International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Extraction of Aspects from Drug Reviews Using Probabilistic Aspect Mining Model

Pooja Gawande[1], Sandeep Gore[2]

[1]PG Student of Computer Engineering, G. H. Raisoni College of Engineering and Management, Pune, India

[2]Assistant Professor Of Computer Engineering, G. H. Raisoni College of Engineering and Management, Pune, India

Abstract: ***Reviews of medication from patients are numerous on the internet. This review provides a brief overview of approaches to aspect mining as they relate to drug discovery. Many adverse drug reactions on chronic diseases are not discovered during limited pre-marketing clinical trials, they are only observed after long term post-marketing investigation of drug usage. The detection of adverse drug reactions, as early as possible, is an important topic of research for the pharmaceutical industry. Mining significant topics from short and noisy reviews is big challenge. In light of this, such problem is addressed by proposing the probabilistic aspect mining model (PAMM) for identifying the aspects/topics relating to class labels. Because of unique feature of PAMM it focuses on finding aspects relating to one class only rather than finding aspects for all classes simultaneously in each execution. Besides the aspects found also have the property that they are class distinguishing, that means they can be used to distinguish a class from other classes. It helps to reduce the chance of having aspects formed from mixing concepts of different classes; hence the identified aspects are easier to be interpreted by people.***

Keywords: Drug review, opinion mining, aspect mining, text mining, topic modeling.

1. Introduction

WITH the advent of Web 2.0 [1], [2], people get encouraged to contribute their contents to the Internet. For sharing the information and user interaction many user-centered platforms are now available, which include Epinion, Amazon, Facebook and Twitter. Nowadays when people are interested in a product or a service, they usually not only look for official information from product manufacturers or service providers but also check for practical opinions from the customer's who have used that. As a result, online reviews, blogs and forums dedicated for different kinds of products are pervasive, and how to effectively analyze and exploit such immense online information source is a challenge.

Opinion mining (or sentiment analysis) [3]–[6] mainly consist of extraction of specified information (e.g., positive or negative sentiments of a product) from a large amount of text opinions or reviews authored by Internet users. In many situations, an overall rating for a review cannot reflect the conditions of different features of a product or a service. For instance, a camera may come with excellent image quality but poor battery life. As a result, more sophisticated aspect level opinion mining approaches have been proposed to extract and group aspects of a product or service and predict their sentiments or ratings [3], [7]–[09]. Recent state-of-the-art approaches such as frequency-based approach [5], relation-based approach [8], [10], supervised learning [11] and topic modeling [7], [12] showed that favorable results could be obtained.

Previous studies of opinion mining generally deal with popular consumer products or services such as digital cameras, books, electronic gadgets, etc. but Entities of medical domain are of far less concerned. One reason may be because patients are minority groups on the Internet and they are only concerned with particular illnesses or drugs that they are experiencing. Furthermore, people tend to solicit opinions from medical professionals rather than patients. Nevertheless, recent studies have shown that patient generated contents 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 alleviate symptoms and side effects of drugs. These online communities were found to have positive impacts on patient health.

2. Basic Definitions

Definition of Opinion
An Opinion is a belief or judgment of a large number or majority of people formed about a particular thing, not necessarily based on fact or knowledge. In general, opinion refers to what a person thinks about something

Opinion holder: it is the person that gives a specific opinion on an object.

Object: it is entity on which an opinion is expressed by user.

Document, Topic and Sentiment
A Document D is a piece of text in natural language. We assume that each document discusses at least one topic, and not all topics discussed in the same document have to be related to each other. Topic T is a named entity, event or abstract concept that is mentioned in a document D and a Sentiment S is the author's attitude, opinion or emotion expressed on topic T.

Opinion mining or Sentiment analysis
Opinion mining is a technique to detect and extract subjective information in text documents. In general, sentiment analysis tries to determine the sentiment of a writer about some aspect or the overall contextual polarity of

a document. The sentiment may be his or her judgment, mood or evaluation. A key problem in this area is sentiment classification, where a document is labeled as a positive or negative evaluation of a target object (film, book, product etc.)

3. Literature Survey

Yao Wu and Martin Ester [13] introduced A Probabilistic Model Combining Aspect Based Opinion Mining and Collaborative Filtering In this paper, we study the problem of estimating personalized sentiment polarities on different aspects of the items. Unified probabilistic model called Factorized Latent Aspect Model (FLAME) is proposed to solve this problem which combines the advantages of collaborative filtering and aspect based opinion mining. FLAME learns users' personalized preferences on different aspects from their past reviews, and predicts users' aspect ratings on new items by collective intelligence.

Wei Jin , Hung Hay Ho and Rohini K. Srihari [14] suggest OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction.

