VOL. 11, NO. 13, JULY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

CATEGORIZATION OF DRUGS BASED ON POLARITY ANALYSIS OF TWITTER DATA

S. Rijo Meris, R. Raja Singh and J. Andrews
Faculty of Computer Science, Sathyabama University, Chennai, India
E-Mail: rijomeris@gmail.com

ABSTRACT

According to the present scenario, social media have emerged as major platforms for sharing information in medical field, business, education etc. In the existing system, the system will generate warning for adverse drugs reactions based on the negative comments. Social media provides limitless opportunities. The drug and disease related tweets are extracted from twitter API and web crawler based on the given input. The drugs can be predicted whether it is a best drug using polarity by extracted tweets are pre-processed by removing stop words, abbreviations and replacing. By this system, the user will not know which is the good medicine. In the proposed system, the consumer will gain knowledge about the best medicines.

Keywords: twitter data, polarity analysis, social media, pre processed, crawler, drugs.

1. INTRODUCTION

When the user post tweets, those. Retweeting is the processing of forwarding the tweet posted by other user in twitter. Clients can do assemble posts by point or sort by utilization of hash tags - words or expressions prefixed with a "#" sign. Essentially, the "@" sign is trailed by a username is utilized for specifying or answering to different clients. The way of gigantic tweets makes it convoluted for scientists to approve information produced from it. With the prevalence of online networking, twitter is the propelled approach to share information for customers to share their experience in light of medications and infections. To extract the tweets, first the connection should be established with twitter account using the twitter API called twitter4j. Then create the twitter developer application in twitter developer site. From the developed application we get the keys and tokens. Using these keys and tokens, it is Configured and connected with twitter.

2. REVIEW OF RELATED WORK

Xiao Liu and sinchun Chen discovered that online networking can be utilized for recognizing the adverse drug events. The patients information is collected from online community. [1]

Carissa Hilliard proposed Telemedicine concept. Telemedicine can be extremely effective through the social media in regions where local medical care is insufficient and lack of high caring healthcare. This shows that older users are becoming increasingly more comfortable with social media and are beginning to take advantage of its accessibility. [2]

Ming Yang et al, developed social networking communication based on their experience after using the drugs in Web Forum. [3]

Gopal have picked the Twitter4j library in our engineering since it is a standout amongst the most utilized open source libraries as a part of Twitter information administration and has successive updates, making it a solid decision. [4]

Traditional adverse event reporting systems have been slow in adapting to online AE reporting from patients, instead of waiting in clinics and drug safety groups. In mean time, the users in the social media to share their experiences with drugs etc. Tweets lack the structure and standardized reporting that comes from identified adverse events in a clinic. Patient may forget to tell some important details while speaking to the clinical physicians [5].

Robert Leaman *et al*, they proposed a system for mining the relationship between drugs and adverse drug reaction based on the patience post from various health-related website. . The comments are annotated into four categories: adverse effect, beneficial effect, indication and other. Annotation practice is done in four steps. The number of tokens in the longest term found by the annotators. [6]

Lorie and Richard analyzed that during H1N1 pandemic, Twitter was used as a source of communication to update the public at flu vaccination clinics and for the distribution of government alerts. More than clinical information, participants within the network posts experiential health information, reporting daily events. [7]

Adam Sadilek *et al,* they focused on fine- grained modeling of the spread of infectious disease throughout a large real-world social network. In public Twitter data, we receive the tweets with the combination of sick and health related message. So, they achieved sick tweets by developing an SVM model that is robust even in the presence of class imbalance. This model uses significant negatively and positively features for identifying the spreading of diseases. [8]

Joao Cunha *et al,* Twitter data must be analyzed carefully because the semantics of sentence may change the meaning of words. [9]

Joao Cunha et al, Twitter data must be analyzed carefully because the semantics of sentence may change the meaning of words. Twitter4j API extract tweets and stores in MongoDB. Then tweets are processed, the data must be serialized before it is converted into structured

data. In this system they have developed an interface to both user and framework components. They have used Mahout which consists of build-in libraries to apply algorithm and sent to database for analysis. [10]

3. SYSTEM ARCHITECTURE

Architecture on polarity analysis of twitter data is used to analysis the best medicine. Initially, user name and password is given to login, if the user is a registered user otherwise the user has to register first, then login to search page.

When user fill the information during registration it gets stored in database so during the login it gets validates from the database. If the username and password is correct then it enters into twitter search page. In twitter search page, the user can give the input to extract and crawler the tweets from the twitter.

To extract the tweets, first we have to connect the application with twitter using Twitter API. Using Twitter4j the application get connected to twitter application based on consumer key, secret key, Access token and token secret key. The tweets are then pre processed by removing stop words, replacing short form with full words and replace with their respective meaning.

