Research Paper
Volume : 5 | Issue : 8 | Special Issue August-2016 • ISSN No 2277 - 8179

Intelligent Data Mining Based on Sentimental Analysis for Improving Care

Engineering

KEYWORDS: Index Terms—Data Mining, Sentimental Analysis, Clustering, Classification

S.Kavitha — Asso.Professor, Computer Science and Engineering Vedavyasa Institute of Technology Karad,Malappuram

Lubna V — PG Scholar, Computer Science and Engineering Vedavyasa Institute of Technology Calicut, Kerala

ABSTRACT

Various websites provides the feature to put the reviews and their experiences on drugs and devices they are using. So these reviews can make use by the companies to make an analysis on their product and to know the customer satisfaction. Data mining is a powerful technology which helps companies to focus on the most important information they have collected. The system can perform sentimental analysis on the data for patient and consumer opinion on various drugs. It helps to reduce the cost of consumer generated opinion, also help the hospitals, pharmaceutical companies, and medical staffs to get up to date information on the effectiveness and ineffectiveness of the drugs. By using clustering and classification technique, we can predict the drug name on arriving a new review or for some side effects.

I. INTRODUCTION

Health care providers like medical staffs and pharmaceutical companies can collect feedback from doctors and patients to improve their services and results. Patients could use other consumers' opinion to make better knowledge on healthcare decisions. There are many traditional methods which provide information on latest technologies in healthcare field. Some methods are interviews, survey and questionnaires. But these methods are time consuming and not very effective since they are performed manually and the efficiency decreases with increased amount of data. Here is the need of data mining technology which is a "process of extracting knowledge from large amount of data" [1]. The data mining involves data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation [1]. Sentimental analysis is a process of determining whether a piece of writings is positive, negative or neutral. It helps to determine the attitude of a patient towards a drug is positive or negative. Clustering is the task of grouping elements in to different groups, so that the elements in the same group have similar features than the other group. So on arriving a new review we can classify the review into any one of the cluster groups to predict the drug name for that review. Classification is the process of categorizing the data into any one of the cluster groups to which it has maximum similar features.

The outline of the paper is as follows. Then in section 2 we describe the related techniques which contribute to the development of sentimental analysis on the user reviews. The Section 3 deals with the methods used in implementing the system. In section 4 we give our evaluation results and in section 5, we conclude our results.

II. RELATED WORK

The Term Frequency-Inverse Document Frequency (TF-IDF) [2] is a weighting method used for determining the importance of a term in a collection of document. It is used in text mining and information retrieval.

Natural Language Tool Kit (NLTK) [3] is a platform to work with human language data which is implemented in python. It provides a suit of text processing for stemming, tokenizing, stop word removal, tagging etc. it can be also used to find the number of words, number of unique words and to count the most frequent words in a corpora.

Many datasets are described as a linked collection of inter related objects. They may represent homogeneous networks in which there is a single object type and link type or heterogeneous networks in which there may be multiple objects and link types. The link mining [4] considers these links on data mining and these links are the relationship among data instances and it can indicate the properties of the data instances

Artificial neural network (ANN) [5] is a type of network which sees the nodes as artificial neurons. It consist of inputs which are multiplied by some weight and then computed by a mathematical function which determines the activation of the neuron. By increasing the weight the input become more stronger. We can also adjust the weight to obtain the desired output.

Self Organizing Map (SOM) [6] is a type of ANN used to produce low dimensional, discretized representation of the input space. Each node has a weight value which indicate the cluster content. The SOM bringing together similar data weights to similar neurons.

Frequent Subgraph Mining [7] is an active research topic in data mining. A graph is a general model to represent data. Mining patterns from graph database is challenging because of its higher time complexity. Fast Frequent Subgraph Mining is a new technique which reduce the number of redundant candidates to improve the time complexity.