The main aim of OpinionMiner system designed in this work is to mine customer reviews of a product and extract high specified product entities on which reviewers express their opinions. Opinion expressions are determined and opinion orientations for each recognized product entity are classified as positive or negative. This Approach is different from previous one that employed rule-based or statistical techniques; we introduced a novel machine learning approach which is built under the framework of lexicalized HMMs i.e. Hidden Markov Model. The approach naturally integrates multiple important linguistic features into automatic learning. In this paper, we describe the architecture and main components of the system.

Victor Cheng, Chao Tang and Chun-hung Li [15] introduced Drug Review Mining with Regressional Probabilistic Principal Component Analysis. In this paper, problem of mining significant topics from short and noisy reviews is addressed by proposing the Regressional probabilistic principal component analysis (RPPCA) to correlate the sentiment values of the review while simultaneously optimizing the probabilistic generative process of words into reviews. Besides the classification of sentiment in reviews, a sentiment word identified by RPPCA allows the delineation of the core aspects in taking such medications from the patients' perspectives.

Aurelie Neveol and Zhiyong Lu [16] developed a model for Automatic Integration of Drug Indications from Multiple Health Resources. Most drug indication information is only available in free text as opposed to structured format, thus making it difficult for further automatic analysis by computers. In response, herein focus on automatically extracting and integrating drug indication information from multiple resources such as DailyMed and MeSH Scope notes. Then select trustworthy resources of drug/disease relationships and apply state-of-the-art relationship extraction methods, customized to improve recall and perform ellipsis and anaphora resolution.

4. System Architecture

The block diagram of Aspect/feature based sentiment analysis model is given in the Following figure.

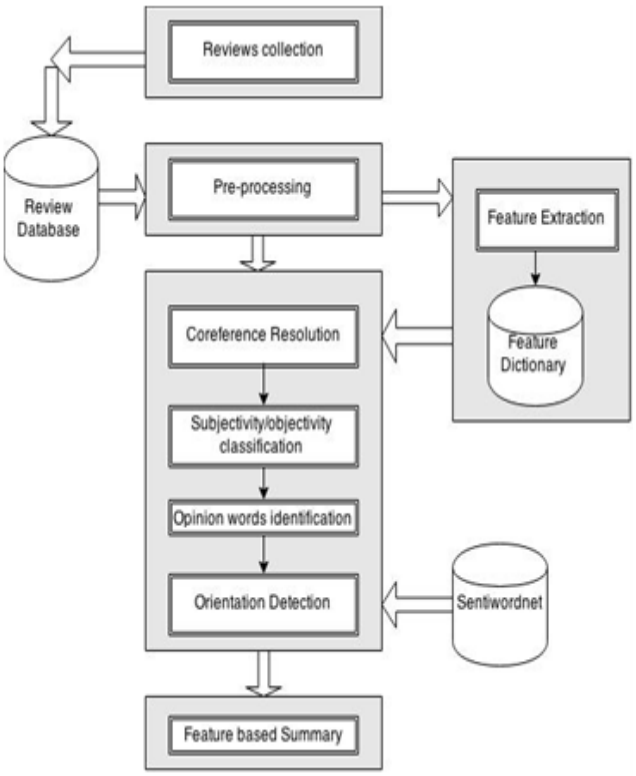

Figure: Block diagram of Feature based Sentiment Analysis Model

Reviews Collection
Reviews about drugs from online website WebMD are collected as the data-set for sentiment analysis. Role of reviews collection module is to download the opinions and reviews about drugs from the specified URL. Then these collected reviews are stored in the database.

Pre-Processing
The second step of the technique involves preprocessing or filtering of reviews, which improve the accuracy and also avoid the unnecessary processing overhead of opinion mining process. The pre-processing steps include stop words removal. Non alphabetic characters like numbers and symbols and smiley's are removed before sentiment analysis. This can increase the speed of the opinion mining process.

Feature Extraction
The aspects or features of a product can appear as a single word or a phrase. For example, picture quality of a camera is one among its features while size is another feature. In this, we have a feature dictionary which is domain specific. We manually add the known features of a drug that is to be needed for the creation of sentiment profile. These features are also stored in the database for further use.

Co-reference Resolution
Co-reference Resolution is the task of identifying the mentions to entities that they refer to. The Stanford Deterministic Co-reference Resolution System is used for

resolving all noun phrases that refer to the same entities. For instance, consider the two sentences given below. Picture quality of the camera is very good. It is amazing. We could not relate the aspect used in second sentence with the aspect in first sentence without using co-reference resolution. The co-reference resolver produces an output that "Picture quality" in first sentence and "It" in second sentence as it is co-referred. So we replace the pronouns that got resolved, with the corresponding nouns. This replacement is limited to the pronouns that got resolved to aspect names of the product.

