Volume 5, Issue 5, May 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Automatic Integration of Drug Indications from Multiple Health Resources

Satchinathan P M.C.A., **Dr.T.Stephen Thangaraj** Ph.D, **R. Nandakumar** M.Tech.
Computer Science and Engineering, T.J. Institute of Technology
India

Abstract— Recent findings show that online reviews, blogs, and discussion forums on chronic diseases and drugs are becoming important supporting resources for patients. Extracting information from these substantial bodies of texts is useful and challenging. We developed a generative probabilistic aspect mining model (PAMM) for identifying the aspects/topics relating to class labels or categorical meta-information of a corpus. Unlike many other unsupervised approaches or supervised approaches, 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. This 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. An efficient EM-algorithm is developed for parameter estimation. Experimental results on reviews of four different drugs show that PAMM is able to find better aspects than other common approaches, when measured with mean point-wise mutual information and classification accuracy. In addition, the derived aspects were also assessed by humans based on different specified perspectives, and PAMM was found to be rated highest.

Keywords— Social Media, Data Mining, Gig Data, Illicit Drug Use, Map Reduce, Feature Selection, Drug Review, Opinion Mining, Aspect Mining, Text Mining, Topic Modelling

I. INTRODUCTION

Opinion mining (or sentiment analysis) deals with the 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, solely 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. Previous studies of opinion mining usually deal with popular consumer products or services such as digital cameras, books, electronic gadgets, etc. Entities of medical domain are of far less concerned. It may be because patients are minority groups on the Internet and they are only concerned with specific illnesses or drugs that they are experiencing. Furthermore, people tend to solicit opinions from medical professionals rather than patients.

Unlike general products or services, drugs have a very limited number of kinds of aspects: price, ease of use, dosages, effectiveness, side effects and people's experiences. There are other more technical aspects such as chemical or molecular aspects, but they are almost not mentioned in drug reviews. A difficulty in dealing with drug reviews is that the wording in describing effectiveness, side effects and people's experiences are very diverse. In particular, side effects are drug dependent: a set of side effect symptoms for a drug is very unlikely applicable to another drug.

In this paper, we address the opinion mining problems for drug reviews. As many drug review websites are equipped with rating functions, prediction of sentiments is not the task. Instead a model for identifying a set of aspects relating to class labels or meta-information of drug reviews is proposed. For example, if the reviews are associated gender information, people may be interested in studying the aspect difference between female patients and male patients.

This task is different from general aspect-based opinion mining in which the task aims to extract all aspects and their sentiments from reviews. Referring to the problem definition, not all the aspects but only relevant aspects need to be extracted. Sometimes, an aspect may need to be segmented further (in finer granularity) because only limited components of it are required. For instance, considering the aspect of side effects of a drug, male patients may be anxious about a specific side effect while other side effects are of less concerned.

We propose a novel probabilistic aspect mining model (PAMM) to mine the aspects of drug reviews correlated with categorical information. This can be regarded as a topic model with the derived topics treated as aspects.

II. RELATED WORK

Aspect-based opinion mining is becoming popular in recent years. Frequency based approach [5] extracts high frequency noun phrases which meet the specified criteria or constraints from the reviews as aspects. On the other hand, relation based approach [9], [12] identifies aspects based on the aspect-sentiment relation in the reviews. These two kinds

of approaches, however, may not be applicable to drug reviews as aspects are often not indicated explicitly by authors and descriptions of side effects and people's experiences is diverse. Moreover, grouping of the extracted noun phrases is another challenge as they cannot be grouped just based on semantic meanings. In contrast, topic modelling identifies aspects based on the co-occurrence of words in reviews. It has an advantage that aspect identification and grouping are performed simultaneously. Topic modelling (e.g., LDA) is a popular probabilistic approach in understanding a corpus. With this approach, a set of topics, which are represented by multinomial distributions over vocabulary words, are inferred. When the words of a topic are sorted according to the probabilities, high probability words of a topic are usually semantically correlated and the concept or aspect of the topic can be captured manually. For example, Topic Sentiment Mixture (TSM), Joint Sentiment/Topic (JST) model and Aspect and Sentiment Unification Model (ASUM) were proposed to extract both the aspects and predict their associated sentiments. Nevertheless, these aspectbased opinion mining methods may not be appropriate to address the problem defined in the previous section as the extracted aspects may not be related to the specified class labels and the performance depends on the manual selection of seed words. Recently, topic modelling with supervised label information has become an interest of research. Blei and McAuliffe proposed the supervised LDA (sLDA) that can take care of different forms of supervised information during topic inference. Mimno and McCallum introduced Dirichlet-multinomial regression to handle different kinds of meta information. Ramage et al. proposed DiscLDA to process discriminative information and find.

