ISSN:0975-9646

Extraction of Drug Reviews by Specific Aspects for Sentimental Analysis

Vrushali Moon[#1], Prof. Prashant Borkar[*2]

#Department of computer science,
G.H. Raisoni College of Engineering,
Nagpur University, Nagpur, India

***Abstract*- Nowadays medical data has been search by people for gathering the more and more information about drugs. The dialog discussions on chronic diseases and drug, as well as online audits and sites are getting to be more important assets for patients. Patients read online reviews, blogs and discussion forum ideas to have knowledge from the other patients with identical condition. Reviews about medicine which is obtaining from the patients are largely available at internet. Partitioning data from these significant collections of writings is helpful in testing. Extracting these huge medical data is challenging. In this research, drugs reviews are extracted by the specific aspect such as age and gender and then summarize all the reviews. For reviews extraction and summarization, various techniques are also discus in this paper.**

Keywords: **Text mining, opinion mining, aspect mining, sentiment analysis, opinion orientation.**

I. INTRODUCTION

In this world, people are connected to each other and they share their experience across over an internet. People shows enthusiasm in the conventional information about the product as well as services provided by the internet is also facilitating. Now internets become important part of day to day life through which people access information and also interact with each other. People share their experience about product, service about particular association through online reviews, mini blogs and forums, so that they can analyze various kinds of domain and aspects [1]. These reviews and blogs now also seen in the medical field where person with medical problems share their opinion which is advantageous for lifelong disease as well as medicine with affecting side effects [2]. With tons of information regarding health available, patients know what kind of or which drugs they are taking and how helpful it would be to them [3].

Patients write their experience and opinion about particular drug and also the review about drugs which make strong influence on patient's purchasing decision. Some well-known drugs get hundreds of reviews at some website and blogs [4]. Some reviews have long message with very limited useful information regarding drug [5]. It is very problematic for the people to read and to make a decision on drugs effectiveness. People have diverse experience in terms of effectiveness, side effect, dosages.etc, which also affect the drug reviews analysis [6]. On the online blogs, patients just express the symptoms, feelings and comments. They do not illustrate which aspect they are describing [7]. For drug reviews analysis, number of aspect needs to be determined which are correlated not only with categorical information but also sentiment classifications which are further used to reveal two types of orientation text i.e. positives and negative ones and it also should provide satisfactory result[8].

Opinion mining (or sentiment analysis) extracts large amount of texts containing different opinions and reviews which are authored by various Internet users to get the specified information (positive or negative response of a product).

Previous studies of opinion mining [9] focus mainly on famous customer services and product such as digital camcorder, electronic appliances, books, etc. but subsistence of medical region are not much considered. One reason for this, only some group on patients interested in sharing their experiences and those are limited to particular illnesses or drug. Although, that patients originated contents are helpful for other people and important also [10]. Dealing with the reviews on the aspect of side effect is very difficult because side effects are dependent on number of aspects because symptoms for one drug may not be the same for another .

In the proposed paper, drug review system solves the problem related to aspect for review analyzing will be focused. Here aspect broadly considered as age and gender of the commenter's. In the drug review system, a set of patient's reviews of a some drugs are considered, in which the major task is associated with three subtasks: (1) extract that reviews age wise and gender wise that have been commented on by patient's and identification of opinion words in each review. (2)After opinion word identification, the next task is to identify the orientation for deciding the each review sentence is positive or negative. This is depending on the opinion words present in the reviews. (3) The last subtask displays the summarized review result to the user.

The layout of the paper is as follows, section II covers the literature survey, while section III proposed work and section IV contain experimental evaluation and lastly provides conclusion and references.

II. LITERATURE SURVEY

A. Online reviews mining:

Review mining indicates the process of that drive the product features and opinions from the subjective contents and estimate the user's opinions, sentiment, and comments, etc. Boulos and Wheeler [11], gives lots of social software systems and also define their importance in health care. The online blogs and social networking facilities are also provided. It gives the growing spectrum of application. Collier et al. [12] introduce a setup for text mining to finds

out sudden changes in occurs in the infectious diseases and keep track of that outbreak using webpage documents reported by Rich Site Summary (RSS) feeds. Steinberger et al. [13], proposed a system which is quickly accessible and keep the track of disease database by monitoring textual report. The important parts of the system are Medical Information System (MedISys) as well as Pattern base Understanding and Learning System (PULS). Jingiing Liu and Stephanie Senelf [14], derived the latent information and give a multimodal interface for review scanning and inquiring. In this paper, spoken dialogue system involves three methods: Speech processing, Language understanding and dialogue management. Carolin Kaiser, Freimut Bodendorf [15], find the online review data to diagnose the strength as well as the weakness of the drugs. The Paper describes main three methods: text mining, aggregation method, and last data mining algorithm.

