A Comparative approach for Sentiment Analysis from Summarized User Health posts

Mr Ajay A V, Dr.Chethan H K

P G Scholar, Department of Computer Science and Engineering, Maharaja Institute of Technology (MIT), Mysore, Karnataka, India

Professor, Department of Computer Science and Engineering, Maharaja Institute of Technology (MIT), Mysore, Karnataka, India

Abstract:Now a day there are number of online health communities provide a huge amount medical information that are posted by patients. The health post contains patients view, opinion or experience on particular drug. In this work we proposed to collect real time posts from websites like askapatient.com, healthboards.com etc. We proposed to summarize user posts using simplified Lesk algorithm. While summarizing the importance of each sentence is calculated using online dictionary called WordNet. Depending upon the percentage of summarization given by the user, the top ‘m’ sentences will be selected as summary. We proposed to classify users based on the sentiment expressed in their posts. To classify users we proposed to use Machine Learning Algorithms such as Naïve Bayes classifier and Support Vector Machine(SVM) and comparing both the classifiers.

Keywords:Summarization, Classification, Simplified Lesk algorithm, Naïve Bayes, SVM.

I. INTRODUCTION

With the enormous increase in web, electronic information is also increasing in huge amount which, although good with respect to Information Age, creates overhead of time and space. Also understand ability of information and consequent knowledge continues to be big challenges.

The patients are expressed their views on particular drug on online health forums like healthboards.com. These posts are difficult to analyse and predict what sentiment is being expressed. These problems can be resolved by using Summarization and Sentiment Analysis.

Summarization is the process oftaking an information source, extracting the content from it, and presenting the most important content to the user in a condensed form and in a manner sensitive to the user's application needs. There are two main techniques for Text Document Summarization: Extractive summary and Abstractive summary. While Extractive summary copies information that is very important to the summary, Abstractive summary condenses the document more strongly than extractive summarization and require natural language generation techniques.

Sentiment Analysis (SA) or Opinion Mining (OM) is the computational study of people’s opinions, attitudes and emotions toward an entity. The entity can represent individuals, events or topics. Sentiment Analysis can be considered a classification process. There are three main classificationlevels in SA: document-level, sentence-level, and aspect-levelSA. Document-level SA aims to classify an opinion documentas expressing a positive or negative opinion or sentiment. Itconsiders the whole document a basic information unit(talking about one topic). Sentence-level SA aims to classifysentiment expressed in each sentence. The first step is toidentify whether the sentence is subjective or objective. If thesentence is subjective,

Sentence-level SA will determinewhether the sentence expresses positive or negative opinions. Aspect-level SA aims to classify the sentiment with respect to the specific aspects of entities. The first step is to identify the entities and their aspects. Our word falls on Document Level SA since we considered the each user post as document.

II. NEED

Today social networking sites like Facebook, twitter are very popular. Like these social sites there are also health related social networking sites like healthboards.com, patientslikeme.com etc.

These sites contain large amount of data. Along with vital information, these 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.

Most of the times, doctor-patient interaction is not friendly. Medical language used by doctors is difficult to be understood by patient community. In other case, patient-patient interaction language is simple and easily understandable. This motivates us to find association among disease-drug-symptoms.

Companies cannot cite all side effects of drugs within short period of time, for multiple reasons. Also they should know effect of the same drug on different users. Hence sentiment analysis is required. According to sentiments, we propose to classify the users into three classes- normal, depressed and satisfied. This classification will help us further to give indirect feedback to company.

III. RELATED WORK

A summarization approach using simplified Lesk algorithm was used in [4]. The importance of a sentence in an input text is evaluated by the help of Simplified Lesk algorithm. As an online semantic dictionary WordNet is used. First, this approach evaluates the weights of all the sentences of a text separately using the Simplified Lesk algorithm and arranges them in decreasing order according to their weights. Next, according to the given percentage of summarization, a particular number of sentences are selected from that ordered list.

Extractive text summarization based on sentence scoring techniques is done in [9]. Word based, sentence based and graph based scoring methods are combined together to give weightage. Evaluation of summary is done by ROUGE as a quantitative measure and by counting number of sentences selected by system matching human gold standard as qualitative measure.This method is given for dataset of news, blogs and articles. However, sentence scoring technique is not suitable for posts.

