International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 3, Issue 2, February 2016

Clustered Probabilistic Aspect Summarization for Medical Reviews

[1]Devi Venugopal, [2]Remya R
[1] PG Scholar, [2] Assistant Professor in IT
College of Engineering Perumon
[1] devivenugopal06@ gmail.com [2] remya.cep.it@gmail.com

Abstract- **With the rising popularity of internet, online drug reviews have been proved to be extremely helpful for patients suffering from chronic diseases. Most of the patients search upon online reviews before taking any medicine. Online reviews, blogs, and discussion forums such as WebMD on chronic diseases and medicines are becoming important supporting resources for patients. Extracting useful information from these reviews is very difficult and challenging. Opinion mining or aspect mining involves the extraction of useful information (e.g. positive or negative sentiments of a product) from a large quantity of text opinions or reviews given by Internet users. Various algorithms had been proposed to extract information from the opinion of web users. Some of the algorithms are LDA, sLDA, NMF, SSNMF, DiscLDA and PAAM. Using clustering based probabilistic aspect summarization technique every user and medical experts can view the positive and negative aspects separately generated from large number of medical reviews. So common people can rate the medicines for chronic diseases which has high side effects.**

Index Terms— **Aspect Mining, Drug Reviews, Opinion Mining, Text Mining, Topic Modeling.**

I. INTRODUCTION

The internet is a vast repository of various kinds of knowledge. Due to the emergence and impact of the internet in our day to day lives, people are encouraged to contribute their opinions and reviews to the Internet. Many user centered platforms are now available for sharing information and user interaction, such as Amazon, Face book and Twitter. Nowadays when people are interested in a product or service, besides consulting the product manufacturers and service providers, they refer to the experienced and practical opinions prepared by end users. This has turned out to be very beneficial since it helps people to be more aware of the products and services.

Previous studies in opinion mining [1] deal with popular consumer products such as books, electronic gadgets, etc. Opinion mining in the medical domain has not been explored in detail. It is because patients belong to the minority groups of Internet and are only concerned with specific illnesses or drugs that they experience. Furthermore, people tend to prefer acceptance of opinions from medical professionals rather than patients. Nevertheless, recent studies show that reviews posted by patients are useful and important especially for chronic diseases and drugs with afflicting side effects. Many patients hope to get more information from other patients with similar conditions. They can also share their experience and propose practical ways to identify the symptoms of different diseases and the side effects of various drugs. Online communities provide a positive impact on patient health.

The identification of different features of a product cannot be done by considering just the overall rating of a review. For instance, a camera might provide excellent image quality, but on the other hand, its battery life may be very poor. Various opinion mining approaches have been proposed to extract and group aspects of products and services so as to predict their sentiments and ratings. Approaches that rely on frequency, relation approach, supervised learning and topic modeling are made use of for this purpose. Dealing with the diverse wordings that are used for describing effectiveness, side effects and people's experiences from the drugs is one of the prime challenges. In particular, side effects are drug dependent: a set of side effect symptoms for a drug is very unlikely to be applicable to another drug. This impedes some opinion mining approaches based on lexicons. Most importantly, authors provide descriptions of symptoms, feelings and comments without specifying which aspects are being described. Even though a number of techniques have been proposed for mining correct opinions from the drug reviews given by the users, each technique can be revised so as to increase their efficiency and throughout.

The rest of this paper is organized as follows: Section 2 discusses the related works, Section 3 describes the proposed methodology and finally, Section 4 concludes the paper.

II. RELATED WORKS

Aspect-based opinion mining is becoming popular in recent years. Topic modeling [2] (e.g., LDA [3]) is a popular probabilistic approach, a set of topics which are represented by multinomial distributions over vocabulary words, are inferred. When sorting the words of a topic based on

probabilities, high probability words of a topic are usually semantically correlated. By doing this concept or aspect of the topic can be captured manually. These aspects which are extracted may not be related to the specified class labels and the manual selection of seed words will determine the performance. Words about an aspect tend to co-occur within close proximity to one another in reviews. One of the methods is called Sentence LDA (SLDA), which is a probabilistic generative model that assumes all words in a single sentence are generated from one aspect.

