ISSN 2395-1621

Evaluating Public Opinion Variation on twitter

[#1]Pallavi Darekar, [#2]Kalyani More, [#3]Tanuj Deshmukh, [#4]Prof. R.S.Shirbhate

[1]pallavidarekar3@gmail.com
[2]kalyanimore11@gmail.com
[3]tanujadeshmukh1693@gmail.com
[4]radha.shirbhate@gmail.com

[#123]Department of Computer Engineering and Information Technology

[#4]Assistant Prof. Department of Computer Engineering

Bhivarabai Sawant Institute Of Technology Research
Pune, INDIA

ABSTRACT

Millions of user shares their opinion and or emotional felling on twitter. This platform valuable and effective tracking and analysing public sentiments. It is useful for decision making in various domains therefore industry attracted to work on twitter platform. Previous work mainly focuses on modelling and tracking public sentiment but in this project we move one step farther to interpreted sentiment variation. This variation period related on genuine reason variation for this we used LDA model that is foreground and background FB-LDA model it is used to separate foreground topic and filter out longstanding background topic and another model is reasons candidate and background (RCB-LDA) this model is used for finding the changeable reasons in variation period .

Keywords: **Sentiment analysis, Twitter Latent Dirichlet Allocation, Public sentiment Emerging topic mining, Naïve bayes.**

ARTICLE INFO

Article History

Received: 9th October 2015

Received in revised form :

10th October 2015

Accepted : 14th October 2015

Published online :

15th October 2015

I. INTRODUCTION

Twitter is launch in 2006 and at the end of 2014 hundred million of user are used this side and authorized user are only read and post the tweets. Unauthorized user only read the post. All over the world this side is used. Twitter is valuable platform because it contains very useful information. That information is used in decision making for various domains. Twitter platform are used for tracking and analysing public sentiments. It is very effective way to expose the public opinion. There is no any restriction we can share our opinion or emotion on twitter. It mainly focus on public opinion. For example company can launch our new product in market and they can study the public opinion towards its products. If public opinion are positive then they knows product are popular in market and if negative feedback are obtain then they concluded that some drawback are in our product. They obtain the positive and negative tweets towards its products. They can overcome the drawback and improve the products quality. To finding the opinion variation LDA model are used. Foreground and background LDA can used to filter out background topics. Reason candidate and background LDA are used to removing the interference of background topics.

II. EXISTING SYSTEM

Sentiment Variation tracking is the procedure to find the change in people's opining about a particular product. Opinion changing from negative to positive or positive to negative. The existing system combines two sentiment analysis tools i.e. SentiStrength and Twitter Sentiment to track the polarity of the tweets. After obtaining the sentiment labels of all extracted tweets about a target, we can track the sentiment variation using some descriptive statistics. In this work, we are interested in analysing the time period during which the overall positive (negative) sentiment climbs upward while the overall negative (positive) sentiment slides downward. In this case, the total number of tweets is not informative any more since the number of positive tweets and negative tweets may change consistently. Here we adopt the percentage of positive or negative tweets among all the extracted tweets as an indicator for tracking sentiment variation over time. Based on these descriptive statistics, sentiment variations can be found using various heuristics (e.g., the percentage of positive/negative tweets increases for more than 50%).To generated reasons for the variation we use two algorithms called Foreground and background LDA (FB-LDA) and Reason candidate and Background LDA (RCB-LDA).FB-LDA can distinguish the foreground topics out of the

background or noise topics. Such foreground topics can help reveal possible reasons of the sentiment variations, in the form of word distributions. The tweets generated by FB-LDA are considered as the reason candidate for the RCB-LDA. It finds the most relevant tweets for each foreground topic learnt from FB-LDA. The candidate with maximum number of relevant tweets is the reason behind the variation.

III. DESIGN PROCESS

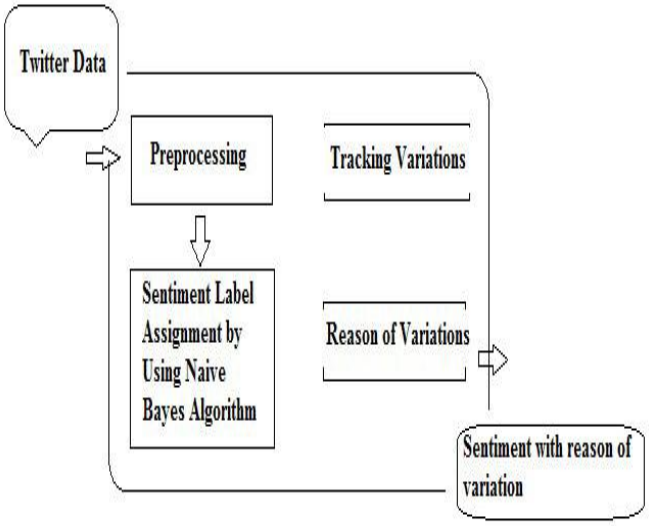

Fig 1: Architecture of the system

Following are the different modules to be performed to achieve the reasons behind the variation in sentiment of the public expressed in the form of tweets are:

Preprocessing: Twitter messages are taken as a input data, as twitter messages are extremely casual, these massages are filter out using various methods like, URL filtering, Slang words translation, Non-English tweets filtering, stop word removal .To extract tweets identified with the target, we can experience the entire dataset and concentrate all the tweets which contain the catch phrases of the target.

