International Journal of Computer Applications (0975 - 8887)
Volume 133 - No.15, January 2016

A Review on Different Opinion and Aspect Mining Techniques

Devi Venugopal
P G Scholar
Department of Computer Science & Engineering
College of Engineering, Perumon(CUSAT), Kerala, India

Remya R.
Assistant Professor in IT
Department of Information Technology
College of Engineering, Perumon(CUSAT), Kerala, India

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. A detailed review of the most important opinion mining algorithms is presented and a comparison among the discussed techniques is given.

General Terms

Opinion Mining, Text Mining

Keywords

Aspect Mining, Drug Reviews, Opinion Mining, Text Mining, Topic Modeling

1. 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, Facebook 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 peoples 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. This paper is organized as follows. Section 2 describes the various opinion mining algorithms. Section 3 presents a performance analysis of the various algorithms for opinion mining. Section 4 concludes the review.

2. REVIEW OF VARIOUS OPINION MINING TECHNIQUES

In this section, some of the techniques used for mining information from the user reviews are discussed. The methodologies categorized based on their basis of operation for the survey.

2.1 LDA (Latent Dirichlet Allocation)

An example of a topic modeling approach [2] is LDA [3] in which only the words in the documents are modeled. With this 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. When writing each document using this method at first decide the number of words, N that the document will have and then choose a topic mixture for the document according to the probability. LDA has two advantages that are the content spread of each sentence can be inferred by a word count and can derive the proportions that each word constitutes in given topics.

2.2 Aspect and Sentiment Unification Model

The positions of individual words are neglected for topic inference in LDA. Words about an aspect tend to co-occur within close proximity to one another in reviews. The first method is called Sentence LDA(SLDA) ,which is a probabilistic generative model that assumes all words in a single sentence are generated from one aspect. The advantage of SLDA is,it finds aspects that match the details of the reviews, which is better than LDA.

An extension of SLDA called Aspect and Sentiment Unification Model (ASUM), [4] which incorporates aspect and sentiment together to model sentiments toward different aspects. As illustrated in the following scenario ASUM models the generative process of writing a review. The reviewer first decides to write a review of a restaurant that expresses a distribution of sentiments. If 70% satisfied and 30% unsatisfied, the reviewer decides the distribution of the aspects for each sentiment, say 50% about the service, 25% about the food quality, and 25% about the price for the positive sentiment. Then the reviewer decides, a sentiment to express and an aspect for which he feels that sentiment for each sentence, he may be satisfied with the friendly service of the restaurant. A set of general affective and evaluative words are taken to find the aspect-specific evaluative words in ASUM. Without labeled data this is a simple sentiment word expansion and adaptation. ASUM finds sentiment words related to specific aspects from a small set of general sentiment words.

2.3 Joint Sentiment Topic model

In this JST Model [5] sentiment is integrated with a topic in a single language model. JST does not limit individual words JST is different from ASUM in that individual words may come from different language models. Both JST and ASUM make use of a small seed set of sentiment words, but the exploitation is not explicitly modeled in JST. ASUM integrates the seed words into the generative process, and this provides ASUM with a more stable statistical foundation.

2.4 Supervised LDA (sLDA)

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. This response variable can be indicated by the number of stars given to a movie, a count of the users in an on-line community who marked an article interesting, or the category of a document. In order to find latent topics that will best predict the response variables for future unlabeled documents, sLDA model the documents and the responses. Predictive power of sLDA is more better than unsupervised LDA features.

2.5 DiscLDA

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. The parameter estimation can be done by maximizing the conditional likelihood. Here can obtain a supervised dimensionality reduction algorithm that uncovers the latent structure in a document collection while preserving predictive power for the task of classification by using the transformed topic mixture proportions as a new representation of documents.

2.6 Labeled LDA

Another generalization of LDA is Labeled LDA [8]. A probabilistic model is Labeled LDA that describes a process for generating a labeled document collection. A one-to-one correspondence between LDAs latent topics and user tags is defined in labeled LDA as a topic model that constrains Latent Dirichlet Allocation.