III. PROBABILISTIC ASPECT MINING MODEL

Probabilistic Aspect Mining Model (PAMM) is a generative model which generates the observed data x ? RM and the class label y ? {0, 1} from the Gaussian latent variable z = (z1, . . . , zK)T (i.e. z ? RK) with zero mean and identity covariance matrix, i.e. z ~ N(0, I). Fig. 1 describes the data and label generation process. Referring to the figure, data points and the associated class labels are generated as follows.

1) Draw z ~ N(0, I);

2) Draw x ~ N(Wz + μ, s2I);

3) Draw y ~ (p(y = 0|z), p(y = 1|z)),

where μ is the mean of the observed data, s2 is the Gaussian noise level on x, W ? RM×K + is a matrix having non-negative entries, p(y = 1|z) and p(y = 0|z) are given by

p(y = 0|z) = 1 - p(y = 1|z) , (1)

p(y = 1|z) = f(vTz) = f

f(t) = 1

1 + e-t , (3)

Where f is a logistic function and c is a constant. The label y is binary and drawn from the Bernoulli distribution with probabilities p(y = 1|z) and p(y = 0|z). The aspects of the model can be obtained from W as it can be regarded as the basis of generating the observed data. By inspecting high probability/value words of individual columns of W, the underlying concepts of the aspects can be interpreted.

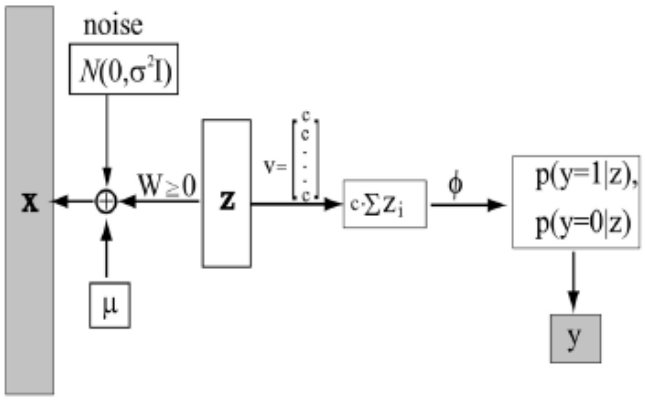

Fig. 1. PAMM for generating observed data x and label y from latent variable z.

IV. EVALUATION USING CLASSIFICATION ACCURACY

The reviews of each drug were divided into training data and test data: 80% of reviews were randomly drawn to form the training data and the rest 20% reviews were the held-out test data. The training data were used to derive the aspects of the drugs. As previously, only 20 words with top probabilities/values were preserved for each derived aspect. Then a subspace was formed from the aspects and the evaluation of classification accuracy was performed by projecting both the training data and test data into the subspace.Assuming there are K aspects, the k-th basis vector **w**k ∈ RM of the subspace is formed from the k-th aspect: the component i of **w**k (i.e., wk,i) is set to 1.0 if the associated word appears in the k-th aspect, and is set to 0.0 otherwise.

The least square projection matrix **P** ∈ R*K*×*M* is given by

$\mathbf{P} = (\mathbf{WTW} + \lambda \mathbf{I})^{-1}\mathbf{WT}$, (23)

where $\mathbf{W} = [\mathbf{w1}, \ldots, \mathbf{wK}]$ and λ is a regularization constant for avoiding the matrix inverse computation problem when **WTW** is singular or near singular. In the experiments, $\lambda = 0.01$. Let $\mathbf{x}i$ denote the i-th review, the projected vector $\mathbf{z}i \in R^K$ is given by

$\mathbf{z}i = \mathbf{P}\mathbf{x}i$.

The steps for preforming the above described projection can be summarized as follows:

1) For each aspect, preserve only the 20 words with highest Probabilities/values and set their values to 1.0. Values of other words are set to 0.0.
2) Form the matrix W with columns having the values obtained from step 1.
3) Form the projection matrix P by using (23).
4) Project both the training reviews and test reviews with P and learn a SVM with the projected training reviews to classify the projected test reviews.