POLAR(Probabilistic Object oriented Logics for Annotation based Retrieval) [8] is a frame work for supporting document and discussion search based on user comments. Also it provides a means to specify whether a reply is positive or negative. The discussion search not only regard a comment from an atomic value, but also take into account the context coming from the surrounding structure.

Machine learning is a major topic in Artificial Intelligence, which has increasing importance in recent years. YALE (Yet Another Learning Environment) [9] is a frame work for knowledge discovery and data mining and machine learning. It is implemented in java and it offers features such as integrated easy preprocessing of data, automatic selection of the best classifier and flexible attribute creation and selection.

S. Wasserman and K. Faust [10], proposes a social network analysis software used to identify, represent, analyze or simulate nodes and edges from various types of input data. The output data can be saved in external files. Visual representations of social network are important to understand network data and to convey the result of analysis. The network analysis tool is used by the researchers to investigate representation of networks of different size.

Altug Akay, Andrei Dragomir and Bj¨orn-Erik Erlandsson, [11] proposes a method for generating the feedback of drugs by using data mining and sentimental analysis techniques on the user reviews

posted on social medias. These feedback can be used by the hospitals, pharmaceutical companies etc to get the up to date information.

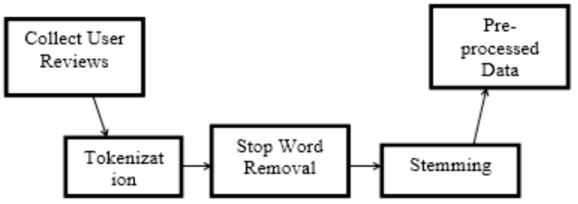

Fig 1. Pre-processing of the Collected User Reviews

II. IMPLEMENTATION

3.1 Initial Data Collection and Preprocessing

The first step is to collect the user reviews from websites which provide discussion facility for the consumers (www.drugs.com and www.webmd.com) . we focused on the number of posts on lung cancer, because it is the most frequently diagnosed cancer in the world for both sexes. We then find a list of drugs that lung cancer patients were most discussed in the sites. In the next step all the reviews are

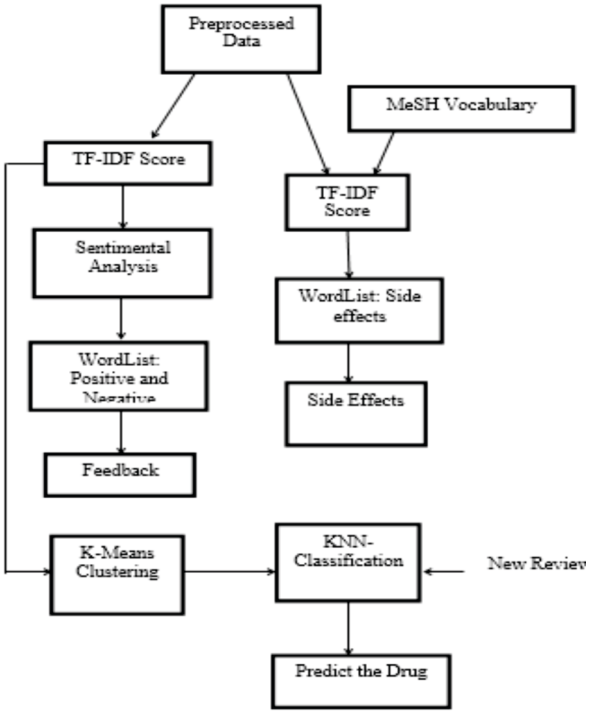

Fig 2. System Architecture

preprocessed for unwanted words, symbols and characters. The architecture of preprocessing is shown in figure 1. It includes tokenizing the reviews into tokens, removing stop words and stemming the words to get the root form of each word.

