Intelligent Information Extraction from Big Data using Self Organizing Map

Senthamarai.R[1] Mary Shamala.L[2]
[1]P.G. Student [2]Assistant Professor
[1,2]Department of Computer Science and Engineering
[1,2]IFET College of Engineering, Villupuram, Tamilnadu, India

Abstract— Knowledge extraction from social media has recently attracted great interest from the biomedical and health informatics community. Sentiment analysis has emerged as a popular and efficient technique for information retrieval and web data analysis. Such intelligent system improves healthcare outcomes and provides self-awareness using consumer opinion. The Proposed system uses natural language processing (NLP) which involves a two-step analysis framework that focuses on positive and negative sentimental analysis, as well as the side effects of treatment through users' forum posts. Regression process is used to merge the data from two-step analysis by using NLP approach. Finally self-Organizing Map (SOM) is enabled to classify the merged data. After classification SOM analyze the data by knowledge learning and token value is assigned for each medicine. Here token values play a vital role to list the appropriate medicines as per their priority. The proposed system may provide self-awareness to pupil by checking whether they are using the unbanned and effective medicine, thereby increasing healthcare outcomes by using user opinion from the medical web forum data.

Key words: Sentiment analysis, Natural Language Processing, Self-Organizing Map, Tokens

I. INTRODUCTION

In the last few years the growth and use of internet increases and sharing of user's opinions also increases. social media platforms like Twitter, Facebook, MySpace and other forums provides a platform for people to express their emotions in the digital world, which provide valuable information[5]. By using this information we can make several decisions which are suitable to our work. Opinion can be collected from any individual in the world about anything through review sites, blogs, web forums. Organizations and product owners makes this feedback of users or customers to improve their products/services. Generally used sources for finding opinion are Blogs, review sites, raw dataset, and Micro-blogging web sites. Online messages that are posted by individual in the internet are mostly informal. Analysis and handling of this kind of text is often more difficult when compared with formal texts[4]. The main difference between formal and informal text is in data preprocessing for formal text require less preprocessing whereas informal text contains emotions, sarcasm, utilization of weak grammar, and non-lexicon-standard words. Therefore, extraction of informal content is more troublesome.

People frequently ask their friends, relatives, and field specialists for suggestion during the decision-making procedure, and their opinions and perspectives are based on experiences and observations. One's point of view a subject can either be positive or negative, which is known as the polarity detection of the sentiment. During sentiment analysis process, it requires very fast and concise information so that any individual can make quick and accurate decisions [4]. Nowadays, large attention has been given to sentiment analysis because of its wide range of possible applications[5].Sentiment Analysis is an emerging field for research which deals with information extraction and knowledge discovery from text using Natural Language Processing (NLP) and Data Mining (DM) techniques. Supervised learning and unsupervised learning is the type of machine learning techniques used for sentiment classifications. Self-Organizing Map is one of the most popular neural network models. It is based on unsupervised learning, which means that no human intervention is needed during the learning and that little needs to be known about the characteristics of the input data. This map is used for classifying the input data that are given by the user.

II. EXISTING METHODOLOGIES

Network based modeling and Intelligent data Mining of Social Media for Improving Care[3] has a two-step analysis framework that focuses on positive and negative sentiment, as well as the side effects of treatment, in users' forum posts, and identifies user communities (modules) and influential users for the purpose of ascertaining user opinion of cancer treatment. Here a Self -Organizing Map is used to analyze word frequency data derived from users' forum posts. Then employed a network-based approach for modeling users' forum interactions and employed a network partitioning method based on optimizing a stability quality measure. This allows consumer opinion and identifying influential users within the retrieved modules using information derived from both word-frequency data and network-based properties. These are processed using the following

A. Initial Data Search and Collection

Searching for the most popular cancer message boards for data collection and initially focused on the number of posts on lung cancer.

B. Initial Text Mining and pre-processing

After data collected from the message boards it is processed for the common positive and negative words and their term frequency –inverse document scores within each post.

C. Cataloging and Tagging Text Data

Text data containing the highest TF-IDF scores were tagged with a modified NLTK toolkit using MATLAB to ensure that they reflected the negativity word and the positivity of a positive word in the context. This approach was used before using negative tags on positive words. Then they added a positive tag on negative words and used the NLTK toolkit for the analysis, and classification, of words to match their exact meanings within the contextual settings.

D. Side-Effects

In a parallel automatically browsed the user posts to look for side effects. For this used the National Library of Medicine's Medical Subject Heading (MeSH), which is a controlled vocabulary that consists of a hierarchy of descriptors and qualifiers that are used to annotate medical terms. A custom designed program was used to map words in the forum to the MeSH database. A list of words present in forum posts that were associated to treatment side effects was compiled.