Subjectivity/Objectivity Classification

All the sentences in the reviews do not contain an opinion. A sentence of the review is analyzed only when it contains an opinion. Such sentences are called subjective sentences and non- opinionated sentences are called objective sentences. The subjective sentences should be identified and other (objective) sentences should be removed before the analysis. It helps for avoiding the further processing overhead. This is done using feature dictionary containing feature words. If the sentences taken for this purpose contain the feature words in the feature
dictionary, then these sentences are taken as subjective.

Opinion Words Identification

Opinion words are usually adjectives, adverbs, and verbs which expresses the positive or negative polarity of a feature of a product.. By the incorporation of Stanford dependency parser, we could get dependency relations of opinion words related to a particular feature in feature dictionary. This parser outputs the adverb, verb and adjectives having dependency relations with the aspect in a sentence. These opinion words are further used for calculating the polarity of different features of the product in reviews.

Orientation Detection

In Orientation Detection, positive or negative score of every feature in the reviews are determined. Reviews contain one or more sentences. First we calculate the sentence level score of the opinion by analyzing each sentence in the reviews. By combining all the sentences in a review and calculate the overall score of features in reviews. Then calculate the overall score of each feature by combining all reviews. For calculating the opinion score of each feature, we have used the algorithm for determining the score of adverb adjective and adverb verb combinations in sentences.

Existing System

In existing system given corpus of reviews (every review is in bag of words format), words highly correlated with the class label can be identified by various approaches such as association rule, information gain, pointwise mutual information(PMI).These approaches unfortunately suffer from sever problem that is nothing but the difficulty in understanding the underlying aspects or concept from just set of words correlated with the class label. There is no intuitive algorithm to group the words so that each group conveys one or few easily understandable concepts.

Aspects correlated to different class labels are found simultaneously. This formulation identifies aspects having mixed contents from different classes. Existing system extracts all the aspects and their sentiments from the reviews but we want only relevant aspects.

Proposed System

We propose a probabilistic model for finding the aspects correlated to class labels. The work differs from other previous approaches, however, in that each time the model focuses on finding aspects correlated to one class label only. Aspects correlated to different class labels are found separately. This formulation avoids the identified aspects having mixed contents from different classes. By focusing the task on one class, better and more specific aspects can be found. This approach is also different from the intuitive approach of which reviews are first grouped according to their class labels and followed by inferring aspects for the individual groups.

The proposed model uses all the reviews and find the aspects that are specific to the target class and are helpful in differentiating reviews of different classes. For the intuitive approach, the identified aspects may not be only related to the contents of individual groups. They may be common to all the classes and not useful. For example, the dosages of a drug can be a common aspect to all the classes but it may not be useful in differentiating classes.

PAMM Model

PAMM is a generative model which generates the observed data $\mathbf{x} \in R^M$ and the class label
$y \in \{0, 1\}$ from the Gaussian latent variable $\mathbf{z} = (z1, \ldots, zk)^T$ (i.e. $\mathbf{z} \in R^K$) with zero mean and identity covariance matrix, i.e. $\mathbf{z} \sim N(0, \mathbf{I})$. Data points and the associated class labels are generated as follows:-

1) Draw $z \sim N(0, \mathbf{I})$

2) Draw $x \sim N(Wz + \mu, \sigma^2\mathbf{I})$

3) Draw $y \sim (\rho(y = 0 \mid z), \rho(y = 1 \mid z))$

Where μ is the mean of the observed data, $\sigma 2$ is the Gaussian noise level on $\mathbf{x}$, $\mathbf{W} \in R^{M*K}$ is a matrix having non-negative entries, $\rho(y = 1 \mid z)$ and $(\rho(y = 0 \mid z)$ are given by
$\rho(y = 0 \mid z) = 1 - \rho(y = 1 \mid z)$,
$\rho(y = 1 \mid z) = \phi(V^T z) = \phi(c \sum_{i=1}^{k} zi)$,
$\phi(t) = 1/\ 1 + e^{-t}$

Where, ϕ is logistic function and C is constant The label y is binary and drawn from the Bernoulli distribution with probabilities $\rho(y = 1 \mid z)$ and $(\rho(y = 0 \mid z)$. The aspects of the model can be obtained from $\mathbf{W}$ as it can be regarded as the basis of generating the observed data.