B. Summarization:

Y. Jo and A. Oh [16], focus on the reviews which are written in plain text and extract the information about the relevant data. Automatically discovering of aspect problem is deal in this paper. This paper shows the sequence of SLDA and ASUM. Chenghua Lin, Yulan He, Richard Everson[17], develop the data which works on opinion search and automated sentiment analysis to find out the hidden information on unstructured text data. A novel probable modeling structure based on Latent Dirichlet Allocation (LDA) also kwon as Joint Sentiment-Topic (JST), Sentiment. Simultaneously prior weakly supervised is performed. Julien, F. Sha, and M.Jordan[18], for dimensionality reduction image document and text, describe the probabilistic method which is more important. This is done for the data analysis. This paper focuses on the different technique used for discovery of lessen dimensionality representation on a discriminative formation. The output of this paper is a complex model which can be trained with the unsupervised method. W. Jin, H. Ho, and R. Srihari [19] use the web opinion mining and extraction tools for mining the customer reviews about the product because collecting all reviews of cutomer's gets difficult for a particular item. Victor C. Cheng, Leung, Jiming Fellow, [20], online drugs reviews are listed and opinion mining concentrate on polarity classification. The Author introduced Regression Probabilistic Principal Component Analysis (RPPCA) which examines medical data used for document analysis and review sentiment values. The advantage is Sentiment words can be recognized by RPPCA and medication is given by patient viewpoint.

C. Sentiment analysis:

Sentiment analysis use different ways to analyze and extract useful information from database such as the natural language processing, text analysis as well as computational linguistic. Sentiment analysis classifies the polarity of the text present at the document level or at the aspect level. The two main type of sentiment analysis: subjective/objective identification as well as feature /aspect based sentimental analysis [21].

a. *Subjective/Objective Identification:*

In subjective/objective identification, given text is classified into two classes: subjective and objective where subjectivity of the word depends on their context while the objective document may consist of the subjective sentence. The target of Subjective/objective identification is to determine whether a language unit express a private state or opinion and what will be the polarity i.e. positive or negative[22].K. Sarvabhotla, P. Pingali, and V. Varma[23], propose the statistical methodology called review summary (RSUMM) for identification of subjective in the given text for sentiment classification. S. Zhou, Q. Chen, and X. Wang[24], proposed the semi-supervised machine learning sentiment classification method which is based on active deep network (ADN. S. Li, C. Huang, G. Zhou, and S. Y. M. Lee[25], handle sentiment classification depend on a co-training framework by linking two views "personal and impersonal views". The author uses unsupervised mining for personal and impersonal views extraction and after that this views are used in supervised and semi-supervised sentiment classification which improves the model. S. Li and J. Hao [26], propose the algorithm which is based on spectral clustering. This algorithm improves the accuracy of sentiment classification. The method proposed in this paper shows best results when compared with the method Self-learning sentiment classification.

b. *Featured/Aspect Based Sentiment Analysis:*

A.-M. Popescu and O. Etzioni [27], introduced OPINE which is an unsupervised information retrieval system. OPINE system mines review to develop a model of important product features and use the web as a corpus to find out the features of the product with advance precision. Semantic orientation of the possible opinion words is find out by using relaxation labeling technique. C. Li Zhuang, Feng Jing [28] presented a paper which gives the idea of the Web and online audit which is converted into a helpful as well as critical data asset for individuals. In this paper, the author focuses on a peculiar space film audit. Lin and Y. His [29] main focus is on fast text data and text mining growth which is in turn help full for recognized hidden knowledge from more than one domain where customer's sentiment and opinions are conveyed in form of a free text for the companies, however out of which the big amount of textual data is needed to find out the applications. This paper focuses on the document level sentiment classification. Yao Wu and Martin Ester [30] proposed A Probabilistic Model linking with Collaborative Filtering and Aspect-Based Opinion Mining. The author, discuss the issues which is occurs in personalized sentiment polarities evaluation perform on numerous aspects of the items. Factorized Latent Aspect Model (FLAME) which is also known as unified probabilistic model is introduced to resolve this problem.

III. THE PROPOSED WORK

The drug review system shown in figure 1, dived the task into three major subtasks: (1) extract that reviews age wise and gender wise that have been commented on by patients and identification of opinion words in each review. (2)After opinion word identification, the next task

is to detect the orientation of the opinion word for deciding the positive sentence or negative sentence of the reviews. (3) Producing a summary.

