Sentiment Analysis using LDA on Product Reviews: A Survey

Suvarna D.Tembhurnikar
PG Student, Department of Computer engineering, SES's R.C. Patel Institute of Technology, Shirpur, North Maharashtra University, MS, India

Nitin N.Patil
Head & Associate Professor, Department of Computer Engineering, SES's R. C. Patel Institute of Technology, Shirpur, North Maharashtra University, MS, India

ABSTRACT
This paper presents a survey of sentiments analysis for product review. Online social and news media has become a very popular for users to share their opinions and generate prosperous and timely information about real world events of all kinds. Several efforts were dedicated for mining opinions and sentiments automatically from natural language in social media messages, news and commercial product reviews. For this task a deep understanding of the explicit and implicit information are needed. Social media like facebook, twitter, online review websites like Amazon are popular sites where millions of users exchange their opinions and making it a valuable platform for tracking and analyzing public sentiments. This provides important information for decision making in various domains. A lot of research has been done on modeling and tracking public sentiment. Here main focus is given to interpret sentiment variations. It has been observed that emerging topics within the sentiment variation periods are greatly related to the actual reasons behind the variations.

In this paper we are discussing LDA based model for interpreting sentiments. This model is used for giving rank to the tweets with respect to their popularity within the variation period. This method efficiently finds foreground topics and rank reason candidates and also used to find topic differences between two sets of documents.

Keywords
Public Sentiments, Sentiment Classification, Latent Dirichlet Allocation, Sentiment Analysis.

1. INTRODUCTION
Amazon, flipkat, jabang, etc are the popular social network platforms where millions of users can give their views about any product. Sentiment analysis gives an effective and efficient means to expose public opinion timely which gives vital information for decision making in various domains. For obtaining users feedback towards any product, different companies can study the public sentiment in tweets. Many research studies and industrial applications have been done in the area of public sentiment tracking and modeling. If any company gets a greatly changed public sentiment towards its product then the company wants to know why their product gets such types of feedback. It is very difficult to find the correct causes of sentiment variations because they involve complicated internal and external factors. It has been observed that the emerging topics discussed in the variation time could be greatly linked to the actual reasons behind the variations. Mining emerging events is challenging because the tweets gathering in the variation time could be very noisy and contains irrelevant background topics which had been discussed for a long time and did not contribute to the changes of the public's opinion. How to clean out such background topics is an important topic.

In this paper we are discussing LDA based model for interpreting sentiments.

2. LITERATURE REVIEW
Shulong Tan et al., had proposed a two LDA based model (FB-LDA and RCB-LDA) for analyzing public sentiments variations and finding the possible reasons causing this variation [1].

Previous research like O'Connor *et al*., focused on tracking public sentiment on Twitter and studying its correlation with consumer confidence and presidential job approval polls [2]. M. Thelwall et al. and Y. Tausczik et al., describe the SentiStrength tool which is based on LIWC sentiments lexicon. These two tools are used to assign sentiment labels for each tweet [3], [4].

Pang *et al.,* conducted a detailed survey of the existing methods on sentiment analysis. Sentiment analysis, also known as opinion mining which are widely applied to various document types, such as movie or product reviews. Online public sentiment analysis is gradually more popular topic in social network related research. There has been some research work focusing on assessing the relations between online public sentiment and real-life events [5].

G. Mishne et al. and A. Tumasjan et al., reported such type of events in real life indeed have a significant and immediate effect on the public sentiment in Twitter [6], [7].
Sentiments signal used in blogs and tweets to predict movie sales and elections. Online public sentiment is indeed a good indicator for movie sales and elections. D. Chakrabarti et al., summarized the events for better understanding. They try to characterize events using work tweets. He proposed to study the correlation between tweets and events [8].

Hu *et al.* proposed novel models to map tweet to each segmentation in a public speech [9]. RCB-LDA model also focuses on finding the correlation between tweets and events. It is different from previous work. RCB-LDA utilize a background tweets set as a reference to remove noises and background topics. The interference of noises and background topics can be eliminated. The reason mining task is used to show specific information hidden in the text data. It is correlated to data visualization techniques.

D. Tao et al., and X. Tian et al., proposed data visualization technique ranking. Ranking is core techniques in the information retrieval domain which can help find the most relevant information for given queries. The reason mining task cannot be solved by ranking methods because there are no explicit queries in this task [10],[11].

3. CHALLENGES IN SENTIMENT ANALYSIS

Sentiment Analysis or Opinion Mining is a recent subtask of Natural Language processing. There are several challenges in Sentiment analysis field. The first is that, an opinion word which is considered as positive in one situation can be considered as negative in some other situations. Ex.: take the word "long", if a customer said that the battery life of laptop was long, it indicates a positive opinion for laptop. But, if the customer said that the start-up time of laptop was long, then it indicates a negative opinion.