The pre processed data is then stored in database. In future, the pre processed tweets are classified using SVM classification. The polarity of the words such as good or bad etc is identified from classified tweets to classify whether it is positive tweets or negative tweets. Polarity can be predicted based on the number of positive comments and negative comments.

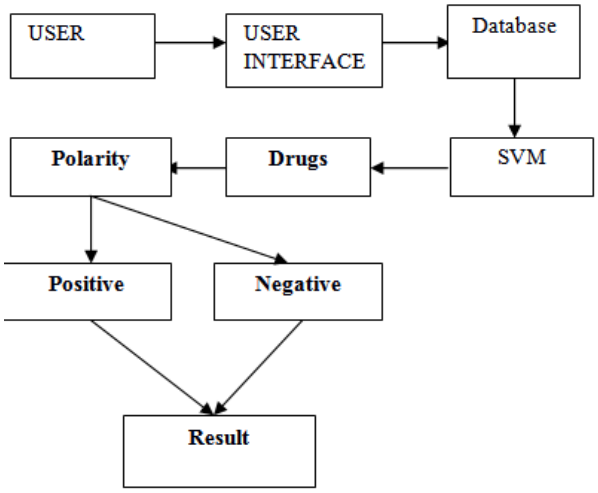

Figure-1. Architecture.

4. SYSTEM IMPLEMENTATION

- Twitter Extraction
- Preprocessing
- SVM classification
- Polarity prediction

4.1 Twitter extraction

User can interact as interface between the user and the system. New user have to create an account by giving the username and password, the registered user can directly login and can enter into the system twitter search space. In search space user can give the input, and user get the tweets from the twitter.

To extract the tweets, first the connection should be established with twitter account using the twitter API called twitter4j. Then create the twitter developer application in twitter developer site. From the developed application we get the consumer key, secret key, Access token and token secret key. Using these keys and tokens, it is Configured and connected with twitter. In this API it contains many parameters to extract and read from the Twitter Factory by using query search and have to manage the query search concludes in Query Result. Using get Tweets method we can get the tweets, from which we can extract the tweet username.

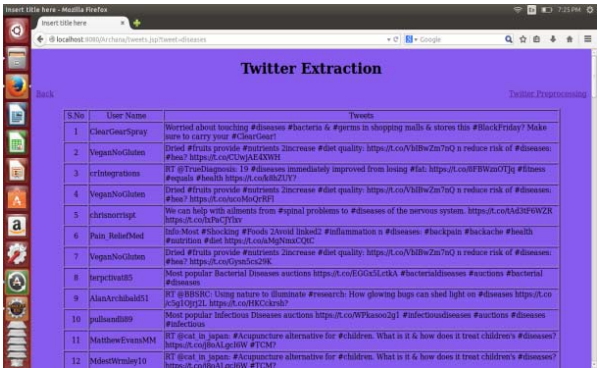

Figure-2. Twitter extraction.

4.2 Preprocessing

The extracted tweets are the pre-processed by removing stop words, short form and All meaningless words in the tweets such as stop words are been removed. All short forms will be replaced with full words so that it is understandable for all the users. Are known as smiley's, there are varies kinds of smiley's. For each smiley's there are some emotional feelings in it, which the user use to communicate in much easier manner but it is not necessary all the user will know the meaning of all. So, all the is replaced with their respective meaning.

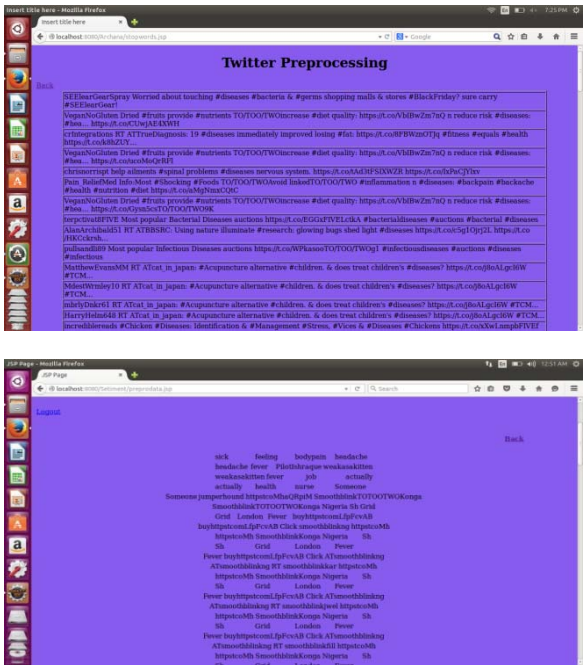

Figure-3. Twitter preprocessing.