A. *Review Collection:*

The drug reviews are collected from the website WebMd.com which is considered as a database for sentimental analysis. For review collection web scrapping technique is used which direct scrape the whole web page of the website using URL http://www.webmd.com. This database is unstructured which can be converted into structured data using a regular expression (regex). Regex used to modify the text by defining the search pattern.

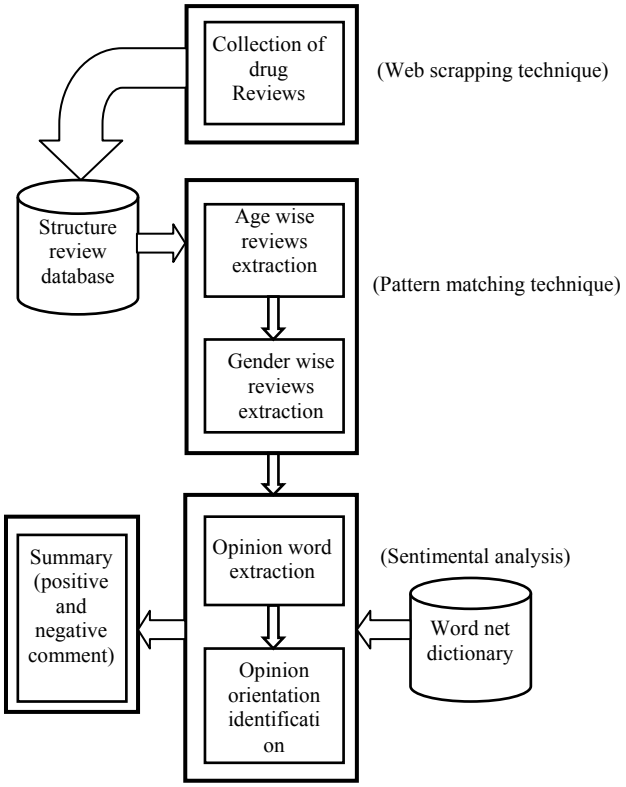

Fig1. Flow of drug review system

B. *Age wise and gender wise extraction of reviews.*

The reviews collected by the web scrapping from online website WebMd contain the age and gender of the reviewer which is given the in the review. Using this information about age and gender, reviews are extracted by using pattern matching technique.

All reviews about drugs read by the drug review system and then firstly extract on the basis of age. Then for each age group, reviews are extracted gender wise. Exact pattern matching algorithm is used for this process.

C. Sentimental Analysis

The sentimental classification involves opinion word identification and orientation identification of that opinion word. The Orientation of opinion word determines the positive and negative impact of the reviews. The opinion orientation of word will be decided by performing two subtasks. The first task is to identify opinion words from the reviews. These opinion words are adjectives. The second task is to identify the orientation of that opinion word e.g. positive or negative.

The wordnet dictionary is used for the sentimental analyses which possess the list of adjectives. This adjective contains the synonym and antonym set which is used for semantic orientation of the opinion word. Adjectives share the equal orientation for their synonyms and opposite orientations for antonyms. Finally summarized result will be generated on the basis of sentimental analysis.

IV. EXPERIMENTAL EVALUATION

A system, which is name as drug review system has been implemented in dot net. In the experiment, patient's reviews about three drugs: Abilify, Carisoprodol and Bactrim DS are taken into consideration from webmd.com. Each review contains age, and the gender of the commenter as well as text reviews. For each drug, first crawled and download the 4-5 reviews. These reviews save in the database were then clean and remove all the HTML tag using a regular expression. After that, pattern matching technique is used to extract reviews age wise and gender wise. Finally, drug review system implements the sentimental analysis for final result generation.

For the interpretation, all reviews are read manually by the expert. User reviews with different age and gender are extracting and the opinion words from the reviews are identified. Finally, whether the extracted opinion words are positive or negative is also identified. The analysis of the drug reviews is performing by counting positive words and negative words from the complete review sentence.

$$pcomm = \frac{pword}{count} * 100 \quad (1)$$

$$ncomm = \frac{nword}{count} * 100 \quad (2)$$

Above Equation.1 shows the total positive comment i.e. pcomm in percent by dividing positive word (pword) by total word count (count).The Same Equation.2 calculates total negative comments i.e. ncomm in percentage by dividing negative word (nword) by total word count (count).

