Volume 4, Issue 11, November 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Analyzing Variation of Public Sentiment on Twitter

Chandar. M. Jadhav, Ashwini P. Patil
Department of Computer Science & Engg
BIGCE College, Solapur
Maharashtra, India

Abstract—Social Networks have become one of the admired communication medium used over internet. Millions of text messages are appearing daily on popular web-sites that provide web services such as Twitter. Millions of users share their opinions on variety of topics and discuss several current issues on Twitter, making it a valuable platform for tracking and analyzing public sentiment. Twitter is a novel micro-blogging platform with more than 25 million unique monthly visitors. On Twitter, any user can publish a message referred to as tweet, which is visible on the public display. Such tracking and analysis can provide critical information for decision making in various domains. In this work, we move one step further to interpret sentiment variations. We observed that emerging topics (named foreground topics) within the sentiment variation periods are highly related to the genuine reasons behind the variations. we propose a Latent Dirichlet Allocation (LDA) based model, Foreground and Background LDA (FB-LDA), to distill foreground topics and filter out longstanding background topics. These foreground topics can give potential interpretations of the sentiment variations. we select the most representative tweets for foreground topics and develop another generative model called Reason Candidate and Background LDA (RCB-LDA) to rank them with respect to their "popularity" within the variation period.

Keywords— Twitter, public sentiment, emerging topic mining, sentiment analysis, latent Dirichlet allocation.

I. INTRODUCTION

Twitter is a worldwide popular website, which offers a social networking and micro blogging services, enabling its users to update their status in tweets, follow the people they are interested in retweet other's posts and even communicate with them directly. Sentiment analysis on Twitter data has provided an economical and effective way to expose timely public sentiment, which is critical for decision making in various domains. For instance, a company can study the public sentiment in Tweets to obtain users' feedback towards its products. As one of the most popular social networking websites, Twitter is drawing more and more attention from researchers from different disciplines. There are several streams of research investigating the role of Twitter. Therefore it has attracted attention in both academia and industry. Previous research mainly focused on tracking public sentiment.

There have been a large number of research studies and industrial applications in the area of public sentiment tracking and modeling. Previous research like O'Connor [1] focused on tracking public sentiment on Twitter and studying its correlation with consumer confidence and presidential job approval polls. Twitter is a novel micro-blogging platform launched with more than 25 million unique monthly visitors. On Twitter, any user can publish a message referred to as tweet , which is visible on the public display.

Similar studies have been done for investigating the reflection of public sentiment on stock markets [4] and oil price indices[3]. They reported that events in real life indeed have a significant and immediate effect on the public sentiment on Twitter. One valuable analysis is to find possible reasons behind sentiment variation, which can provide important information for decision-making. For example, 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. Another example is, if public sentiment changes greatly on some products, the related companies may want to know why their products receive such feedback.

II. LITERATURE SURVEY

Several Researchers carried out research work in Social Network Analysis and sentiment analysis. Sentiment analysis is a text processing technique to derive an opinion or mood intention based on the terms used in a real language sentence. The numbers of researchers have concentrated on generating statistical inference from social network data using sentiment analysis models. Bo Pang and Lilliam Lee [2] provided an insight full discussion on sentiment analysis. They considered the ratio of positive words to total words to estimate the opinion.

Today's users can easily obtain information but also they can actively generate content. News reports, BBS, forums, blogs, and etc are the main sources of public opinion information. The text contains both facts and opinion which could be extracted using natural language processing. Opinions are usually subjective expressions that describe people's sentiments or feelings toward entities and events , it is a sub-discipline of computational linguistics that focuses on extracting people's opinion from the web.

Social media technologies take on many different forms including magazines, Internet forums, weblogs, social blogs, micro blogging, social network, photographs , video, rating and social bookmarking. Twitter is a worldwide popular website, which offers a social networking and micro blogging services, enabling its users to update their status in tweets, follow the people they are interested in, retweet other's posts and even communicate with them directly. Micro blogging websites have evolved to become source of varied kind of information. This is due to nature of micros blogs on which people post real time messages about their opinions on a variety of topics, discuss current issues and express positive, negative sentiment for products they use in daily life. Companies manufacturing such products have started to poll these micro blogs to get a sense of general sentiment for their product.

Public and private opinions about variety of subjects are expressed and spread continually via numerous social media. Sentiment analysis is uesd to determine the attitude of a writer with respect to some topic . The attitude may be his or her judgment, the intended emotional communication or the emotional state of the author when writing. A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence — whether the expressed opinion in a document, a sentence feature is positive, negative, or neutral. Sentiment classification looks, for instance, at emotional states such as 'angry,' 'sad,' and 'happy.'