The algorithm has been developed extracts key words from Kannada text documents, for which they combine GSS (Galavotti, Sebastiani, Simi) coefficients and IDF(Inverse Document Frequency) methods along with TF(Term Frequency) for extracting key words and later uses these for summarization. The important objective this work is to assign a weight to each word in a sentence, the weight of a sentence is the sum of weights of all words, based on the scoring of sentences; we choose top 'm' sentences. A document from a given category is selected from our database custom built for this purpose [1].

Online health communities are a valuable source of information for patients and physicians. However, such user generated resources are often plagued by inaccuracies and misinformation. In this work we propose a method for automatically establishing the credibility of user-generated medical statements and the trustworthiness of their authors by exploiting linguistic cues and distant supervision from expert sources. Important aspects arising from users posting their views online, like – credibility, trustworthiness, language objectivity etc [6].

Collecting opinions of people about products and about social and political events and problems through the Web is becoming increasingly popular every day. The opinions of users are helpful for the public and for stakeholders when making certain decisions. Opinion mining is a way to retrieve information through search engines, Web blogs and social networks. Because of the huge number of reviews in the form of unstructured text, it is impossible to summarize the information manually. Accordingly, efficient computational methods are needed for mining and summarizing the reviews from corpuses and Web documents. This study presents a systematic literature survey regarding the computational techniques, models and algorithms for mining opinion components from unstructured reviews [8].

The various Machine Learning algorithms for Sentiment Analysis such as Naïve Bayes and Support Vector Machine were given in [5].

IV. PROPOSED WORK

The proposed system architecture is as shown in Fig 1.

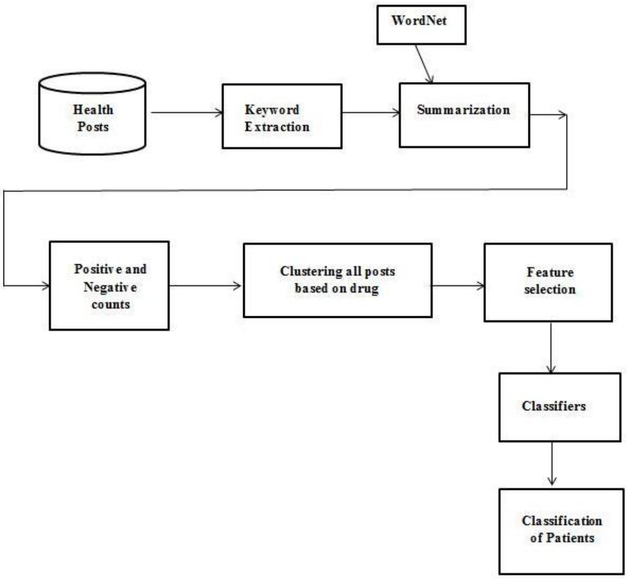
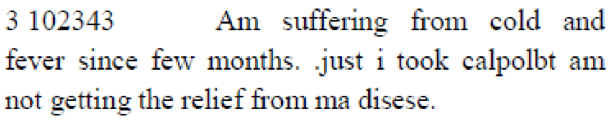

WordNet
Health Posts
Keyword Extraction
Summarization
Positive and Negative counts
Clustering all posts based on drug
Feature selection
Classifiers
Classification of Patients

Fig 1.The Proposed System Architecture

The proposed work has the main phases like Keyword Extraction, Summarization, Feature Selection and Classification of Patients.

i. Keyword Extraction

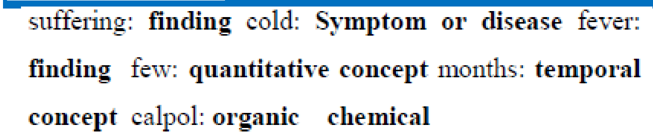
In this phase, input data is taken as user posts obtained from healthboards.com. Each keyword from post is assigned to particular UMLS (Unified Medical Language System) category like organic chemical, sign, symptom, disease, feeling etc. Then keywords having the category disease, drug or symptoms are extracted for further processing.



3 102343 Am suffering from cold and fever since few months. .just i took calpolbt am not getting the relief from ma disese.