Following tables shows the different drugs with different age groups. Each drug is manually computed for evaluation with system computed value. In the table, "P" refers to positive comments and "N" refers to the negative comments. For each age group manually computed result shows huge difference than the value computed by drug review system.

Below Table no. 1 shows the information about Abilify drug. It describes the difference between system computed value and manually computed value of positive comments (P) and negative comment (N) for male and female. For example, age group 10-20, system computed for male shows the P= 62% and N= 38%. For female this values

become P= 88% and N= 12%. While manually computed value of same age is different i.e. for male P= 75% and N= 25% and for female P=45% and N= 55%. In the age group of 20-30, for male system computed values are P=55% and N=45% while manually computed value are P=65% and N=35%. And for the female the system computed value is P = 78% and N=22% which is compared with manually computed value P=70% and N=30%.

TABLE NO.1:
SENTIMENTAL ANALYSIS OF ABILIFY DRUG FOR DIFFERENT AGE GROUP

Abilify					
Sr. No.	Age groups	System computed value		Manually computed value	
		Male	Female	Male	Female
1	10-20	P =62% N=38%	P=88% N=12%	P=75% N=25%	P=45% N=55%
2	20-30	P=55% N=45%	P=78% N=22%	P=65% N=35%	P=70% N=30%
3	50-60	P=83% N=17%	P=36% N=64%	P=53% N=47%	P=69% N=31%
4	70-80	P=74% N=26%	P=90% N=10%	P=42% N=58%	P=51% N=41%

Table no. 2, gives the system computed value and manually computed value for the Carisoprodol drug, which is computed for different age group of male and female. The system value for age group 50-60 of male is P= 74% and N=26% compared with manually computed value P=65% and N= 35%, shows the difference of 9% in the value of P and N. Considering for female of age group 30-40, the system computed value is P=72% and N=28% while manually computed value is P=58% and N=42%. As shown in table 2, for age group 70-80 comparison between system computed value and manually computed value shows the difference of 6%.

Similarly, the table no. 3, shows the comparison between system computed value and manually computed value for different age group of male and female for the drug Bactrim.

TABLE 2:
SENTIMENTAL ANALYSIS OF CARISOPRODOL DRUG FOR DIFFERENT AGE GROUP

Carisoprodol					
Sr. No.	Age groups	System computed value		Manually computed value	
		Male	Female	Male	Female
1	20-30	P=77% N=23%	P=82% N=18%	P=69% N=31%	P=75% N=25%
2	30-40	P=89% N=11%	P=72% N=28%	P=85% N=15%	P=58% N=42%
3	50-60	P=74% N=26%	P=67% N=33%	P=65% N=35%	P=60% N=40%
4	70-80	P=79% N=21%	P=84% N=16%	P=73% N=28%	P=77% N=23%

TABLE 3:
SENTIMENTAL ANALYSIS OF BACTRIM DS DRUG FOR DIFFERENT AGE GROUP

Bactrim DS					
Sr. No.	Age groups	System computed Value		Manually computed value	
		Male	Female	Male	Female
1	10-20	P=84% N=16%	P=61% N=39%	P=76% N=24%	P=66% N=34%
2	40-50	P=46% N=54%	P=94% N=6%	P=56% N=44%	P=87% N=13%
3	50-60	P=26% N=74%	P=64% N=36%	P=31% N=69%	P=70% N=30%
4	60-70	P=88% N=7%	P=78% N=22%	P=85% N=15%	P=83% N=17%

V. CONCLUSION

Social networking sites provide a platform for online reviews through various discussion forums and different blogs for a different variety of products and services. Getting information out from these databases of texts is helpful as well as challenging. In this paper, the literature survey on drug reviews mining is presented various information about technique and how the techniques are evolved on drug reviews.

The objective of this research is to gives the summary of patients reviews which is obtain by the specific aspect. It is very useful to find the aspects of a product which gives people satisfactory result. Aspect mining is the part of the sentimental analysis. The method of extracting aspect from the drug review system using aspect mining method is working on actual data so this give an actual result of reviews to the user.The experimental result shows the comparison between system computed value and manually computed value which shows the techniques used in the drug review system are very promising in performing their task. Summarizing the patient's reviews is not only helpful for the people with the same condition but also important for pharmaceutical companies.

REFERENCE

[1] S. Baccianella, A. Esuli, and F. Sebastiani, "Multi-facet rating of product reviews," in Proc. 31st ECIR , Berlin,, Germany, 2009, pp. 461–472.

[2] 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.

[3] D.Giustini, "How web 2.0 is changing medicine," BMJ, vol. 333, no. 7582, pp. 1283–1284, 2006.

[4] J. Sarasohn-Kahn, "The wisdom of patients: Health care meets online social media," California Healthcare Foundation, Tech. Rep., 2009.

[5] 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. 1870–1880, 200

[6] 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. 297–306.

[7] Victor C. Cheng, C.H.C. Leung, Jiming Liu, Fellow, IEEE, and Alfredo Milani, "Probabilistic Aspect Mining Model for Drug Reviews", Hong Kong Baptist University, Kowloon 1234, Hong Kong

[8] V.Ranjani Gandhi, N.Priya "Literature Survey on Data Mining and Statistical Report for Drugs Reviews" International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 3, March 2015.

[9] 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.

[10] 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 onlin ecommunities," Inform. Technol. People, vol. 21, no. 4, pp. 350–374, 2008.

[11] Boulos, M. N. K., and Wheeler, S., "The emerging Web 2.0 social software: an enabling suite of sociable technologies in health and health care education", Health Information & Libraries Journal Volume 24, Issue 1, 2007, pp. 2–23.

[12] Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., Ngo, Q.-H., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., and Taniguchi, K., "BioCaster: detecting public health rumors with a Web-based text mining system", Bioinformatics Volume 24, Issue 24, 2008, pp. 2940-2941.

[13] Steinberger, R., Fuart, F., van der Goot, F., Best, C., von Etter, P., and Yangarber, R., "Text Mining from the Web for Medical Intelligence", F. Fogelman-Soulié et al. (Eds.), Mining Massive Data Sets for Security, IOS Press, 2008, pp. 295-310

[14] Jingiing Liu and Stephanie Senelf "A dialogue system for accessing drug review" Automatic Speech Recognition and Understanding (ASRU), IEEE Conference Publications, Pages: 324 – 329, Year: 2011

[15] Carolin Kaiser, Freimut Bodendorf "Mining Patient Experiences on Web 2.0"

[16] 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.

[17] Chenghua Lin, Yulan He, Richard Everson, "Weekly Supervised Joint Sentiment Topic Detection from Text" IEEE Transaction on Knowledge and Data Engine Ring. Vol: 24 No:6 Jun2012.

[18] Julien, F.Sha, and M.Jordan,"DiscLDA:Discriminative learning for dimensionality reduction and classification," in Proc. Adv. NIPS, 2008, pp. 897–904.

[19] W. Jin, H. Ho, and R. Srihari, "Opinionminer: A novel machine learning system for web opinion mining and extraction," in Proc. 15th ACM SIGKDD Int. Conf. KDD, New York, NY, USA, 2009, pp. 1195–1204

[20] Victor C. Cheng, Leung, Jiming Fellow, "Drug Review Mining with Regression Probabilistic Principal Component Analysis".. HI-KDD'12, 2012, Beijing, China. Copyright 2012 ACM 978- 1- 4503-1548-7/12/08.

[21] Shailendra Kumar Singh1, Sanchita Paul2 and Dhananjay Kumar "Sentiment Analysis Approaches on Different Data set Domain: Survey", International Journal of Database Theory and Application Vol.7, No.5 (2014), pp.39-50.

[22] Pang and L. Lee, "Opinion mining and sentiment analysis", Foundations and trends in Information Retrieval, vol. 2, (2008), pp. 1-135.

[23] K. Sarvabhotla, P. Pingali, and V. Varma, "Sentiment classification: a lexical similarity based approach for extracting subjectivity in documents", Inf. Retr., Vol. 14, No. 3, pp. 337–353, 2011.

[24] S. Zhou, Q. Chen, and X. Wang. "Active deep networks for semi-supervised sentiment classification", In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10, pp.1515–1523, 2010.

[25] S. Li, C. Huang, G. Zhou, and S. Y. M. Lee. "Employing personal/impersonal views in supervised and semisupervised sentiment classification", In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, pp. 414–423.

[26] S. Li and J. Hao. "Spectral clustering-based semisupervised sentiment classification", In Proceeding of 8th International Conference of Advanced Data Mining and Applications, pp. 271–283, 2012.

[27] 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.

[28] 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.

[29] Lin and Y. He, "Joint sentiment/topic model for sentiment analysis", in Proc. 18th ACM CIKM, New York, NY, USA, 2009, pp. 375–384.

[30] Yao Wu and Martin Ester "FLAME: A Probabilistic Model Combining Aspect Based Opinion Mining and Collaborativ Filtering" University Burnaby, BC, Canada.