An extension of SLDA called Aspect and Sentiment Unification Model (ASUM) [4] which incorporates aspect and sentiment together to model sentiments toward different aspects. In JST Model [5] sentiment is integrated with a topic in a single language model. During topic inference supervised latent Dirichlet allocation (sLDA) [6] takes the different forms of supervised information. With LDA a response variable is associated with each document in this approach. DiscLDA or Discriminatory LDA [7] is a discriminative variation of Latent Dirichlet Allocation (LDA).In this method, class dependent linear transformation is introduced on the topic mixture proportions DiscLDA first process the information and find topics specific to individual classes as well as topics shared across different classes.

Another generalization of LDA is Labeled LDA [8]. A probabilistic model is Labeled LDA that describes a process for generating a labeled document collection. NMF is Non negative Matrix Factorization (NMF) [9] which is a deterministic method for topic modeling. Topics can be identified by decomposing the data matrix into two low rank matrices. To incorporate the supervised information into NMF a Semi-supervised NMF (SSNMF) [10] is used which is an extension of NMF. Probabilistic Aspect Mining Model (PAMM) [11] is a probabilistic model for finding the aspects which are correlated to class labels from the drug reviews given by the users. Reviews are generated by the patients [12]-[14] suffering from chronic diseases and having drugs with afflicting side effects.

III. PROPOSED SYSTEM

Aspect summarization model is a probabilistic model for finding the aspects which are correlated to class labels from the drug reviews given by the users. Reviews are generated by the patients suffering from chronic diseases and having drugs with afflicting side effects. Many patients are happy to get more information from other patients with similar conditions. The patients having chronic diseases can also share their experience and can suggest practical ways to alleviate symptoms and side effects of drugs. Experiments shows that these online communities were found to have positive impacts on patient health [15].By comparing with other previous approaches, this model focuses on finding aspects correlated to one class label only.

Aspects correlated to different class labels are separately identified. This method avoids the identified aspects which are having mixed contents from different classes. Better and more specific aspects can be found by focusing the task on one class. This approach is different from the method of which reviews are first grouped according to their class labels and followed by inferring aspects for the individual groups. Parameter estimation can be done by using an efficient EM-algorithm. By examining the experimental results of four different drugs show that aspect summarization model is better to find relevant aspects than other common approaches, when measured with mean point-wise mutual information [16] and classification accuracy.

The following summarizes the features of drug reviews.

- Drug reviews have a small number of kinds of aspects: price, ease of use, dosage, effectiveness, side effects and people's experiences.
- Aspects are usually not mentioned explicitly.
- Descriptions of effectiveness, side effects and people's experiences are diverse.
- Side effect and effectiveness descriptions are different from drug to drug.

The data set containing user reviews are the input of this Probabilistic Aspect Summarization Model. Mainly four algorithms are used. They are

LDA: Latent Dirichlet allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.

Gibbs sampling method: The Gibbs sampling algorithm generates an instance from the distribution of each variable in turn, conditional on the current values of the other variables.

Dimensionality Reduction algorithm: Dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration.

EM algorithm: Expectation maximization (EM) algorithm is an iterative method for finding maximum likelihood of parameters in statistical models.