E. Consumer Sentiment Using A Self-Organizing Map

Here analyzing, all posts were manually labeled according to the general user opinion observed within the post as positive and negative before feeding the collected data for exploratory analysis via Self- Organizing Maps. SOMs are neural networks that produce low-dimensional representation of high-dimensional data. Within this network, a layer represents output space with each neuron assigned a specific weight. The weight values reflect on the cluster content. The SOM displays the data to the network, bringing together similar data weights to similar neurons.

F. Modelling Forum Postings Using Network Analysis

Influential users were the next step in our analysis. For this built in networks from forum posts and their replies. Networks are composed of nodes and their connections: they are either non-directional or directional.The nodal degree of the latter measures the number of connections from the origin to the destination.

G. Identifying Sub-Graphs

The modeled framework has consequently converted the forum posts into several large directional networks containing a number of densely connected units (or modules). These modules have the characteristic that they are more densely connected internally (within the unit) than externally (outside the unit). They chosen a multi-scale method that uses local and global criteria for identifying the modules, while maximizing a partition quality measure called stability.

H. Module Average Opinion and User Average Opinion

To refine the information modules with the information obtained from the forum posts (using the wordlist vectors). The global measure (pertaining to the whole information module) is represented by the module average opinion (MAO). It examined the TF-IDF scores of postings matching the nodes in a specific models.

I. Information Brokers within The Information Module

Ranked the individual nodes in terms of their total number of connecting edges (in and out-degree) to identify influential users within the modules. Then looked nodes in each module based on the following criteria

- The nodes have densest degrees within the module (highest number of edges).
- The UAO scores equate the signs of the MAO of the containing module.

J. Network-Based Identification Of Side Effects

In the second step of network-based analysis, they devised a strategy for identifying potential side effects ooccurred during the treatment and which user posts on the forum highlight. The TF-IDF scores within each module will directly reflect how frequent a certain side-effect is mentioned in module posts.

III. PROPOSED SYSTEM

The method of Intelligent Information Extraction from Big data using Self-organizing map is to merge a two-step analysis framework that focuses on positive and negative sentiment, as well as the side effects of treatment, from user's forum posts. Regression analysis is used to merge the data from two-step analysis of Natural Language Processing (NLP) approach and make a clustering of that merged data. Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). Then the clustered data is given to the Self Organizing Map (SOM) for analysis and classification. After classifying the merged data token value is assigned to each medicine to know their priority.

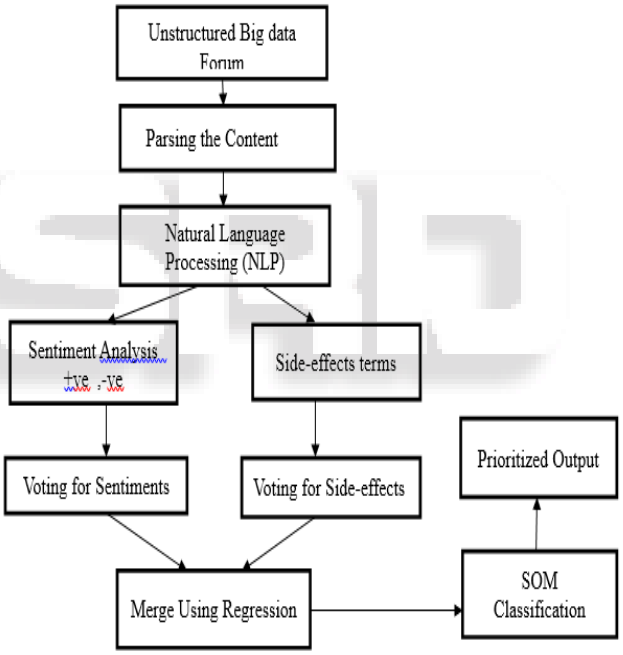

Fig. 1: System Architecture of Intelligent Knowledge Extraction

Fig.1 gives the system architecture of intelligent knowledge Extraction from the unstructured data that is collected from the big data forum. There available lots of blogs, web forums and websites are made used to collect the input data which is then pre-processed through parsing those content. NLPTK is then used to perform NLP as well as voting the processed terms. Then finally the processed terms from NLPTK are merged using regression analysis and then forwarded to further classification and assigning priority to the drugs using SOM.

Modules

The proposed system includes the following modules

a) Information Extraction Using Natural Language Processing
b) Clustering using Regression
c) SOM classification

A. Information Extraction Using Natural Language Processing

Information extraction is the task of automatically extracting structured information from unstructured and semi-structured machine readable documents. In most of the cases this activity concerns processing human language texts by means of NLP. NLTK toolkit is used for the analysis, and classification of words to match their exact meanings within the contextual settings. It means pre-processing the text data that comes from the user medical forum. In pre-processing the stop words and stemming words are filtered and then the voting is given for the analysis data

