ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 4, April 2017

Polarity based Analytics from Social Media to Mitigate Adverse Drug Reactions

[1] S.Kavishree [2] Dr.P.Shanmugapriya [3] R.Saravanan
[1].PG student, [Associate Professor Assistant Professor
Dept of IT, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya University Enathur Kanchipuram

Abstract:-- **In the recent years, social media have been emerged as major platforms for sharing information in the medical field, business, education etc. The existing system will generate a warning for adverse drugs reactions based on the negative comments from forums. The use of Latent Dirichlet Allocation modelling(LDA) in the existing system makes the data labelling process time-consuming and the polarity analysis is not done. But the proposed system uses Twitter to get the information and process on it. The information from the Twitter is extracted using Twitter API. Pre-processed tweets are stored in the database and those tweets are identified and classified whether it is based on drugs related tweets and diseases related tweets using Support Vector Machine classification(SVM). The user keywords can be predicted whether it is the best suggestion using polarity. Polarity detection is done by the keywords. Based on the number of positive tweets and the number of negative tweets it analyzes the best medicine. This system is very useful for the users to gain knowledge of the medicine.**

***Keywords*— Big Data; Drugs; Polarity; Preprocessing; Support Vector Machines; Twitter; Twitter API;**

I. INTRODUCTION

Big data is a term that describes a huge volume of data which includes structured, semi-structured and unstructured data. Mostly 90% of big data contains unstructured data. Big data can be analyzed for insights that lead to better decisions, efficiency and to reduce the risk. A Large amount of data from various fields like Logistics, Finance, Education, Social media, Science etc. can be analyzed. There are various social networking websites such as Facebook, Twitter etc., which can be used to analyze the data.

Twitter is one of the largest micro-blogging and social networking site which allows the user to post tweets such as real-time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life.

The Tweets can be analyzed to assess the drugs. Although rigorous clinical studies are required before a drug is placed on the market, it is impossible to predict all side effects for the approved medication. Adverse drug Reactions (ADRs) are injuries resulting from drug-related medical events. Drugs can cure diseases and save lives, yet, they may also cause different ADRs, and in some serious cases, threaten people's health. However, some side effects might not be revealed during this stage due to the limited size of clinical trials[9]. The user opinion about drugs and feedback from consumers of a drug can be obtained from the tweets, polarity prediction and support vector machine classification is applied to assess the drugs in an efficient manner.

II. REALTED WORKS

Altug Akay, Andrei Dragomir, Björn-Erik Erlandsson[1] presented a self-organizing map to analyze word frequency data derived from users forum posts and a novel network- based approach for modeling users forum interactions and employed a network partitioning method based on optimizing a stability quality measure. The drawback is that the customer opinions are analysed only from a certain forums.

Chih-Hua Tai, Zheng-Han Tan, Yue-Shan Chang[2] proposed a Feeling Distinguisher system based on supervised Latent Dirichlet Allocation (sLDA), Latent Dirichlet Allocation, and SentiWordNet methodologies for detecting a person's intention and intensity of feelings through the analysis of his/her online posts. But the intentions are grouped based on inter-related topics only.

Dingcheng Li, Cui Tao, Hongfang Liu[3] presented that Latent Dirichlet Allocations (LDA) can integrated into MapReduce framework.As a non-parametric Bayesian model, it integrates Hidden Markov Model LDA (HMM-LDA). A MapReduce based variational method is employed to do parameter estimation and inferences. The downside is

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 4, April 2017

that the classification of topics is based on co-referring events.

Julio Cesar Louzada Pinto, Tijani Chahed[4] introduced a framework to model information diffusion based on linear multivariate Hawkes processes and the latent Dirichlet allocation topic model. An estimation algorithm based on nonnegative matrix factorization techniques is used, and a modified variational Bayes method to identify data-driven estimation of hidden influences in social networks. The dis-advantage is that Inter-relationship between topics are analysed but the polarity of each individual topic are not identified.