Measuring the quality of the generated aspects is used for performance analysis. It can be achieved with mean point wise mutual information (PMI). PMI is a measure of association between a feature (in this case aspect or word) and a class (i.e. label).How much information we gain. Consider a set of 2K aspects, with each aspect is sorted descending order according to the individual probabilities/values of the words, the top 20 words of the k^{th} ($k = 1, 2, \ldots, 2K$) aspect are selected .The mean point wise mutual information (PMI) of this set of aspects is defined as

$$mean\ PMI = \frac{1}{40K}\sum_{k=1}^{2K}\sum_{i=1}^{20} log\frac{p(w_{k,i}, C_k)}{p(w_{k,i})p(C_k)}$$

where C_k is the class label associated with the aspect k. The probabilities $p(w_{k,i}, C_k)$,$p(w_{k,i})$ and $p(C_k)$ (assuming all probabilities are greater than zero) are empirical probabilities obtained by counting the words and the reviews in the data set. Therefore, mean PMI gives the mean of PMI between a word in the aspect and the class label. In computing mean PMI, a category label ought to be appointed to every derived topic. For supervised algorithms PAMM, SSNMF and DiscLDA, the data was promptly offered. For unsupervised algorithms LDA and NMF, since it absolutely was not clear that category label ought to be related to a derived facet, half the aspects were labeled one and therefore the rest were labeled zero. Aspects derived by aspect summarization model have considerably higher association with the category labels than different algorithms. The mean PMI produces the mean of PMI between a word in the aspect and the class label. This model gives the best performance in comparison with the previous aspect mining algorithms.

The mined aspects can be processed by the following two steps.

A. *Clustering of aspects*

Clustering of the generated aspects can be done. It is done by calculating the senti strength [17] of the each aspect. So each aspect is clustered into two groups i.e. positive aspects and negative aspects. Many approaches to sentiment analysis rely on lexica where words are tagged with their prior polarity i.e. if a word out of context evokes something positive or something negative. Since words can have multiple senses, we address the problem of how to compute the prior polarity of a word starting from the polarity of each sense and returning its polarity strength as an index between

-1 and 1.For example, wonderful has a positive prior polarity, and horrible has a negative prior polarity.

The advantage is that they don't need deep semantic analysis or word sense disambiguation to assign an affective score to a word and are domain independent. The problem of assigning affective scores (between -1 and 1) to words is harder than traditional binary classification tasks (assessing whether a word or a fragment of text is either positive or negative), for an overview. We want to asses pretty, beautiful and gorgeous are positive words. But gorgeous is more positive than beautiful and is more positive than pretty. This is fundamental for tasks such as affective modification of existing texts, where not only words polarity, but also their strength, is necessary for creating multiple graded variations of the original text.

One of the most widely used resources for sentiment analysis is SentiWordNet. SentiWordNet is a lexical resource in which each word is associated with three numerical scores: Obj(s), Pos(s) and Neg(s). These scores represent the objective, positive and negative values of the entry respectively.

B. *Aspect Summarization*

There are a huge number of review documents that include user's opinions for products. To summarize the opinions [18] is one of the relevant topics in natural language processing. We focus on aspects of a medical product in the summarization process. First, identify a relation between aspects and each word in review documents. Our method employs an unsupervised approach for the identification process. Next, need to generate a summary by using the relations. In our system, users obtain the summary by an interactive approach.

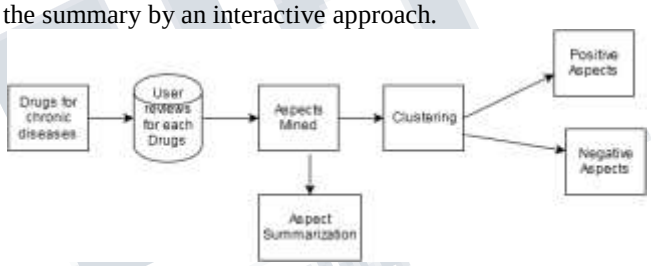

Figure 1 Aspect Mining Architecture

IV. CONCLUSION

Nowadays the online reviews, blogs and discussion forums are very popular for different kinds of products and services. People can write their opinion and experiences through these online communities about the various products including drugs. It is useful and challenging to extract information from these texts. In particular, it is helpful to identify the aspects of a product that will help the people. Every drug is a product of a pharmaceutical company. So they can also view the relevant aspects generated from the user reviews. By the use of dimensionality and classification reduction algorithms, patients can be able to know the relevant aspects from medical reviews. A patient review provides valuable reference from other patient's points of view. Medical domain data mining become one of the focused research areas because of increasing the number of patients and our living environment becomes increasingly polluted. Thus, opinion mining is a field of study which helps to extract aspects from the opinions of the internet users.