After projecting both the training data and test data, a support vector machine (SVM) with linear kernel was used to classify the test data. The above process was repeated with 5-fold cross validation and the mean classification accuracy is shown in Table 3. Furthermore, classification accuracy (SVM_input) computed by directly applying SVM on the input space (i.e. the bag-of-words representation) of the reviews is also given in the table for reference. Referring to Table 3, it is clear that the aspects derived from the supervised algorithms perform better than the unsupervised algorithms. NMF is marginally better than LDA and SSNMF performs closely with DiscLDA with the former give better results in more cases. PAMM gives the best accuracy in all cases. Comparing with the SVM_input, all models seem inferior, except for the drug simvastatin processed with PAMM. This is reasonable because only 20 distinct words, from each aspect, were used to evaluate the projection matrix P while all the words of the vocabulary were used to learn the SVM to give the SVM_input results. Moreover, SVM is a classifier dedicated for classification and not for deriving aspects. For testing the statistically significance of the results, the Wilcoxon signed-rank test was used rather than the paired t-test because the latter needs the normal distribution assumption and has serious weaknesses mentioned in. From Table 3 there is four drug datasets and each is tested with five different values of K, thus this gives a total of 20 tests. Since PAMM has the highest accuracy in all the cases, the value of Wilcoxon test statistic is 0 when PAMM compared with other algorithms. Hence, PAMM significantly performs better than others even with 0.01 significance level.

V. SYSTEM IMPLEMENTATION

Comparing with other supervised topic modelling algorithms, PAMM has a unique feature that it focuses on deriving aspects for one class only. This feature reduces the opportunities of forming aspects from reviews of different classes and hence the derived aspects are easier for people to interpret. The focus areas include the following:

I. Drug Information
II. Allergic Information
III. Drug Trial Information
IV. Individual Trial Information
V. Drug Trial History
VI. Individual Trial Information

VI. DISCUSSION AND CONCLUSION

Nowadays, online reviews, blogs and discussion forums for different kinds of products and services are pervasive. Patient experiences and concerns are often insufficiently represented despite the abundance of reviews of medication from patients on the internet. Extracting information from these substantial bodies of texts is useful and challenging. In particular, it is helpful to identify the aspects of a product that people are happy to with or finding the aspects that may anger customers. As human lifespan becomes longer and our living environment becomes increasingly polluted, medical domain data mining becomes one of the focused research areas. In this paper, we propose PAMM for mining aspects relating to specified labels or groupings of drug reviews. Opportunities of forming aspects from reviews of different classes and hence the derived aspects are easier for people to interpret. Unlike the intuitive approach in which reviews are first grouped according to their classes and followed by inferring aspects for individual groups, PAMM uses all the reviews and finds the aspects that are helpful in identifying the target class. Apart from the quantitative assessments, the aspects were assessed by a group of people based on four different perspectives and PAMM obtained the highest score. The model was also applied to finding those aspects relating to the genders of patients. Its performance advantage over other approaches is more prominent as very specific aspects are discovered. Parameter estimation of PAMM is not complex as only one matrix needs to be estimated from the training data. On the other hand, clinical trials are costly and very time consuming. It usually takes a few years or even over a decade to finish. Their sample sizes are usually not large enough to give significant conclusions. Thus, studying of patient reviews provides a value reference from the patient's points of view. Algorithm design and efficiency analysis become more important when one studies how to efficiently mine all possible rare event sets and association rules based on minimal support. To different segmentation of data such as different age groups or other attributes. It is also useful to work with aspect interpretation as aspects are now represented by a list of keywords. If a few sentences can be extracted or generated automatically to summarize the keywords, interpretation and understanding will be greatly improved.

REFERENCES

[1] O'Reilly, "What is web2.0: Design patterns and business models for the next generation of software," Univ. unich, 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'l Conf. Knowledge Discovery and Data Mining, 2004, pp. 168–177.

[4] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, pp. 1–135,2008

[5] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in Proc. Conf. on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 339–346.

[6] L. Zhuang, F. Jing, and X. Zhu, "Movie review mining and summarization," in Proc. 15th ACM Int'l Conf. on Information and Knowledge Management, 2006, pp. 43–50.

[7] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: modelling facets and opinions in weblogs," in Proc. 16th Int'l Conf. on World Wide Web, 2007, pp. 171–180.

[8] S. Moghaddam and M. Ester, "Aspect-based opinion mining from online reviews," Tutorial at the 35th Int'l ACM SIGIR Conf., 2012.

[9] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analysing and comparing opinions on the web," in Proc. 14th Int'l Conf. on World Wide Web, 2005, pp. 342–351.

[10] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in Proc. 18th ACM Conf. on Information and Knowledge Management, 2009, pp. 375–384.

[11] E. Cambria, A. Hussain, and C. Havasi. Towards the crowd validation of the british national health service. 2010.

[12] L. Xia, A. Gentile, J. Munro, and J. Iria. Improving patient opinion mining through multi-step classification. In Text, Speech and Dialogue, pages 70{76. Springer, 2009.