Sentiment Labeling: Utilizing different tools like Sent strength, Twitter sentiment, and the messages are labeled as positive or negative or unbiased sentiment, based on tools.

Sentiment Variation Tracking: Once the sentiment labels are obtained of all extracted tweets about a target, the proposed system tracks the sentiment variation using some descriptive statistics.

Foreground Topics from FB-LDA: To mine foreground topics, there is a need to filter out all background topics from the foreground tweets set. The generative model FB-LDA is used to achieve this goal.

Reason Ranking of RCB-LDA: In this method we automatically select reason candidates by finding the most relevant tweets (i.e., representative tweets) for each foreground topic learnt from FB-LDA.

IV. MODULE DESCRIPTION

1. Pre-processing

Tweets are less formal compared to the general document and often written in adhoc manner. Sentiment analysis performed on raw tweets often achieves poor performance in most cases. So we need to pre-process the tweets before performing the opinion analysis.
The main pre-processing tasks we considered are the following:

- Removing Uniform Resource Locator. References to user names, and hash tags.
- Reduction of replicated characters (e.g. wahhhhhhhhhh→wah).
- Identifying emotions and interjections and replacing them with polarity or sentiment expressions (e.g. :-) → good).
  Non English word filtering in advance using English dictionary

2. 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.

Step 1: Go through each tweet, and randomly assign each word in the tweet to one of the K topics.
Step 2: Notice that this random assignment already gives you both topic representations of all the tweets and word distributions of all the topics.
Step 3: So to improve on them, for each tweet d...
Step 4: for each 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.
Step 5: Reassign w a new topic, where you choose topic t with probability p(topic t | tweet d) * p(word w | topic t) (according to our generative model, this is essentially the probability that topic t generated word w, so it makes sense that we resample the current word's topic with this probability). (Also, I'm glossing over a couple of things here, such as the use of priors/pseudo counts in these probabilities.)
Step 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.
Step 7: After repeating the previous step a large number of times, you'll 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).

3. Reason Ranking of RCB-LDA

A subset of tweets is manually labelled in foreground set as the ground reality. Each label contains two elements: one tweet and one candidate (or the background). For each case, 1,000 tweets are manually labelled. Then we extend the labelled set by comparing labelled tweets contents with the unlabelled tweets. If an unlabelled tweet has the same content with a labelled tweet, it should inherit the label from the labelled one.

We automatically select reason candidates by finding the most relevant tweets for each foreground topic.

4. 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. Prior work on burst detection usually chooses the variation of the total number of tweets during time period . Though, in this work, we are interested in analyzing the time period during which the overall positive (negative) opinion climbs upward while the overall negative (positive) opinion slides downward. The total number of tweets is not helpful in this case since the number of positive tweets and negative tweets may change literally. Here the percentage of positive or negative tweets among all the extracted tweets is adopted as an indicator for tracking sentiment variation over time.

V. EQUATIONS

The Bayes Rule

To understand and code for the naive Bayesian algorithm, we will do some math to understand the procedure. The primary equation for the Bayes rule is:

$$P(A|B) = (P(B|A) * P(A)) / P(B)$$

This states mathematically that the posterior probability or probability of future occurrence can be calculated by the product of previous belief P(A) and the likelihood of B if A is true; i.e., P(B|A). P(A|B) is called posterior probability, P(A) is called prior probability, and P(B) is normalization constant. This equation enables us to calculate the probability that A would occur providing that B has happened.

VI. CONCLUSION

The project goal is achieved that relational data classified at three main categories like as positive, negative and neutral by using sentiment analysis labeling. Using the FB-LDA model to filter out background topics and extract the foreground topics. Opinion variation graph is generated and it display the positive, negative and neutral feedback during particular time period then system find out the possible latest reason behind the opinion variation.

REFERENCE

[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 identification 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 scientific 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",In J.Comput. Sci., vol. 2, no. 1, pp. 18, Mar. 2011.

[10] G. Heinrich, "Parameter estimation for text analysis", Fraunhofer IGD, Darm-stadt, Germany, Univ. Leipzig, Leipzig, Germany, Tech. Rep., 2009.

