC, USA, 2004.
[14] Y. Hu, A. John, F. Wang, and D. D. Seligmann, Et-lda: Joint topic modeling for aligning events and their twitter feedback, in Proc. 26th AAAI Conf. Artif.Intell., Vancouver, BC, Canada, 2012.
[15] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, Target-dependent twitter sentiment classification, in Proc. 49th HLT, Portland, OR, USA, 2011.
[16] J. Leskovec, L. Backstrom, and J. Kleinberg, Meme-tracking and the dynamics of the news cycle, in Proc. 15th ACM SIGKDD, Paris, France, 2009.
[17] C. X. Lin, B. Zhao, Q. Mei, and J. Han, Pet: A statistical model for popular events tracking in social communities, in Proc. 16th ACM SIGKDD, Washington, DC, USA, 2010.
[18] F. Liu, Y. Liu, and F. Weng, Why is SXSW" trending? Exploring multiple text sources for twitter topic summarization, in Proc. Workshop LSM, Portland, OR, USA, 2011.
[19] T. Minka and J. La_erty, Expectation-propagation for the generative aspect model, in Proc. 18th Conf. UAI, San Francisco, CA, USA, 2002.
[20] B. OConnor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, From tweets to polls: Linking text sentiment to public opinion time series, in Proc. 4th Int. AAAI Conf. Weblogs Social Media, Washington, DC, USA, 2010.
[21] G. Mishne and N. Glance, Predicting movie sales from blogger sentiment, in Proc. AAAI-CAAW, Stanford, CA, USA, 2006.
[22] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, “Sentiment strength detection in short informal text,” *J. Amer. Soc.Inform. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010.
[23]J. Weng and B.-S. Lee, “Event detection in twitter,” in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, Barcelona, Spain, 2011.

BIOGRAPHY

Pankaj Bhalerao is a PG Student in Department of Computer Engineering, RMD SinhgadSchool of Engineering, SavitribaiPhule Pune University, India. He has received B.E. in Computer Engineering from University of Pune,India. His Research interest is Data Mining, Big Data.

Prof.TruptiDange received the B.E. and M.Tech Degrees in Computer Engineering from University of Mumbai,India. She is working as Assistant Professor in Department of Computer Engineering, ,RMDSinhgad School of Engineering, SavitribaiPhule Pune University, India. She is having more than four year experience. Her research interest is Data Mining, Web Technology.