Another challenge is that a person doesn't always express opinions in the similar way. Most of the traditional text processing uses the approach which relies on the fact that minute differences between pieces of text don't change their meaning very much. Ex.: "the product was good" is very different from "the product was not good".

People can be contradictory in using their statements. Some reviews will have both positive as well as negative comments. For example: "the movie bombed even though the lead actor rocked it". This is not much difficult for a human to understand, but for a computer it is not easy to parse.
One drawback of the sentiment analysis using combination of lexicon based and learning based approaches at document level is that reviews with a lot of noise are often assigned a neutral score. The reason for this is that the method fails to detect any sentiment.

4. PUBLIC SENTIMENT TRACKING

For tracking public sentiment the first task is to collect reviews of products from different e-shopping sites. Preprocessing plays an important role in sentiment analysis. It helps to give the more accurate result. Some preprocessing methods are also discussed. Then combine two state-of-the-arts sentiment analysis tools for assigning a sentiment label to every individual tweet. After obtaining sentiment labels for every tweets, used some descriptive statistics for tracking the sentiment variation concerning the related target.

4.1 Extraction of Reviews

Amazon, flipkart, home shop 18, jabong, snapdeal, etc are popular sits for e-shopping. Million of user share their opinion about product and services in these sites. This becomes the great source for information gathering. Reviews of different products can be collected from these sites.

4.2 Preprocessing of Reviews

User generated massages are very noisy. They are less formal and generally used non English words and symbols. If sentiment analysis tools applied on raw tweets then it frequently get very poor performance. Therefore for removing noise and unwanted things different preprocessing techniques are used. They are very important for obtaining satisfactory results on sentiment analysis. Different preprocessing techniques are as follows.

4.2.1 Tokenization

Tokenization is the process of extracting bags of cleaner terms from raw tweets by removing stop words and punctuation, compressing redundant character repetitions and removing IDs or name used in the text for messaging purposes.

4.2.2 Slang words translation

User generated tweets often contains the slang words. Slang word translation means converting the slang words like lol, omg, etc, into their standard form using the Internet Slang Word Dictionary1 and then add them to the tweets.

4.2.3 Non-English tweets filtering

Stemming means a group of different words share the same meaning. Stemming is the process of reducing inflected words to their stem. This process reduces the number of words which share the same meaning.

4.2.4 Stemming

Stemming means a group of different words share the same meaning. Stemming is the process of reducing inflected words to their stem. This process reduces the number of words which share the same meaning.

4.2.5 URL removal

Many users include URLs in their tweets. These URLs make the sentiment analysis process more complex. So URLs are removed from the tweets.

4.3 POS Tagging

"POS (Parts of speech) tagging" means a type to which a word is assigned in according to its syntactic functions. In English language the main "parts of speech" are pronoun, noun, adjective, verb, adverb, etc. "POS tagging" means assigning the labels (tags) to words in sentence according to its function in the sentence. For assigning a label (tag) to each word, "Stanford POS (Parts of Speech) tagger" is used. A tag is allocated to each word, like, NNS, NN, JJS, JJ, RB, VB, etc.

4.4 Foreground and Background LDA

Foreground and Background LDA (FB-LDA) is used for filtering out background topics and mine foreground topics from tweets in the variation time, with the help of an auxiliary set of background tweets generated just before the variation [1].

STEPS FOR FB-LDA

1) Choose a word distribution $\phi_{f;k_f} \sim Dirichlet(\beta_f)$ for each foreground topic k_f.
2) Choose a word distribution $\phi_{b;k_b} \sim Dirichlet(\beta_b)$ for each background topic k_b.
3) For each tweet b in the background data, $b \in \{1, \ldots, B\}$:
   a) Choose a topic distribution $\mu_b \sim Dirichlet(\alpha_\mu)$.
   b) For each word w_b^i in the tweet, $i \in \{1, \ldots, N_b\}$:
      i) Choose a topic $z_b^i \sim Multinomial(\mu_b)$.
      ii) Choose a word $w_b^i \sim Multinomial(\phi_{b;z_b^i})$.
4) For each tweet t in the foreground data, $t \in \{1, \ldots, T\}$:
   a) Choose a type decision distribution $\lambda_t \sim Dirichlet(\alpha_\lambda)$.
   b) For each word w_t^i in the tweet, $i \in \{1, \ldots, N_t\}$:
      i) Choose a type $y_t^i \sim Bernoulli(\lambda_t)$.
      ii) if $y_t^i = 0$:
         A) Choose a foreground topic distribution $\theta_t \sim Dirichlet(\alpha_\theta)$.
         B) Choose a topic $z_t^i \sim Multinomial(\theta_t)$.
         C) Choose a word $w_t^i \sim Multinomial(\phi_{f;z_t^i})$.
      iii) else (i.e., $y_t^i = 1$.):

Figure 1.1: Modified FD-LDA.