3.2 Term-Matrix Creation

The term-matrix is created by calculating the TF-IDF [2] score of each word in the preprocessed data using the formula:

$$weight_{t,d} = tf_{t,d} + \log \frac{n}{D_{t,d}}$$

Where $tf_{t,d}$ is the frequency of the term t in document d, n is the total number of documents and $D_{t,d}$ is the total number of documents containing the term t.

3.2 Sentimental Analysis and Tagging data

Text data containing the highest TF-IDF scores are annotated using the POS (Part Of Speech) Tagger with noun (NN), verb (VB), adjective (JJ) and adverb (RB). SentiWordNet [12], a lexical resource in which each synset of WordNet is associated to three numerical scores Obj (s), Pos (s) and Neg(s), describing how Objective, Positive and Negative the terms contained in the synset are. Each of the three scores range from 0.0 to 1.0 and their sum is 1.0 for each synset. Create two word lists containing the positive and negative terms. Exclude the words which appear less than ten times and the final word list is shown in the Table I.

TABLE I. Post Analysis Word List

Positive	Negative
Great	Fear
Effective	Died
Good	Bad
Positive	Suffering
Improve	Problem
Comfort	Damage
Help	Hurt
Well	Worst
Greater	Hard
Right	Terrible

In parallal procedure we automatically look for the side effects drugs. For this we used the National Library of Medicine's Medical Subject Heading (MeSH), which is a co-ntr-olle-dvo-ca-bu-lar-y(http://www.nlm.nih.gov/mesh/) that consist of a collection of descriptors and qualifiers used to annotate medical terms. The list of words present in the user reviews that were associated to treatment side effects are thus compiled using a custom designed program to map the words in the reviews to MeSH database. The full list of side effects then preprocessed and kept the side effects with highest TF-IDF score. The final list of side effects is shown in Table II.

3.2 Identification of Side Effects and Feedback generation.

To this goal, we overlay the word list in table 1 and table 2 to the preprocessed data obtained in section 3.1 for feedback and side effects. A statistical method can be used to compare the values in word lists to the overall population and identify the variables that have high scores.

3.3 Clustering

The term-matrix obtained in section 3.2 are clustered using the K-Means clustering Algorithm shown in Algorithm 1 into different cluster groups. Each cluster represents the features of a particular drug.

3.4 Classification for a new review

The system can predict the drug name for some side effects or for a new review. To this goal, we classify the new review using KNN-Classification algorithm shown in Algorithm 2 to any one of the cluster groups to which it has maximum similar features.

TABLE II. Side Effects Word List

discomfort
nausea
fatigue
bleed
dermatitis
redness
constipation
Diarrhoea
rash

Algorithm 1:

1. Initialize the centre of clusters
μi= some value ,i=1,...,k

2. Attribute the closest cluster to each data point
$\mathbf{c}i=\{j:d(\mathbf{x}j,\mu i)\leq d(\mathbf{x}j,\mu l), l\neq i, j=1,...,n\}$

3. Set the position of each cluster to the mean of all data points belonging to that cluster
$\mu i=1/|ci|\sum j\in ci\,\mathbf{x}j,\forall i$

4. Repeat steps 2-3 until convergence

Algorithm 2:
1.Determine parameter k=number of nearest neighbours

2. Calculate the distance between the query-instance and all the training samples

3.Sort the distance and determine nearest neighbours based on the k^{th} minimum distance

4. Gather the category Y of nearest neighbours

5. Use simple majority of the category of nearest neighbours as the prediction value of the query instance

4 EVALUATION

The preprocessed data are weighted with the TF-IDF score as shown in Table III. The words containing scores greater than 0.06 are collected and tagged. Then calculated the sentiscore of each word to determine the positivity and negativity of that word as shown in Table IV. The results are obtained as the percentage of satisfied and unsatisfied users, and the side effects can cause from that drug as given the Table V. The K-means clustering algorithm efficiently clustered the given dataset into cluster centroids. The KNN Classification algorithm classifies the new review better into the one of the cluster centroids and predicted the drug name for that review as shown in Table VI.