Fig.2 Sample post

Fig.2 shows sample post for drug calpol. In this figure '3' is author id, '102343' is document id and rest of the part is actual post. Fig.3 shows assignment of UMLS categories to each keyword.



suffering: **finding** cold: **Symptom or disease** fever: **finding** few: **quantitative concept** months: **temporal concept** calpol: **organic chemical**

Fig.3 UMLS categories

ii. Summarization

Summarization is nothing but presenting the data in condensed form. There several algorithms available for this purpose. We proposed to use the simplified Lesk algorithm which finds the actual sense of each word using online dictionary called WoedNet.

iii. Feature Selection

Feature Selection in Sentiment Classification is nothing but extracting and selecting text features.Some of the current features are Terms presence and frequency, Parts of speech (POS), Opinion words and phrases, Negations etc. For this purpose the proposed system uses the BoW (Bag of Words) approach.

iv. Classification of Patients

We propose to classify user into different classes like normal, depressed and satisfied based on the feature extracted. Here we used two classifiers namely, Naïve Bayes and Support Vector Machine (SVM). For each post we are calculating the sentiment score. Depending on sentiment score both classifiers were compared.

V. Conclusion

Analyzing user posts from health communities for knowledge discovery is an interesting area in research. It will help doctors to find out side-effects of different drugs so they can prescribe better drugs to other patients with similar disease. Pharmaceutical companies will be also benefited as we are classifying users into different classes like normal, depressed and satisfied. This will be indirect input to companies to decide which drug is popular, whether to produce alternate drug to this etc. Thus our work shall equally benefit all three parties–medical fraternity, patient community and pharmaceutical companies.

VI. Future Work

Association rule can be used to find association among drug, disease and symptoms.Social media posts contain a lot of errors or spelling mistakes. We are not considering spelling mistakes and their correction. So this could be further improvement.

References

[1] JayashreeR,Srikanta Murthy K,Basavaraj .S.Anami, "*Categorized Text Document Summarization in the Kannada Language by Sentence Ranking*", 12th International Conference on Intelligent Systems Design and Applications (ISDA), pp 776-781, 2012.

[2] C. Lakshmi Devasenal and M. Hemalatha, "*Automatic Text Categorization and Summarization using Rule Reduction*", IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012), pp 594-598, 2012.

[3] 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.

[4] AlokRanjan Pal, DigantaSaha, "*An Approach to Automatic Text Summarization using WordNet*", IEEE International Advance Computing Conference (IACC), 2014.

[5] WalaaMedhat, Ahmed Hassan, HodaKorashy, "*Sentiment analysis algorithms and applications: A survey*", In press, Elsevier, 2014.

[6] Subhabrata Mukherjee, Gerhard Weikum, CristianDanescu-Niculescu-Mizil, "*People on Drugs: Credibility of User Statements in Health Communities*", KDD '14, August 24 - 27 2014, New York, ACM, 2014.

[7] Sara Keretna, CheePeng Lim, Doug Creighton, "*A Hybrid Model for Named Entity Recognition Using Unstructured Medical Text*", Proc.Of the 2014 9th International Conference on System of Systems Engineering (SOSE), Adelaide, Australia- June 9-13, pp 85-90, 2014.

[8] Khairullah Khan, BaharumBaharudin, Aurnagzeb Khan, Ashraf Ullah, "*Mining opinion components from unstructured reviews: A review*", Journal of King Saud University – Computer and Information Sciences (2014), Elsevier, 2014.

[9] Rafael Ferreira, FredericoFreitas, Luciano de Souza Cabral, Rafael DueireLins, Rinaldo Lima, Gabriel Franca, Steven J. Simske, and Luciano Favaro, "*A Context Based Text Summarization System*",11th IAPR International Workshop on Document Analysis Systems,pp 66-70, 2014.

[10] Yi Chen, Yunzhong Liu, "*Connecting the Dots: Knowledge Discovery in Online Healthcare Forums*", ICEC'14 August 05 - 06 2014, ACM.

[11] Pravesh Kumar Singh, MohadShahid Husain, "Methodological Study of Opinion Mining And Sentiment Analysis Technique", *International Journal on Soft Computing (IJSC)"*, volume 5, No. 1, Feb 2014.

[12] Aamera Z H Khan, Dr. Mohammad Atigue, Dr. V M Thakare, "*International JouranlOf Advance Research In Computer Science And Software Engineering*", Volume 5, Issue 4, April 2015

[13] Miller G, Beckwith R, Fellbaum C, Gross D, Miller K., "*WordNet: an on-line lexical database*", OxfordUniversity Press; 1990.

[14] Vinod L. Mane, Suja S. Panicker and Vidya B. Patil, "*Summarization and Sentiment Analysis from user health posts*", IEEE, 2015.