RESEARCH ARTICLE OPEN ACCESS

Categorization of Drugs Using SVM Classification

Asst.Prof.J.Omana [1], S.Dhanalakshmi [2], V.M.Divyalakshmi [3], S.Mahalakshmi [4]

Assistant Professor [1], Student [2], [3] & [4]
Department Of Information Technology
Prathyusha Engineering College, Chennai
Tamil Nadu - India

ABSTRACT

In recent years, social media have emerged as major platform for sharing information in medical field. Twitter is used as a prominent social media to share their experience based on drugs and disease. These experiences are mined, extracted and analyzed can be converted into drugs or service improvement based on observations derived from user behavior. The medical information shared by users on the social media is valuable and trustable. We focused on analyzing the drugs related tweets using polarity. The drugs and disease related tweets are extracted from twitter using twitter4j API. The collected tweets are classified using svm classification and polarity. The final solution here is analysis of drugs and disease related tweets based on the svm classification

Keywords:- KDD, LDA

I. INTRODUCTION

Data Mining is popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in the databases. It is based on Exploration & analysis, by automatic or semi-automatic means of large quantities of data in order to discover meaningful patterns. Large amount of data in various field like Logistics, Financial, Health data, Social media, Scientific data is analysed. Data Mining is used to analyze social media information. Social media has made a tremendous change in normal life of a person in the society. There are various social networking website such as Facebook, Twitter, ScienceStage, SocialVibe, ShareTheMusic etc. The social network information can be used very usefully. Twitter is most popular social networking website. When the user post tweets, those tweets are publicly visible by default, but one can restrict message delivery to just their followers. Users can do group posts by topic or type by use of hashtags – words or phrases prefixed with a "#" sign. Similarly, the "@" sign is followed by a username is used for mentioning or replying to other users. Today, Twitter being a social media could be a useful measure of public awareness and reaction to disease and drugs information released by users. With the popularity of social media, twitter has become an popular source of data for consumers to share their experience based on drugs and diseases.

User can interact the system through user interface. 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

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.

III. DISADVANTAGES 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.

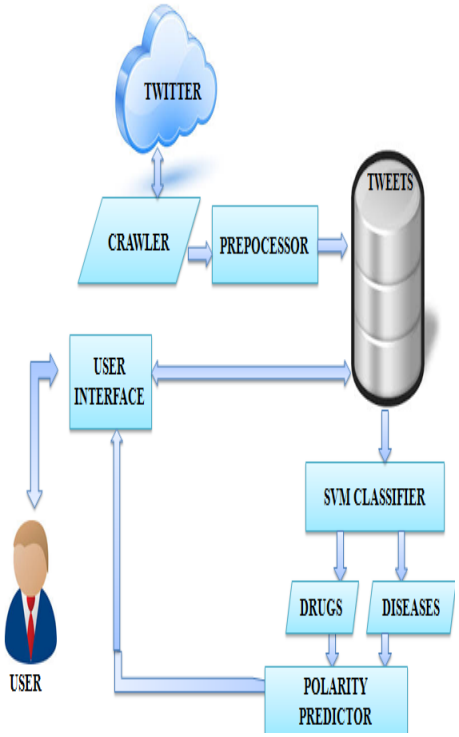
IV. PROPOSED SYSTEM

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 excel or csv format. 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.

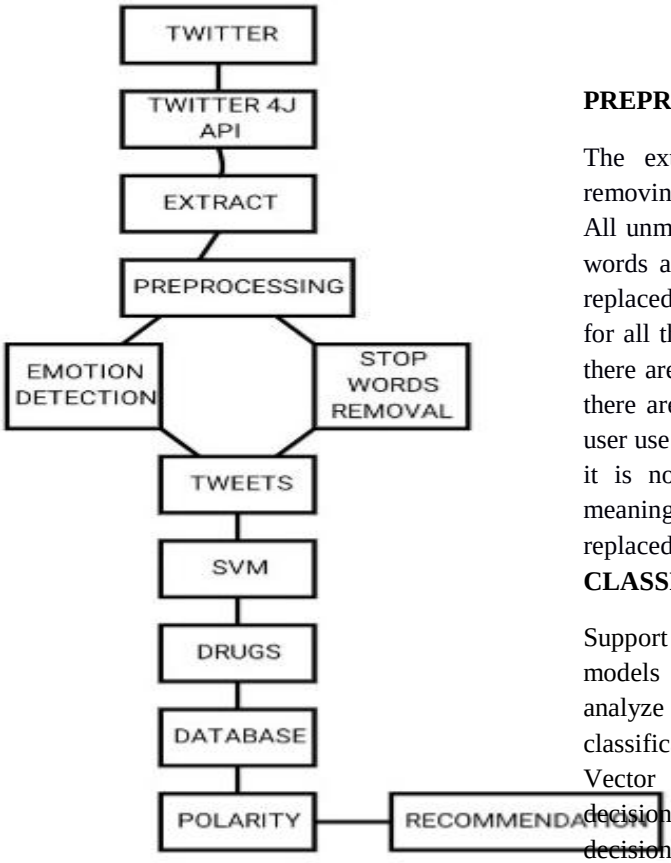
V. ADVANTAGES

- 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.

VI. ARCHITECTURE



DATA FLOW:



VII. MODULE DESCRIPTION

EXTRACTION OF TWEETS:

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 the R studio and r tool we will 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.

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 ANALYSIS:

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.

VIII. CONCLUSION

In this paper 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] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon, "What is Twitter, a Social Network or a News Media?"

[3] Eleonora D'Andrea, Pietro Ducange, 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] Bruno HCh Stricker, Bruce M Psaty, "Detec**tion, verification, and** quantification of adverse drug reactions", BMJ. 2004 Jul 3; 329(7456): 44–47.

[8] 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".

[9] Michael J. Paul and Mark Dredze," You Are What You Tweet: Analyzing Twitter for Public Health".

[10] Nugroho Dwi Prasetyo ,Claudia Hauff, Dong Nguyen ,Tijs van den Broek , Djoerd Hiemstra," 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.