TABLE III. Term Matrix containing TF-IDF scores for each term

Document no	Start	Progress	Injection	Stronger
D1	0.0876	0.1165	0.0765	0.239
D2	0.587	0	0.0465	0.437
D3	0	0.09476	0.1567	0.6834
D4	0.1154	0	0.4962	0.08497
D5	0.9174	0.0698	0.0386	0

TABLE IV. Sentiscore of each synset

Words	Reaction	Great	Steroid
Adj_pos	NA	0.0316	NA
Adj_obj	NA	0.6666	NA
Adj_neg	NA	0.02654	NA
Noun_pos	0.03765	0.0	0.0
Noun_obj	0.9854	1.0	1.0
Noun_neg	0.0	0.0	0.0
Verb_pos	NA	NA	NA
Verb_obj	NA	NA	NA
Verb_neg	NA	NA	NA

TABLE V. User Opinion about Celebrex

Satisfaction	**57%**
Dissatisfaction	43%
Side Effects	Discomfort nausea

TABLE VI. Drug prediction for the new review

User Review: I have used the drug for 16 years and had very little negative result. I felt pain rashes on my skin. Then i reduced the dosage and consulted the doctor.
Drug: Abraxane

5. CONCLUSION

Sentimental analysis on the data collected from various discussion forums is beneficial for mankind in providing them better health care. The work flow of analysis of health care content helps to overcome the limitations of large scale data analysis and the analysis of user generated textual contents in discussion forums. It helps the patients to make better decisions on their drug usage and are able to identify the drugs for certain side effects. The system helps the pharmaceutical companies and healthcare providers to work on the feedback and try to come up with the improvised medicines and treatments for lung cancer.

ACKNOWLEDGMENT

I am indebted to the HOD Mrs. Kavitha Murugesan of our department, and Ms. Ginnu George for their guidance and the workshop reviewers for their comments on my work.

REFERENCES

[1.] J. Hans and M. Kamber, Data Mining: Concepts and Techniques. 2nd ed. Burlington, MassMA, USA: Morgan Kaufmann, 2006.

[2.] P. Soucy and G. W. Mineau, "Beyond TFIDF weighting for text categorization in the vector space model," in Proc. 19th Int. Joint Conf. Artificial Intell., Edinburgh, U.K., 2005, pp. 1130–1135

[3.]]. S. Bird, E.Klein, and E. Loper, Natural Language ProcessingWith Python. Sebastopol, CA, USA: O'Reilly Media, 2009, pp. 504

[4.]]. L. Getoor and C. Diehl, "Link mining: a survey," SIGKDD Explor. Newsl., vol. 7, pp. 3–12, Dec. 2005

[5.] Carlos Gershenson, Artificial Neural Networks for Beginners.

[6.] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, "Self- Organizing Map in MATLAB: The SOM Toolbox," in Proc. Matlab DSP Conf., Espoo, Finland, 1999, pp. 35–40.

[7.] J. Huan and J. Prins, "Efficient mining of frequent subgraphs in the presence

[8.]]. I. Frommholz and M. Lechtenfeld, "Determining the polarity of postings for discussion search," in Proc. Workshop Woche Lernen Wissen Adaptivit ¨at., W¨urzburg, Germany, 2008, pp. 49–56.

[9.] I. Mierswa, M. Wurst, W. Michael, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid prototyping for complex data mining tasks," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Philadelphia, PA, USA, 2006, pp. 935–940.

[10.] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications. New York, NY, USA: Cambridge University Press, 1994, pp. 825

[11.] Altug Akay, Andrei Dragomir,and Bj¨orn-Erik Erlandsson, Network Based Modelling and Intelligent Data Mining of Social Media for Improving Care, IEEE Journal Of Biomedical And Health Informatics, Vol. 19, No. 1, January 2015

[12.] Andrea Esuli_ and Fabrizio Sebastiani, "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining".