Opinion Mining and Summarization of User statements in Health Communities

[1]Cleon Philip Sequeira, [2]Hemanth Kumar H.S, [3]Karthik Kumar, [4]Manu Bhargav S , [5]Prof. Prashanth M V

[1,2,3,4]Dept of Information Science and Engineering, Vidyavardhaka College of Engineering, Mysuru (Affiliated to Visvesvaraya Technological University, Belagavi)

[5] Assistant Professor, Dept of Information Science and Engineering, Vidyavardhaka College of Engineering, Mysuru (Affiliated to Visvesvaraya Technological University, Belagavi)

Abstract— *In the modern world, online health communities provide a variety of medical information helpful for medical practitioners, administrators and patients. In this task we collect real time health posts from reputed websites, where patients express their opinions, including their experiences and side-effects on drugs used by them. We deduce useful conclusions for medical fraternity as well as for the patient community, by performing Summarization of user health post based on the drug. Further, we also classify the users based on their experience with the drug. Also, we shall perform knowledge discovery from users opinion, whereby useful 'patterns' about the 'symptoms-drug-disease' is done by Association Rule Mining.*

Keywords— *Association rule mining, Classification, Keyword extraction, Knowledge discovery, Summarization*

I. INTRODUCTION

With the vast increase in web usage, information technology is also gaining rapid precedence which, although positive with respect to Information Age, creates overhead of time and space. Although the traditional medical methods is to study the human organs through various treatments, understanding of the knowledge is a difficult task. Summarization, opinion and association mining are used on the data for knowledge mining which are obtained by patientslikeme.com

Data mining is a process of extracting valid, previously unknown and actionable information from large data set.

One of the data mining techniques used is summarization technique which takes the information stored and extracting the required information. The user uses this extracted useful information which is in condensed form for their application.

Summarization is very important in different NLP applications like Information Retrieval, Text Comprehension, Quality Analysis, etc. Commonly there are two types of Summarization techniques. First one is extract in which, contents from text i.e. words and sentences are reused. Second one is Abstract which includes regeneration of extracted contents [2].

The most popular and widely-known data mining task is Association rule mining. It is used to find out interesting relations between variables in large database. Rules generated by association have two disjoint set of items. Association Rule is an implication expression of the form X Y, where X and Y are itemsets. Support is defined as the fraction of transactions that contain both X and Y. Confidence measures how often items in Y appear in transactions that contain X. Basically association rule works in two steps:

(1) Generating item sets that pass a minimum support threshold.

(2) Generating rules that pass a minimum confidence threshold.

The rules are then extracted and processed after they have been obtained. The extracted rules from the health boards dataset could take one or more of the following form-

1. Drug->disease
2. Disease->symptoms
3. Symptoms->disease
4. Disease->drug

Sentiment Analysis (SA) or Opinion Mining (OM) is a task of finding sentiments from data or text. These sentiments may take different forms like – opinions from people, attitudes and emotions toward an entity. The entity can represent events or topics, or even individuals. These topics are most likely to be covered by reviews. Classification levels considered were - document level, sentence level and aspect level [5].While doing SA, the forst important features are selected from text, then classified using an appropriate classifier.

We consider opinions from health posts and in our case the represented entity is drug. So our classification falls in this aspect.

II. NEED

Today social networking sites like Facebook, twitter are very popular. Sites like healthboards.com, patientslikeme.com, etc, are in the same way popular health related sites in the medical domain.

These sites contain large amount of data, along with useful information, they also contain some non-useful data. So there is need of summarization to extract important contents which can be viewed and understood at a glance.

Many a times, the interaction between doctor-patient is not friendly. Medical semantics used by doctors are difficult to understand by the patient community. But, patient-patient interaction language is fairly simple and easy to understand. This encourages us to find association among symptoms-drug-disease.

Drug manufacturers cannot cite all side effects of drugs within the stipulated time, for multiple reasons. Also the interactions between the same drug on different users is often not known. Hence sentiment analysis is required. According to this analysis, we propose to classify the users into two classes- depressed and satisfied. This classification will help us further provide indirect feedback to company.

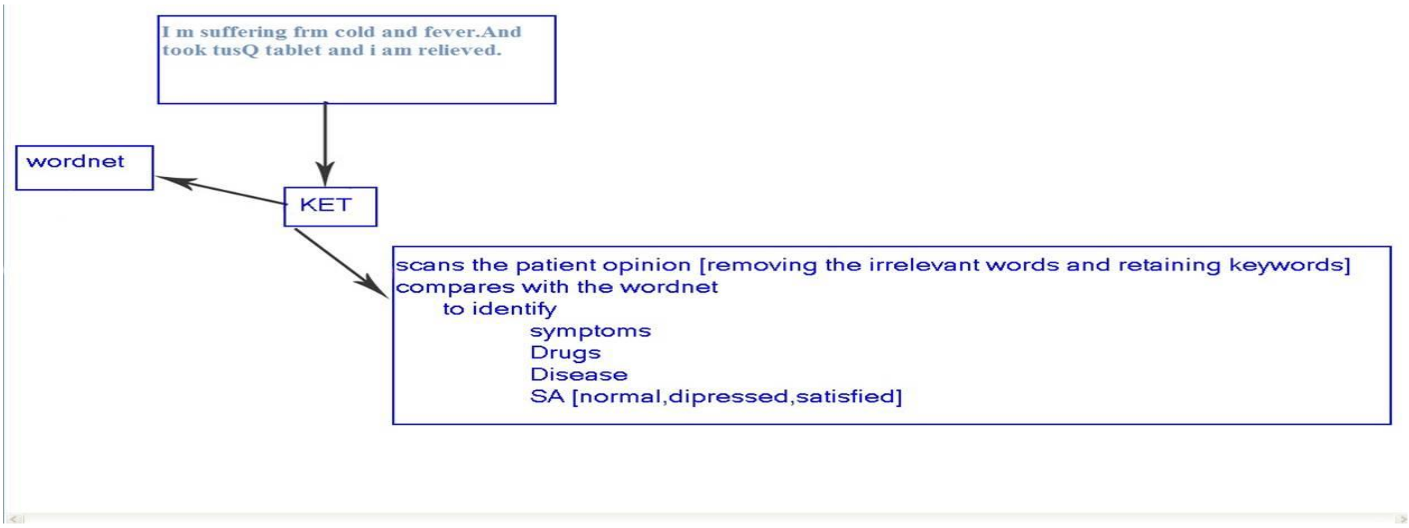
III. MOTIVATION AND RELATED WORK

There are several research topics that are closely related to our research. These topics are online health communities. The goal in summarization is to extract useful information from the large data available in the websites and other blogs. Vinod L. Mane [1] working for generation of useful information from the huge amount of data available in the different websites like patientslikeme.com ,healthboards.com , etc. and summarizing them for the next stage. AlokRanjan Pal simplified lesk algorithm which was one of the summarization technique used. Nandita Rane [2] worked for the association rule mining for type 2 diabetes. Algorithm for mining association rules have been used for extracting information from the posts. G.Vinodhini [4] used Sentiment analysis which involves in building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Sentiment analysis concentrates on attitudes, whereas traditional text mining focuses on the analysis of facts.[6] Text summarization based on scoring technique where the word, sentence and graph based scoring are put together to add on some weight. ROUGE is the evaluation method which is used for counting and selecting number of sentences. This method is not so reliable for the posts. [5.] Nasukawa and Yi. Illustrate a sentiment analysis approach to extract sentiments associated with polarities of positive or negative for specific subjects from a document, instead of classifying the whole document into positive or negative. The essential issues in sentiment analysis are to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject. Powerful functionality for these kinds of issues is used. Lakshmi K S [7] applied another method to apply association rules for medical documents which uses NLP tools, FP-growth and apriori algorithms. Because of the less amount of datasets used the rules was well known. Large datasets should be added to get new set of rules. B. M. Patil [12] the first stage, data preprocessing is done in order to handle missing values and equal interval binning is applied with approximate values based on medical expert advice and the next stage a well-designed association rule mining method Apriori association rule mining that describe item sets that satisfy a minimum support criterion.[9] Ding et al. proposed an effective method for identifying semantic orientations of opinions expressed by reviewers on product features. It is able to deal with two major problems with the existing methods, (1) opinion words whose semantic orientations are context dependent, and (2) aggregating multiple opinion words in the same sentence. For (1), a holistic approach is proposed that can accurately infer the semantic orientation of an opinion word based on the review context. For (2), a new function to combine multiple opinion words in the same sentence is proposed.

IV. PROPOSED WORK

Fig 3. Shows the system Architecture of our proposed system. We have identified different modules like -, Sentiment Analysis, Association, Summarization and Keyword Extraction .

Details of these modules are described below.