2.7 NMF

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. Find non-negative matrix factors Wand H for a given non-negative matrix V, such that: V WH, where W features (rows), H observations/examples/feature vectors (columns). The statistical analysis of multivariate data applies NMF in the following manner. Given m is the number of examples in the data set for a set of multivariate n-dimensional data vectors, the vectors are placed in the columns of an n x m matrix V. This matrix is then approximately factorized into an n x r matrix Wand an r x m matrix H. Wand H are smaller than the original matrix V because r is chosen to be smaller than nor m. This will result in a compressed version of the original data matrix. Topics recovery, feature learning, clustering etc can be done using NMF.

The base topic of a particular document cluster is captured by NMF, and each document is represented as an additive combination of the base topics. Using NMF the cluster membership of each document can be easily identified. To incorporate the supervised information into NMF a Semi-supervised NMF (SSNMF) [10] is used which is an extension of NMF. In this extended version the topics identified are more closely related to the supervised information.

2.8 PAMM

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. 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 com-

paring with other previous approaches, PAMM 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 PAMM is better to find relevant aspects than other common approaches, when measured with mean point-wise mutual information [16] and classification accuracy. The derived aspects were also examined by humans based on different specified perspectives, which show that PAMM was found to be rated highest.

3. PERFORMANCE ANALYSIS

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 of information gained.

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, . . . , 2K) aspect are selected and denoted by $w_{k,i}{}_{i=1}^{20}$. The mean pointwise mutual information (PMI) ofthis set of aspects is defined as

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

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.

Table 1 explains the mean PMI results. It demonstrates that aspects derived by PAMM have considerably higher association with the category labels than different algorithms. The NMF and LDA have similar performance. Comparable results can be obtained with three supervised algorithms, sLDA, SSNMF and DiscLDA, conjointly. SSNMF and DiscLDA perform higher than NMF and LDA in most cases. This mean PMI results are calculated on the basis of citalopram drug which is used for anti-depression. As a result of the category label data this is often smart and is employed in explanation the aspects.

This performance analysis can be represented in graphical form for better interpretation. It is shown in the figure 1.

Table 1. Evaluated mean PMI of the derived aspects using various algorithms

Product	Algorithm	Mean PMI
Citalopram Drug	NMF	2.03
	LDA	2.03
	sLDA	2.07
	SSNMF	2.06
	DiscLDA	2.07
	PAMM	3.20

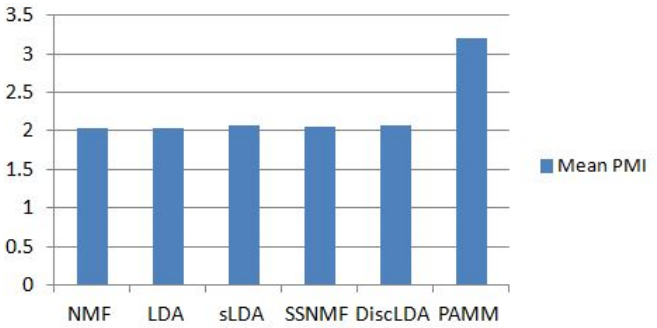

Fig. 1. Mean PMI of the evaluated aspects

Six different opinion mining algorithms are compared here of which some of them are supervised algorithms and some are unsupervised algorithms. Comparisons are done based on calculating mean PMI. The mean PMI produces the mean of PMI between a word in the aspect and the class label. PAMM give the best performance in comparison with the other five algorithms.

4. 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. Literature survey on various opinion and aspect mining methods are discussed here, and also explained how these techniques are useful for mining aspects from drug 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 another patients 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. Designing and implementing opinion mining algorithms is difficult and very complex as the internet data contains large amount of data. When compared to other opinion mining algorithms experimental result shows that PAMM (Probabilistic aspect mining model) gives the best performance.

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

[5] 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.

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

[7] S. Lacoste-Julien, F. Sha, andM. 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.