Satya Katragadda, Harika Karnati, Murali Pusala[5] presented a method to identify adverse drug effects from tweets by modelling it, as a link classification problem in graphs. A link classification model is then used to identify negative edges i.e. adverse drug effects. But the negative comments are only considered.

Sungrae Park, Doosup Choi, Wonsung Lee[6] presented a method based on latent Dirichlet allocation, which is one of the most popular topic models, as the baseline model.. Use of topic network analysis, these specific patterns were proved using centrality measurements. The drawback is that the diseases and their medicines are classified based on specific patterns.

Xiao Xu, Tao Jin, Zhijie Wei[7] used the Latent Dirichlet allocation to cluster items without specifying the topic number. The low-frequency topics are pruned. Finally, fuzzy mining method is applied on these topic sequences. The dis-advantage is treatments are only categorized based diseases, the drugs are not suggested.

Xiaoping Zhang, Xuezhong Zhou , Houkuan Huang[8] proposed a topic model for text analysis and information retrieval by extracting latent topics from text collection. one of the extensions of hierarchical latent Dirichlet allocation model (hLDA) and Link latent Dirichlet allocation (LinkLDA) is used to automatically extract the hierarchical latent topic structures with both symptoms and their corresponding herbs. The drawback is polarity analysis is not done for the usage of corresponding herbs.

Yi-Yu Hsu, Hung-Yu Kao[10] proposed a method based on Named Entity Recognition tools and conditional random fields (CRFs) , to predict chemical and disease mentions in the articles. Finally, action term mentions were collected by latent Dirichlet allocation (LDA). The dis-advantage is the drugs used to cure diseases is only predicted, but polarity of drugs are not stated

Yuji Zhang, Dingchen Li, Cui Tao[11] presented a method based on the Latent Dirichlet Allocation (LDA) to discover topics based on associations and framework through the construction of disease-specific networks, an NLP-generated association database, followed by the analysis of network properties, such as hub nodes and degree distribution. The dis-advantage is the diseases are their drugs are grouped according to specific patterns only.

III. PROPOSED SYSTEM

The proposed system uses Twitter to get the information and process on it. The information from the Twitter is extracted using crawling and Twitter API. The twitter API will crawl the tweets from twitter using Twitter4j. These extracted tweets are then pre-processed by replacing the short form words with full form. It also removes the stop words form the extracted tweets. pre-processed tweets are then stored in the database. The pre-processed tweets are further classified using SVM classification based upon the category. In this system, it is classified user keywords 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 analyses the best suggestion. This system is very useful for the users to gain knowledge about the best suggestion.

A. 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 the twitter account using the twitter API called twitter4j. From the Twitter developer 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

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 4, April 2017

parameters to extract and read from the Twitter by using query search and have to maintain the query search results in the database.

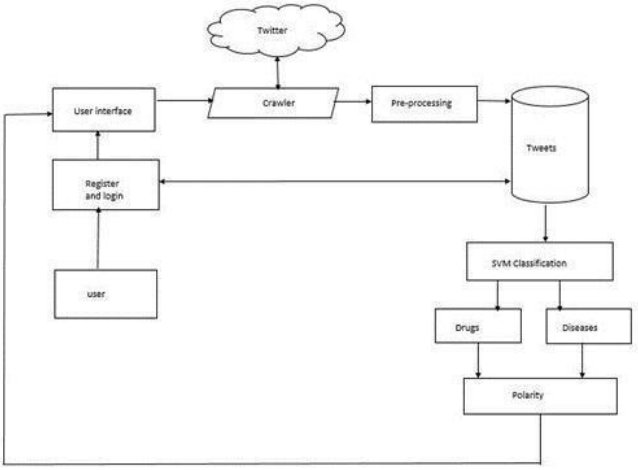

Fig. 1 System Architecture

B. Pre-Processing

The extracted tweets are the pre-processed by removing stop words, short form and emoticons.

1) Stop Words Removal: Stop words are words which are filtered out prior to, or after, processing of natural language data. For example, words like a, an, the, and, or, before, but, while and soon which do not contribute to classify the reviews. These words are removed from the dataset so as to avoid using them as features.