I m suffering frm cold and fever.And took tusQ tablet and i am relieved.

scans the patient opinion [removing the irrelevant words and retaining keywords]
compares with the wordnet
to identify
symptoms
Drugs
Disease
SA [normal,dipressed,satisfied]

Fig.1 Keyword Based Technique

1. Keyword Extraction

In this module, the user opinions are taken as the input which are obtained from healthboards.com. Each keyword from post is assigned to particular UMLS (Unified Medical Language System) category like symptom, disease, feeling, etc. Then keywords falling into the category disease, drug or symptoms are extracted for further processing.

I am suffering from cold and fever from a few months. I was prescribed citrizine and I am not getting any relief from my disease.

Fig.2 Sample Post

Fig.1 shows sample post for drug citrizine. Fig.2 shows assignment of UMLS categories to each keyword (Keyword: **UMLS category**).

cold: **Symptom or disease;** fever: **symptom or disease;** few: **quantitative concept;** months: **temporal concept;** citrizine: **drug;**

Fig.3 UMLS categories

2. Association

Using the extracted keywords from the above module we propose to determine the different types of association. These associations could be among symptoms-drugs-disease. We shall use Apriori algorithm for this purpose.

3. Summarization

3. Summerization

We are focused on providing summarization of topmost drug family only. So first we will group together all the posts related to topmost drug family. Then using Lesk based summarization algorithm [1] we can generate a summary. We are using Wordnetdictionary [13] to detect correct sense of word; it would help to generate better summarized results.

4. Sentiment Analysis (SA)

Text not only provides informative content but also attitudinal information, including possible emotional states of the user who posted his opinions. For SA, we collect all the posts per medical practitioner. To avoid unnecessary pre-processing we take the output of summarization as input for this module.

We are using tf-idf feature selection and bag-of word approach and also a classifier to classify user into different classes like satisfied and depressed.

V. FUTURE SCOPE

Posts of medical data on social media may contain a lot of errors or spelling mistakes. We are not considering spelling mistakes and their correction. So this could be further improvement. Also we could provide a online chat facility for a more interactive session. Posts on social networking may also contain symbolic expressions, which are not considered in this work.