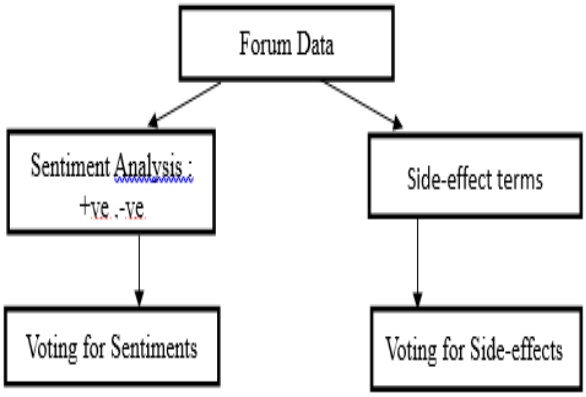

Fig. 2: Process of information extraction Through NLP

B. Clustering Using Regression

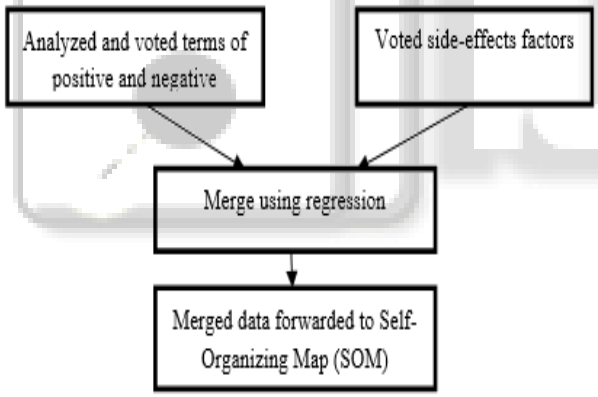

Fig. 3: Clustering

It is the task of grouping a set of objects in the same group. It is not an automatic task, but an iterative process of knowledge discovery. It is a main task of exploratory data mining and a common technique for statistical data analysis used in many fields, including machine learning, image analysis and information retrieval. Here the merged data using regression analysis is clustered and is given to self-organizing map for classification.

C. SOM Classification

The self-organizing map is one of the most popular neural network models. It is trained using unsupervised learning which means that no human intervention is needed during the learning and it needs to be known about the characteristics of input data. Self-organizing map operates in two modes: Training and Mapping. Training builds the map using input. Mapping automatically classifies a new input vector. Process of the self-organizing map is constructed by using the following steps

- Step 1: Initialization - Receive merged data after regression.
- Step 2: Sampling - Drawing a classifying sample training input vector from input space.
- Step 3: Updating - Apply the weight (Tokenizing).
- Step 4: Continuation - Report sampling for another set of input.

IV. CONCLUSION

To measure consumer thoughts on the drugs for lung cancer disease using positive and negative terms alongside another list containing the side effects by using Natural Language Processing (NLP) technique for information extraction. After the analysis of the user sentiments voting is given for the positive and negative words as well as to side-effects terms and this analyzed data are merged using the regression analysis and it forms a clustering. Then this clustered data are given to Self-organizing Map (SOM) for classification. After classifying the input data then the self-organizing map analyses the classified data by his knowledge learning and it maps the large dimensional data onto a lower dimensional space using the SOM. Assigning token value for the analyzed drug and list the appropriate medicines as per their priority.

REFERENCES

[1] BlessySelvam, S.Abirami, "A Survey on Opinion Mining Framework", International Journal of Advanced Research in Computer and Communication Engineering , Vol. 2, Issue. 9, Sep 2013.
[2] Bo Pang, Lillian Lee, ShivakumarVaithyanathan, "Thumbs up Sentiment Classification using Machine Learning Technique," proceedings of Conference on Empirical methods in natural language processing (EMNLP), Vol. 10, pp. 79-86, 2002.
[3] Akay, A.; Dragomir, A.; Erlandsson, B, "Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care," in Biomedical and Health Informatics, IEEE Journal, Vol.19, no.1, pp.210-218, Jan 2015.
[4] Dharmesh Ramani, Hazari Prasun, "A survey: Sentiment Analysis of Online Review", November 2013.
[5] Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, May 2012.
[6] Gaurav S. Chavan, Sagar Manjare, "A survey of various Machine Learning Techniques for Text Classification", International Journal of Engineering Trends and Technology, Volume 15, 2014.
[7] Khan, K.; Baharudin, B.B.; Khan, A.; e-Malik, F., "Mining opinion from text documents: A survey," in Digital Ecosystems and Technologies, 3rd IEEE International Conference on , Vol. 1, no. 3, pp.217-222, June 2009.
[8] Basu. A, Walters. C, Shepherd. M, "Support vector machines for text categorization," Proceedings of the 36th Annual Hawaii International Conference, pp. 6-9, Jan 2003.

[9] G.Angulakshmi,Dr.R.Manickachezian,"An Analysis on Opinion Mining: Techniques and Tools," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue. 7, July 2014.
[10] Desjardins. G, Godin. R, Proulx. R, "A self-organizing map for concept classification in information retrieval," in Neural Networks, Proceedings of IEEE International Joint Conference, Vol.3, pp.1570-1574, Aug 2005.