Table 1: Lexicons Of Social Acronyms Or Slangs

Acronym	Corresponding Word	Polarity
Y	Why?	Neutral
N	And	Neutral
GOT	Go To Hell	Negative
Gud	Good	Positive

2) Emotion Detection: 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.

Table 2: Lexicon Of Emoticons

S.no	Classification	Emoticons	Polarity
1.	Smile/Happiness	:)	Positive
2	Sadness	:(	negative
3	Teasing	:P	positive
4	Kiss	:*	positive
5	Disappointment	:/	negative
6	Love/Heart	<3	positive
7	Wink	;)	positive
8	Cry	:'(	negative
9	Anger	>:(	negative
10	Confused	o.O	negative
13	Thumbs Up	(y)	Positive

C. SVM Classification

Support Vector Machines are supervised learning models with associated learning algorithms that analyse 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 Pre-processing 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.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 4, April 2017

Table 3: Lexicons For Interjections

Interjection	Polarity
Wow	Positive
No way!	Negative
Amazing!	Positive
Thank you!	Positive
Haha, hehe!	Positive

D. 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.

IV. CONCLUSION

In this paper, we have proposed a system for categorizing drugs based on polarity analysis of twitter data. This proposed system will give the solution whether the drug is good or not, it be very efficient for the users to improve and gain their knowledge about the drugs. Based on the users experience comments after their drug usage is very useful for others to consume the best drug for easy recover of their health.

REFERENCES

[1] Altug Akay, Andrei Dragomir, Björn-Erik Erlandsson ,"Network- Based Modeling and Intelligent Data Mining of Social Media for Improving Care",IEEE Journal of Biomedical and Health Informatics,Vol.19, Jan.2015

[2] Chih-Hua Tai, Zheng-Han Tan, Yue-Shan Chang,"Systematical Approach for Detecting the Intention and Intensity of Feelings on Social Network", IEEE Journa of Biomedical and Health Informatics,Vol.20, Jul.2016

[3] Dingcheng Li, Cui Tao, Hongfang Liu,"Ontology-Based Temporal Relation Modeling with MapReduce Latent Dirichlet Allocations for Big EHR Data",Cloud and Green Computing (CGC) Second International Conference,Nov.2012

[4] Julio Cesar Louzada Pinto, Tijani Chahed,"Modeling Multi-topic Information Diffusion in Social Networks Using Latent Dirichlet Allocation and Hawkes Processes",Signal-Image Technology and Internet-Based Systems (SITIS) Tenth International Conference, Nov.2014

[5] Satya Katragadda, Harika Karnati, Murali Pusala,"Detecting adverse drug effects using link classification on twitter data",Bioinformatics and Biomedicine (BIBM),IEEE International Conference,Nov.2015

[6] Sungrae Park, Doosup Choi, Wonsung Lee,"Disease-medicine topic model for prescription record mining",Systems, Man and Cybernetics (SMC) IEEE International Conference,Oct.2014

[7] Xiao Xu, Tao Jin, Zhijie Wei,"TCPM: Topic-Based Clinical Pathway Mining",Connected Health: Applications, Systems and Engineering Technologies (CHASE) IEEE First International Conference,Jun.2016

[8] Xiaoping Zhang, Xuezhong Zhou, Houkuan Huang,"A hierarchical symptom-herb topic model for analyzing traditional Chinese medicine clinical diabetic data",Biomedical Engineering and Informatics (BMEI), 3rd International Conference,Oct.2010

[9] Yang Peng, Melody Moh, Teng-Sheng Moh,"Efficient adverse drug event extraction using Twitter sentiment analysis",Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM International Conference,Aug.2016

[10] Yi-Yu Hsu, Hung-Yu Kao,"Curatable Named-Entity Recognition Using Semantic Relations",IEEE/ACM Transactions on Computational Biology and Bioinformatics,Vol.12,Jul.2015

[11] Yuji Zhang, Dingchen Li, Cui Tao,"An integrative computational approach to identify disease-specific networks from PubMed literature information",Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference,Dec.2013