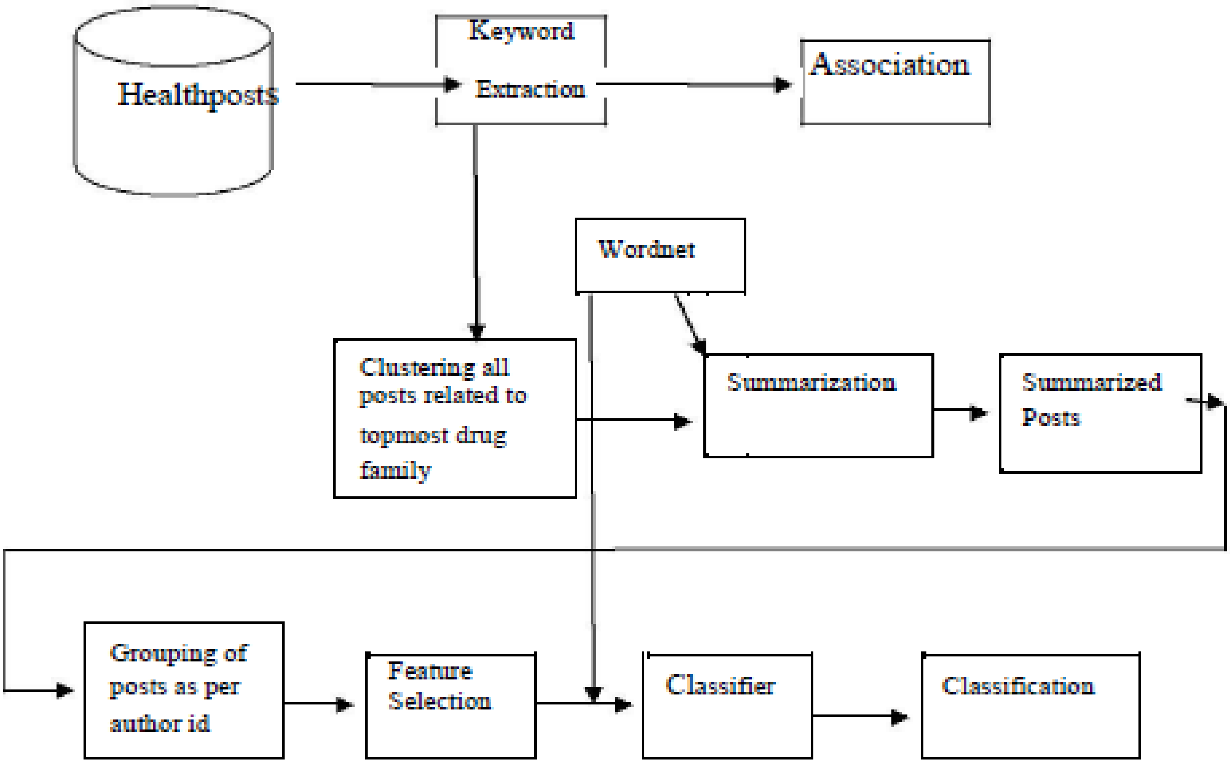


Fig.4 Proposed System Architecture

VI. CONCLUSION

Analysing user opinions from health communities for knowledge discovery is an interesting area in medical field. This work will help patients to find out association among different drugs, symptoms and disease. It will help doctors to find out different drugs and their interactions and side-effects so they can prescribe better drugs to other patients with similar disease. Pharmaceutical companies will also benefit from the classification of users of a particular drug into different classes like depressed and satisfied. This will be indirect feedback to companies to decide which drug is more effective, and/or whether to produce an alternate drug to this etc. Thus our work shall equally benefit all three parties–medical fraternity, pharmaceutical companies and the patient community.

REFERENCES

[1] Vinod L. Mane, Suja S. Panicker, Vidya B. Patil." Summarizationtiment analysis from user health posts", 2015 International Conference on Pervasive Computing (ICPC).

[2] Nandita Rane, Madhuri Rao," Association Rule Mining on Type 2 Diabetes using FP-growth association rule", International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 8 August, 2013

[3] Subhabrata Mukherjeey Gerhard Weikumy Cristian Danescu-Niculescu-Mizil," People on Drugs:
Credibility of User Statements in Health Communities", KDD '14, August 24 - 27 2014, New York, NY, USA
Copyright 2014 ACM 978-1-4503-2956-9/14/0

[4] G.Vinodhini, RM.Chandrasekaran," International Journal of Advanced Research in Computer Science and Software Engineering", Volume 2, Issue 6, June 2012 ISSN: 2277 128X.

[5] Shailesh Kumar Yadav," Sentiment Analysis and Classification: A Survey", International Journal of Advance Research in Computer Science and Management Studies Volume 3, Issue 3, March 2015.

[6] S.M.Nuwangi, C. R. Oruthotaarachchi, J.M.P.P.
Tilakaratna & H. A. Caldera, ‖Usage of Association rules and Classification Techniques in Knowledge Extraction of

[7] Diabetes‖, pp. 372–377.AlokRanjan Pal, DigantaSaha, "An Approach to Automatic Text Summarization using WordNet", IEEE International Advance Computing Conference (IACC), 2014.

[8] Lakshmi K S, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Yi Chen, Yunzhong Liu, "Connecting the Dots: Knowledge Discovery in Online Healthcare Forums", ICEC'14 August 05 - 06 2014, ACM.

[9] Chen, "Association rule mining to detect factors which contribute to heart disease in males and females", J. Nahar et al. / Expert Systems with Applications 40 (2013) 1086–1093, Elsevier, 2012.

[10] Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," Computational Intelligence, vol. 22, pp. 110–125,2006.

[11] Ding et al., Ahmed Hassan, HodaKorashy, "Sentiment analysis algorithms and applications: A survey", In press, Elsevier, 2014.

[12] Rudy Prabowo, Mike Thelwall, "Sentiment analysis: A combined approach .", Journal of Informetrics 3 (2009) 143–157.

[13] Milan Z, Gou M, Peter K et.al. Mining diabetes database
with decision trees and association rules[C]. In: Proceedings of the 15th IEEE Symposium on Computer- Based Medical Systems, 2002, pp. 867—871.

[14] B. M. Patil, R. C. Joshi, Durga Toshniwal, ‖Association rule for classification of type -2 diabetic patients‖, ‖, InProceedings of 2010 Second International Conference on Machine Learning and Computing, pp. 330–334.

[15] SaeedMohajeri,AfsanehEsteki, Osmar R. Zaiane and DavoodRafiei, "Innovative Navigation of Health Discussion Forums based on Relationship Extraction and Medical Ontologies",IEEE International Conference on Bioinformatics and Biomedicine, pp 13-14, 2013.