EXPLORATIONS ON ENGINEERING LETTERS (EEL)
VOLUME 1, ISSUE 1 (2016):PP.180-184
SANA ACADEMIC PRESS

Categorization of Drugs Based on Polarityanalysis of Twitter Data

Packiamary.M [1], S.Brindha [2]
[1]Research Scholar, Dept. of Computer Science & Applications, St.Peter's University, Chennai.
[2]Asst.Prof. ,Dept. of Computer Science & Applications, St.Peter's University, Chennai
maryakilam@gmail.com

Abstract-- *In the recent years, 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 for patients to share experiences with their drug usage. In current scenarios and with available new technologies, twitter can be used effectively for gathering information rather than gathering information in traditional method.Twitter is a most popular online social networking service that enable user to share and gain knowledge. User can post only short messages at the most of 140 character called "tweets". 60% of doctors say social media provides high quality of care to patients. 90% of response is from 18 to 24 years of age said that they would trust medical information shared by others on their social media networks.With the popularity of social media, twitter is used as an important source of data for consumers to share their experience based on drugs and diseases. First the user has to register by giving username and password and then login. The drug and disease related tweets are extracted from twitter using API and web crawler based on the given input. The extracted tweets are preprocessed by removing stopwords, abbreviations and replacing emoticons.*

I. INTRODUCTION

Big data is high-volume, high-velocity, high veracity, high value and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, create, manage and process the data within a tolerable elapsed time.

II. EXISTING SYSTEM

In the existing system, the physicians and pharmacists post the messages related to new drugs which have released in the market recently in web forum. Even the consumers are allowed to post their experience after using the drugs. The system integrates both text and data mining techniques to automatically extract important text features from the posts first, and then classify the posts into positive/negative examples based on a few pre-identified ADR related posts Filtering mechanism is done using text classification. Using partially supervised classification and Latent Dirichlet Allocation modeling, the messages are being categorized. The LDA is a generative probabilistic model that uses a small number topics to describe a collection of documents and it effectively reduce the dimension of the

texts. This system assist Food and Drug Administration (FDA) in identifying ADR(Adverse Drug Reaction) messages on web forum and result can be used as early warning system.

ISSUES IN EXISTING SYSTEM

- Polarity analysis is not done.
- Performance of the classification is not good.
- Data-labeling process is very time consuming and costly.
- Partially supervised learning is not automatically processed.

III. PROPOSED SYSTEM

The proposed system uses Twitter to get the information and process on it. The information from the Twitter is extracted using crawing and Twitter API. The twitter API will crawl the tweets from twitter using twitter4j. Twitter4j will connect with Twitter using configured Application .Twitter application is configured based on the consumer key, secret key, Access token and token secret key. Using these keys and tokens the connection is established with Twitter and Twitter4j will extract the tweets and display to the user I the table format. These extracted tweets are then preprocessed by replacing the short form words with full form. Eg: “r” is replaced with “are” , “2” is replaced with “TO”. It also replace the emoticons with its respective meaning. Eg: #:-) means Smiling with a fur hat. It also remove the stop words form the extracted tweets. Eg: “the”, “of”. These preprocessed tweets are then stored in the database. The preprocessed tweets are further classified using SVM classification based upon the category. In this system it is classified based on drugs related tweets and diseases related tweets. Polarity detection is done by the keywords like good, bad etc. Based on the number of positive tweets and the number of negative tweets it analyse the best medicine. This system is very useful for the users to gain knowledge about the best medicine.

ADVANTAGES

- The Twitter information can be used effectively.
- Users will gain knowledge about the best medicines.
- It is time consuming, can get to know medicine through this system instead of searching a doctor.

IV. ARCHITECTURE

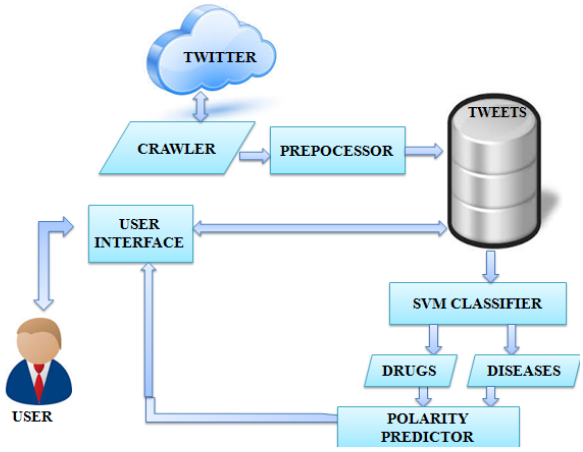

Fig 4.1: Architecture on polarity analysis of twitter data

Architecture on polarity analysis of twitter data is used to analyse the best medicine. Initially, the user should login into the search page by giving user name and password 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 preprocessed by removing stop words, replacing short form with full words and replace emoticons with their respective meaning. The preprocessed data is then stored in database. In future, the preprocessed 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.

V.EXPERIMENTAL METHODS

MODULES

1. Twitter Extraction
2. Preprocessing
3. SVM classification
4. Polarity prediction

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 TwitterFactory by using query search and have to maintain the query search results in QueryResult. Using getTweets method we can get the tweets, from which we can extract the tweet username.