4.5 Reason Candidate and background LDA

Reason Candidate and Background LDA (RCB-LDA) is used for mining representative tweets for the foreground topics as reason candidates and will associate one reason candidates to every remaining tweet in the variation time and rank the reason candidates by the number of tweets associated with them [1].

4.6 Assign Sentiment Label

For assigning sentiment labels for every tweets sentiment analysis tools are used. The SentiStrength tool is based on the LIWC sentiment lexicon. According to the sentiment lexicon a sentiment score is assign to each word in the text and then select the highest positive score and the highest negative score among those of all individual words in the text. Then calculate the sum of the highest positive score and the highest negative score and denoted as Final Score. Lastly, a Final Score is used to indicate whether a tweet is positive, neutral or negative. One more sentiment analysis tool is Twitter Sentiment. It is based on a Maximum Entropy classifier. In this a automatically collected 160,000 tweets with emoticons is used as noisy labels to train the classifier. A sentiment label is assigned based on the classifiers. These two tools are very popular, but still a large proportion of tweets contain noises after preprocessing. Therefore its performance on real datasets is unsatisfactory.

These two techniques are combined for labeling sentiments for improving the performance [1].

1. If the same judgment created by both tool then accept this Judgment.
2. If the judgment of one tool is neutral and other is non neutral then trust the non-neutral judgment.
3. If two judgments conflict with each other means one is positive and one is negative then trust SentiStrength's judgment if the absolute value of FinalScore is larger than 1 if not trust TwitterSentiment's judgment.

4.7 Tracking Sentiment Variation

After finding the sentiment labels of every extracted tweet, the sentiment variations are track using some descriptive statistics. Burst detection generally selected the variation of the total number of tweets over time as an indicator. The main interest has given in analyzing the time period during which the overall positive sentiment climbs upward whereas the overall negative sentiment slides downward. In this the total number of tweets is not useful and the number of positive and negative tweets may change constantly. For tracking sentiment variation over time the percentage of positive or negative tweets among all the extracted tweets as an indicator was adopted [1].

5. CONCLUSION

In this paper the problem of analyzing public sentiment variations and finding the probable reasons causing this variation is studied. Different sentiments analysis tools and techniques are studied for solving problem. The two Latent Dirichlet Allocation (LDA) based models, Foreground and Background LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA) effectively solve the problem. For filtering out background topics and then extracting foreground topics to expose possible reasons the FB-LDA model is used. The RCB-LDA model is used to give a more sensitive representation. To provide sentence-level reasons RCB-LDA rank a set of reason candidates expressed in natural language. This LDA based model is effectively and efficiently used to mine the possible reasons behind sentiment variations. In future this method can also used to find topic differences between two sets of documents.

6. ACKNOWLEDGMENTS

The authors would like to thank fellows of IJCA for their reviews on this paper. I am grateful to my guide Prof. Nitin N. Patil for his valuable suggestions and encouragement. Special thanks to the authors of the reference papers which helps me to understand the different techniques.

7. REFERENCES

[1] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen and Xiaofei He, Interpreting the Public Sentiment Variations on Twitter, *IEEE Transactions On Knowledge And Data Engineering*, Vol. 6, No. 5, May 2013.

[2] 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 Social Media*, Washington, DC, USA, 2010.

[3] 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.

[4] 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.

[5] B. Pang and L. Lee, Opinion mining and sentiment analysis, *Found. Trends Inform. Retrieval*, vol. 2, no. (1–2), pp. 1–135, 2008.

[6] G. Mishne and N. Glance, "Predicting movie sales from blogger sentiment," in *Proc. AAAI-CAAW*, Stanford, CA, USA, 2006.

[7] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, Predicting elections with twitter: What 140 characters reveal about political sentiment, in *Proc. 4th Int. AAAI Conf. Weblogs Social Media*, Washington, DC, USA, 2010.

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

[9] 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.

[10] D. Tao, X. Li, X. Wu, and S. J. Maybank, Geometric mean for subspace selection, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 260–274, Feb. 2009.

[11] X. Tian, D. Tao, and Y. Rui, Sparse transfer learning for interactive video search reranking, *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 8, no. 3, article 26, Jul. 2012.