Sentiment analysis has become popular in judging the opinion of consumers towards various brands [5]. The way in which consumers express their opinion on social networking websites helps to judge this opinion [6].When it comes to sentiment or opinion or emotion we are not concerned with the topic of the text but the positive or negative opinion it express. People can freely express their opinion in social media as reviews, blogs, micro blogs, and forum discussion, social network sites towards any product, service, news or organization. All these platforms provide a huge amount of valuable information that we are interested to analyze.

Twitter making it a valuable platform for tracking and analysing public sentiment. It provide critical information for decision making in various domains. In this work, we interpret sentiment variations. An emerging topics within the sentiment variation periods are related to the genuine reasons behind the variations. Based on this observation, Latent Dirichlet Allocation (LDA) based model, Foreground and Background LDA (FB-LDA), to distill foreground topics. It filter out longstanding *background topics*. These foreground topics can give interpretations of the sentiment variations. we select the most representative tweets for foreground topics and develop model called Reason Candidate and another generative model called Background LDA (RCB-LDA) to rank them with respect to their 'popularity' within the variation period.

Latent Dirichlet Allocation (LDA) based models to analyze tweets in significant variation periods, and infer possible reasons for the variations. This model is called as 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 an auxiliary set of background tweets generated just before the variation.

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 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.

III. PROPOSED METHODOLOGY

A. General Architecture

Social Networking portals have been widely used for expressing opinions in the public domain through internet .Twitter has been the point of attraction to several researchers in important areas. Sentiment analysis over Twitter offers a fast and efficient way to analyze the public sentiment. The main contributions of this paper are two-folds: (1) To the best of our knowledge, our study is the first work that tries to analyze and interpret the public sentiment variations in micro blogging services. (2) Two novel generative models are developed to solve the reason mining problem. The two proposed models are general: they can be applied to other tasks such as finding topic differences between two sets of documents.

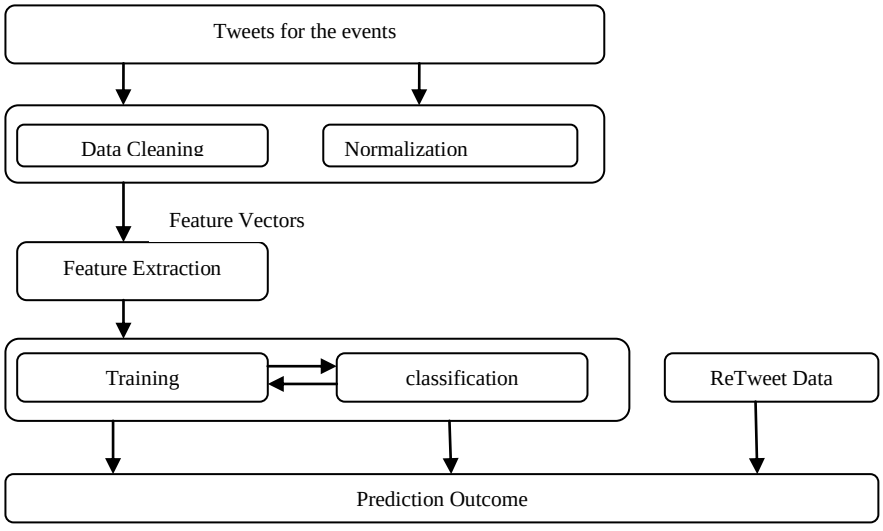

Fig. 1 High Level System Flow

Fig. 1 shows an example of High level system flow. To analyze public sentiment variations, There are two Latent Dirichlet Allocation (LDA) based models: (1) Foregroundand Background LDA (FB-LDA) and (2) Reason Candidate and Background LDA (RCB-LDA). NaïveBayes, SVM, MaxEnt, ANN classifiers with features extracted from Twitter data using feature extraction methods such asUnigram, Bigram and Hybrid (Unigram + Bigrams) for sentiment analysis. In order to remove stop words and extract features, we perform data cleaning and normalization. We extract the target based extended features model [7] by modifying it and twitter user data from the normalized data. vectors are used in part of chunks to train the classifier as a part of incremental training. After utilizing nearly 2/3 rd of the data we test it with 1I3 rd of the data. The sentiment analysis results are incorporated with influence factor to predict the results using prediction model .

B. Proposed Architecture

In our work, sentiment tracking involves the following three steps.