PREPROCESSING

The extracted tweets are the preprocessed by removing stop words, short form and emoticons. All unmeaningful 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. Emoticons are known as smileys, there are varies kinds of smileys. For each smileys 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 emoticons. So, all the emoticons is replaced with their respective meaning.

SVM CLASSIFICATION

Support Vector Machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Support Vector

Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different 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.

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.

VI. 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 preprocessed 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. In future work, it is further classified and analyse 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”.

VII. CONCLUSION

The proposed system for categorizing drugs based on polarity analysis of twitter data. From the twitter developed application all the keys and token are generated, with these information we can connect the twitter with twitter API. The twitter tweets are extracted with twitter API using twitter4j.Then extracted tweets are preprocessed by removing stop words, short forms and emoticons. The preprocessed tweets are stored in database. These preprocessed 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.

REFERENCES

[1]. Ming Yang , Melody Kiang , Wei Shang , “Filtering big data from social media – Building an early warning system for adverse drug reactions” , Journal of Biomedical Informatics 54 (2015) 230–240.

[2]. HaewoonKwak, Changhyun Lee, Hosung Park, and Sue Moon, “What is Twitter, a Social Network or a News Media?”

[3]. EleonoraD'Andrea, PietroDucange, Beatrice Lazzerini, Member, IEEE, and Francesco Marcelloni, Member, IEEE, "Real-Time Detection of Traffic From Twitter Stream Analysis" , IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 16, NO. 4, AUGUST 2015
[4]. Robert G. Fichman, Rajiv Kohli, Ranjani Krishnan, "The Role of Information Systems in Healthcare: Current Research and Future Trends", Vol. 22, No. 3, September 2011, pp. 419–428 issn 1047-7047 .eissn 1526-5536 . 11 .2203 . 0419.
[5]. "TWITTER PROVIDES BIG DATA FOR ADVERSE DRUG EVENT IDENTIFICATION" REFER-HTTP://HEALTHITANALYTICS.COM/NEWS/TWITTER-PROVIDES-BIG-DATA-FOR-ADVERSE-DRUG-EVENT-IDENTIFICATION
[6]. K. Revathy, Dr. B. Sathiyabhama, "A Hybrid Approach for Supervised Twitter Sentiment Classification", International Journal of Computer Science and Business Informatics
[7]. A. GOPAL, "ENHANCED CLUSTERING OF TECHNOLOGY TWEETS," SAN JOSE STATE UNIVERSITY, 2013.
[8]. "TWITTER 'BIG DATA' CAN BE USED TO MONITOR HIV AND DRUG-RELATED BEHAVIOR, UCLA STUDY SHOWS" REFER-HTTP://NEWSROOM.UCLA.EDU/RELEASES/TWITTER-BIG-DATA-CAN-BE-USED-TO-250162
[9]. BALACHANDER KRISHNAMURTHY , PHILLIPA GILL , MARTIN ARLITT," A FEW CHIRPS ABOUT TWITTER".
[10]. Joao Cunha, Catarina Silva, MárioAntunes,"Health Twitter Big Data Management with Hadoop Framework", Procedia Computer Science 64 (2015) 425 – 431.
[11]. FarzindarAtefeh and WaelKhreich, "A Survey of techniques for event detection in Twitter",Computational Intelligence, Volume 31, Number 1, 2015.
[12]. Shamanth Kumar, Fred Morstatter, Huan Liu," Twitter Data Analytics", August 19, 2013.
[13]. Robert Leaman, Laura Wojtulewicz, Ryan Sullivan Annie Skariah, Jian Yang, Graciela Gonzalez, "Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks".
[14]. Marco Viceconti, Peter Hunter, and Rod Hose," Big Data, Big Knowledge: Big Data for Personalized Healthcare", IEEE Journal of Biomedical and Health Information,Vol.19,No.4,July 2015.
[15]. Michael J. Paul and Mark Dredze," You Are What You Tweet: Analyzing Twitter for Public Health".
[16]. NugrohoDwiPrasetyo ,Claudia Hauff, Dong Nguyen ,Tijs van den Broek , DjoerdHiemstra," On the Impact of Twitter-based Health Campaigns: A Cross-CountryAnalysis of Movember", Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis, pages 55–63.
[17]. Ahmed Abbasi and Donald Adjeroh," Social Media Analytics for Smart Health", Published by the IEEE Computer Society.
[18]. Xiao Liu, sinchunChen , "Identifying Adverse Drug Events from Patient Social Media", Published by the IEEE Computer Society
[19]. Andrei Yakushev and Sergey Mityagin," Social networks mining for analysis and modeling drugs usage", Volume 29, 2014, Pages 2462–2471 ICCS 2014. 14th International Conference on Computational Science.
[20]. Victor C. Cheng, C.H.C. Leung, Jiming Liuand Alfredo Milani, "Probabilistic Aspect Mining Model for Drug Reviews", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 8, AUGUST 2014
[21]. Adam Sadilek, Henry Kautz, Vincent Silenzio, "Modeling Spread of Disease from Social Interactions", 2012, Association for the Advancement of Artificial Intelligence.