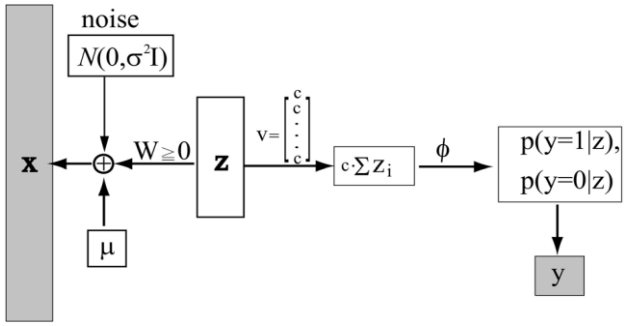

Figure : PAMM for generating observed data x and label y from latent variable z.

Efficiency of Probabilistic Aspect Mining Model:

1) PAMM has a unique feature in that it focuses on finding aspects relating to one class only rather than finding aspects for all classes simultaneously in each execution.
2) Reduces the chance of having aspects formed from mixing concepts of different classes. hence the identified aspects are easier to be interpreted by people.
3) The aspects found also have the property that they are class distinguishing , can be used to distinguish a class from other classes.
4) By focusing the task on one class, better and more specific aspects can be found.
5) By removing the fake reviews also we are going to increase the efficiency of model.

Evaluation Metrics

1) Mean PMI(Pointwise Mutual Information)

PMI is a measure of association between a feature (in this case aspect or word) and a class (i.e. label) .Measure of how much one word tells us about the other. How much information we gain.

Given by pmi (x;y) = log *p(x,y) / p(x)p(y)*

p(x):-Probability of data point x
p(y):- Probability of label y

2) Accuracy

Accuracy is the measure of degree to which the result of mining and classification conforms to the correct value or a standard.

5. Conclusion

The proposed probabilistic aspect mining model (PAMM) which is used for mining of aspects relating to specified labels or groupings of drug reviews is more accurate comparing with other supervised topic modeling algorithms, and This model has a uncommon feature in that it focuses on finding aspects relating to one class only rather than finding aspects for all classes simultaneously in each execution.

This unique feature reduces the chance of having aspects formed from mixing concepts of different classes; hence the identified aspects are easier to be interpreted by people. The aspects found also have the property that they are class distinguishing. They can be used to distinguish a class from other classes. We can apply this model to find aspects relating to different segmentation of data such as different age groups or other attributes.

References

[1] T. O'Reilly, "What is web2.0: Design patterns and business models for the next generation of software," Univ. Munich, Germany, Tech. Rep. 4578, 2007.
[2] D. Giustini, "How web 2.0 is changing medicine," *BMJ*, vol. 333, no. 7582, pp. 1283–1284, 2006.
[3] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. KDD*,Washington, DC, USA, 2004, pp. 168–177.
[4] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Ret.*, vol. 2, no. 1–2, pp. 1–135, Jan. 2008.
[5] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Proc. Conf. Human Lang. Technol. Emp. Meth. NLP*, Stroudsburg, PA, USA, 2005, pp. 339–
[6] L. Zhuang, F. Jing, and X. Zhu, "Movie review mining and summarization," in *Proc. 15th ACM CIKM*, New York, NY, USA, 2006, pp. 43–50.
[7] Q. Mei, X. Ling,M.Wondra,H. Su, and C. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs," in *Proc. 16th Int. Conf. WWW*, New York, NY, USA, 2007, pp. 171–180.
[8] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comaring opinions on the web," in *Proc. 14th Int. Conf. WWW*, New York, NY, USA, 2005, pp. 342–351.
[9] I. Titov and R. McDonald, "A joint model of text and aspect ratings for sentiment summarization," in *Proc. 46th Annu. Meeting ACL*, 2008, pp. 308–316.
[10] S. Baccianella, A. Esuli, and F. Sebastiani, "Multi-facet rating of product reviews," in *Proc. 31st ECIR* , Berlin,, Germany, 2009, pp. 461–472.
[11] W. Jin, H. Ho, and R. Srihari, "Opinionminer: A novel machine learning system for web opinion mining and extraction,", New York, NY, USA, 2009, pp. 1195–1204.
[12] 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. 815–824.
[13] Yao Wu and Martin Ester "FLAME: A Probabilistic Model Combining Aspect Based Opinion Mining and Collaborative Filtering" University Burnaby, BC, Canada.
[14] Wei Jin , Hung Hay Ho and Rohini K. Srihari" OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction" Department of Computer Science North Dakota State University Fargo, ND 58108, June 28–July 1, 2009, Paris, France.
[15] Victor Cheng, Chao Tang and Chun-hung Li "Drug Review Mining with Regressional Probabilistic Principal Component Analysis" Computer Science Department Hong Kong Baptist University, *HI-KDD'12* August 12, 2012, Beijing, China.
[16] Aurelie Neveol and Zhiyong Lu" Automatic Integration of Drug Indications from Multiple Health