- we extract tweets related to our interested targets (e.g. "Obama", "Apple" *etc*), and preprocess the extracted tweets to make them more appropriate for sentiment analysis.
- Second, we assign a sentiment label to each individual tweet by combining two state-of-the-art sentiment analysis tools [9], [8].
- Finally, based on the sentiment labels obtained for each tweet, we track the sentiment variation regarding the corresponding target using some descriptive statistics.

IV. MODULES

A. Tweets Extraction and Preprocessing

To extract tweets related to the target, we go through the whole dataset and extract all the tweets which contain the keywords of the target. Compared with regular text documents, tweets are generally less formal and often written in an adhoc manner. Sentiment analysis tools applied on raw tweets often achieve very poor performance in most cases. Therefore, preprocessing techniques on tweets are necessary for obtaining satisfactory results on sentiment analysis:

1) Slang words translation: A Tweets often contain a lot of slang words (e.g. lol, omg). These words are usually important for sentiment analysis, but may not be included in sentiment lexicons. Since the sentiment analysis tool we are going to use is based on sentiment lexicon, we convert these slang words into their standard forms using the Internet Slang Word Dictionary1 and then 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 in advance. A tweet is considered as non-English if more than 20 percent of its words (after slang words translation) do not appear in the GNU Aspell English Dictionary.

3) URL removal: *A level-3* A lot of users include URLs in their tweets. These URLs complicate the sentiment analysis process. We decide to remove URLs from tweets.

B. Sentiment Label Assignment

To assign sentiment labels for each tweet more confidently, we resort to two state-of-the-art sentiment analysis tools. One is the SentiStrength3 tool [8]. This tool is based on the LIWC [10] sentiment lexicon. It works in the following way: first assign a sentiment score to each word in the text according to the sentiment lexicon; then choose the maximum positive score and the maximum negative score among those of all individual words in the text; compute the sum of the maximum positive score and the maximum negative score, denoted as Final Score; finally, use the sign of Final Score to indicate whether a tweet is positive, a tweet is neutral or a tweet is negative.

V. CONCLUSIONS

Overall, we conclude that social network based behavioral analysis parameters can increase the prediction accuracy . However, presence of all the entities in unbiased and equal manner is necessary to provide accurate results. In this paper, we investigated the problem of analyzing public sentiment variations and finding the possible reasons causing these variations. we proposed two Latent Dirichlet Allocation (LDA) based models, Foreground and Background LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA). These foreground topics can give potential interpretations of the sentiment variations. we select the most representative tweets for foreground topics and develop another generative model called Reason Candidate and Background LDA (RCB-LDA) to rank them with respect to their "popularity" within the variation period.

The FB-LDA model can filter out background topics and then extract foreground topics to reveal possible reasons. To give a more intuitive representation, the RCB-LDA model can rank a set of reason candidates expressed in natural language to provide sentence-level reasons. Our proposed models evaluated on real Twitter data. Moreover, the proposed models are general: they can be used to discover special topics or aspects in one text collection in comparison with another background text collection.

REFERENCES

[1] B. O'Connor, 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 SocialMedia*, Washington, DC, USA, 2010.

[2] Bo Pang. Lilliam Lee, *"Seeing Stars: Exploiting class relationships for sentiment categorization with respect to rating scales"*, 2002

[3] 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.

[4] J. Bollen, H. Mao, and X. Zeng, "*Twitter mood predicts the stock market,"* J. Comput. Sci., vol. 2, no. 1, pp. 1–8, Mar. 2011.

[5] Bernard J. Jansen, Mimi Zhang, Kate Sobel and AbdurChowdury,"*Micro-blogging as online word of mouth branding"*, 27th International Conference Extended Abstracts on Human Factors in Computing Systems, New York, 2009,pages 3859-3862.

[6] J.C. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou."*Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews"*,Advances In Knowledge and organization, 2004, pages 49-54.

[7] T. Minka and J. Lafferty, "*Expectation-propagation for the generative aspect model*", in *Proc. 18th Conf. UAI*, San Francisco, CA, USA, 2002.

[8] 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.

[9] A. Go, R. Bhayani, and L. Huang, "*Twitter sentiment classification using distant supervision*", CS224N Project Rep., Stanford: 1–12, 2009.

[10] Y. Tausczik and J. Pennebaker, "*The psychological meaning of words: Liwc and computerized text analysis methods*" , *J. Lang.Soc. Psychol.*, vol. 29, no. 1, pp. 24–54, 2010.