4.3 SVM classification

Support Vector Machines are superintended learning models with associated learning algorithms that analyzed data's and recognize patterns, used for classifying and regressing the analysis. Support Vector Machines define boundaries based on decision plane studies. A decision plane is one that separates between a set of objects having various type of class memberships. A schematic example: drugs and diseases. After the Preprocessing the tweets are classified into diseases and drugs related tweets. The words are identified based on the keywords to classify the tweets. This lexicon analysis technique is used to find out the preferred category from the large number of tweets.

4.4 Polarity prediction

The classified tweets are analyzed based on polarity of the words like good, bad, not, un etc. Based on the polarity the number of positive tweets and negative tweets are identified. We are using the SVM classifier for classification technique for finding the polarity of the tweets and comments like positive tweets, negative, mixed or neutral.

Figure-4. Polarity calculation.

5. RESULT ANALYSIS

In the proposed system, the system could connect with Twitter by using Twitter application details through Twitter API. Tweets are being extracted so quickly from the Twitter using Twitter4j. These tweets are pre-processed to remove stop words, replace the short form with full form and replace the emoticons with its corresponding meaning for the easy understanding for the users. It classified and analyze the best medicine using polarity.

The input for the system should be given in Twitter Search Space to extract the tweets from the Twitter using Twitter API and Twitter application. Input can be drugs or diseases so that the system will extract the tweets based on the given input. The tweets will be displayed in the table format in a few seconds. The table will contain the username and their posted tweets.

The system is analyzed before preprocessing and after preprocessing. In some cases the number of words in tweets before preprocessing will be more than number of words after preprocessing. In some cases the number of words in tweets before preprocessing will be same as the number of words after preprocessing. In some cases the number of words before preprocessing is less than the number of words after preprocessing. All these cases is based on removal of stop words like "of", "the"etc, replace for short form with full words. eg "u" is replaced with "you" and replace the emoticons with its corresponding meaning. Eg "%-) "is replaced with its meaning "Confused or merry".

Test case:

Test case name	Input	Expect value	Actual value	Result
Registration	Username and password	Username and password	Empty without entering Username and password	Error
Login	Username and password	Username and password	Incorrect Username or password	Not a valid name or password
Twitter Connectivity		Correct consumer key, secret key, Access token and token secret key	Incorrect key values and token values	Tweets will not be extracted

6. CONCLUSIONS

The proposed system for categorizing drugs based on polarity analysis of twitter data. The pre-processed tweets are stored in database. These pre-processed tweets are identified whether it is drug related tweets or disease related tweets using Support Vector Machine classification. The drugs can be predicted whether the posted drug is a best drug or not using polarity. By this, the user will gain knowledge about the best drugs.

REFERENCE

[1] Ming Yang *et al.* 2015. Filtering big data from social media - Building an early warning system for adverse drug reactions. Journal of Biomedical Informatics. 54, pp. 230-240.

[2] Andrei Yakushev and Sergey Mityagin. 2014. Social networks mining for analysis and modeling drugs usage. ICCS. 29: 2462-2471.

[3] Eleonora D'Andrea *et al*. 2015. Real-Time Detection of Traffic from Twitter Stream Analysis. IEEE Transactions on Intelligent Transportation Systems. 16: 4.

[4] Robert G. Fichman *et al.* 2015. The Role of Information Systems in Healthcare: Current Research and Future Trends. 22: 419-428, ISSN 1047-7047.eissn 1526-5536. 11.2203. 0419.

[5] Carissa Hilliard. 2012. Social Media for Healthcare: A Content Analysis of M.D. Anderson's Facebook Presence and its Contribution to Cancer Support Systems. 24 - The Elon Journal of Undergraduate Research in Communications. 3(1).

[6] Victor C. Cheng *et al*. 2014. Probabilistic Aspect Mining Model for Drug Reviews. IEEE Transactions on Knowledge and Data Engineering. 26(8).

[7] Joao Cunha. 2015. Health Twitter Big Data Management with Hadoop Framework. Procedia Computer Science. 64: 425-431.

[8] Bruno HCh Stricker *et al.* 2012. Detection, verification, and quantification of adverse drug reactions. BMJ. 329(7456): 44–47.

[9] A. Gopal. 2013. Enhanced Clustering of Technology Tweets. San Jose State University.

[10] Marco Viceconti, Peter Hunter, and Rod Hose. 2015. Big Data, Big Knowledge: Big Data for Personalized Healthcare. IEEE Journal of Biomedical and Health Information. 19(4).