REFERENCES

[1] B. Pang and L. Lee, ”Opinion mining and sentiment analysis”, Trends Inf. Ret., vol. 2, no. 12, pp. 1135, Jan. 2008.

[2] D.Mimno and A.McCallum,”Topic models conditioned on arbitary features with Dirichlet multinomial regression”, in Proc.24th Conf. Uncertain. Artif. Intell., 2008.

[3] D. Blei, A. Ng, and M. Jordan, ”Latent Dirichlet allocation”, J. Mach. Learn. Res., vol. 3, pp. 9931022, Jan. 2003.X. P. Zhang, Separable reversible data hiding in encrypted image, IEEE Trans. Inf. Forensics Security, vol. 7, no. 2, pp. 826832, Apr. 2012.

[4] Y. Jo and A. Oh, ”Aspect and sentiment unification model for online review analysis”, in Proc. 4th ACM Int. Conf. WSDM, New York, NY, USA, 2011, pp. 815824.

[5] C. Lin and Y. He, ”Joint sentiment/topic model for sentiment analysis”, in Proc. 18th ACM CIKM, New York, NY, USA, 2009, pp. 375384.

[6] D. Blei and J. Mcauliffe, ”Supervised topic models”, in Proc. Adv. NIPS, 2007, pp. 121128.

[7] S.Lacoste-Julien, F.Sha, and M. Jordan, ”DiscLDA: Discriminative learning for dimensionality reduction and classification”, in Proc. Adv. NIPS, 2008, pp. 897904.

[8] D. Ramage, D. Hall, R. Nallapati, and C. Manning, ”Labeled LDA: A supervised topic model for credit attribution in multilabeled corpora”, in Proc. Conf. EMNLP, Stroudsburg, PA, USA, 2009, pp. 248256.

[9] W. Xu, X. Liu, and Y. Gong, ”Document clustering based on non-negative matrix factorization”, in Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Ret., New York, NY, USA, 2003, pp. 267273.

[10] H. Lee, J. Yoo, and S. Choi, ”Semi-supervised nonnegative matrix factorization”, IEEE Signal Process. Lett., vol. 17, no. 1, pp. 47,Jan. 2010.

[11] Victor C. Cheng, C.H.C. Leung, Jiming Liu, Fellow, IEEE, and Alfredo Milani, ”Probabilistic Aspect based mining model for drug reviews”, in: Proceedings of IEEE transactions on Knowledge and Data Engineering, Vol. 26, No. 8, August 2014.

[12] K. Denecke and W. Nejdl, ”How valuable is medical social media data? content analysis of the medical web”, J. Inform. Sci., vol. 179, no. 12, pp. 18701880, 2009.

[13] X. Ma, G. Chen, and J. Xiao, ”Analysis on an online health social network”, in Proc. 1st ACM Int. Health Inform. Symp., New York, NY, USA, 2010, pp. 297306.

[14] A. Nvol and Z. Lu, ”Automatic integration of drug indications from multiple health resources”, in Proc. 1st ACM Int. Health Inform. Symp., New York, NY, USA, 2010, pp. 666673.

[15] J. Leimeister, K. Schweizer, S. Leimeister, and H. Krcmar, ”Do virtual communities matter for the social support of patients? Antecedents and effects of virtual relationships in online communities”, Inform. Technol. People, vol. 21, no. 4, pp. 350374, 2008.

[16] C. Manning and H. Schtze, ”Foundations of Statistical Natural Language Processing” Cambridge, MA, USA: MIT Press, 1999.

[17] Lorenzo Gatti, Marco Guerini, “Assessing Sentiment Strength in Words Prior Polarities”.

[18] Chia-Hui Chang and Kun-Chang Tsai, “Aspect Summarization from Blogsphere for Social Study”.