ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Drug Side Effect Analyser Using Machine Learning

Adarsh Bhat G
Research Scholar
Department of Computer Science
Canara Engineering College
Mangalore, India

Abhilash Prajwal Dsouza
Research Scholar
Department of Computer Science
Canara Engineering College
Mangalore, India

Rajgopal K.T
Assistant Professor
Department of Computer Science
Canara Engineering College
Mangalore, India

Jayadeva Bhat K
Research Scholar
Department of Computer Science
Canara Engineering College
Mangalore, India

Mahesh
Research Scholar
Department of Computer Science
Canara Engineering College
Mangalore, India

Abstract: People are dependent on medicinal drugs on one way or the other for every simple cause such as headache, cold etc. Every drug has a negative impact on a person's body. Some people are unaware of the side effects of the drugs and they consume it without prescription. Social network platforms such as twitter provide an opportunity for people to express themselves. Using twitter as the source of data, this paper aims to find the side effects of drugs with the help of machine learning algorithms.SVM (Support Vector Machine) algorithm is used for drug related classification with an accuracy of 75%.Sentiment analysis is performed using VADER (Valence Aware Dictionary for sEntiment Reasoning) to handle negations, conjunctions and question marks present in the tweets. Keyword Extraction is performed using RAKE (Rapid Automatic Keyword Extraction) to get the side effects.

Keywords: -- Machine learning, Sentiment Analysis, Natural Language Processing, RAKE, SVM

I. INTRODUCTION

Medicinal drugs help in curing diseases but unfortunately there are also negative impacts on the human body. Pharmaceutical companies conduct a series of clinical experiments on drugs before releasing it in the market. These companies depend on clinical experiments to find the side effect of drugs. But some negative impacts may not be revealed because of less number of clinical experiments, and they come to light when the patients use it in a long run.

Twitter is an open platform where the users can share their views on the drugs with the public without complaining the manufacturers of the drugs. This collectively results in huge dataset which includes the views of different users towards the drugs. These views can be analyzed and can be used to extract the side effects based on the concept of Machine Learning.

In machine learning, the machine is trained with data using various algorithms. After training, the machine predicts the output for the unseen inputs. In this paper, supervised Machine Learning is used with twitter as the dataset.

The objective of this paper is to extract side effects. It includes the following steps.

a) Collect the tweets for the selected drugs from twitter.
b) Pre-process the tweets (Feature Engineering).
c) Building model for drug related classification of tweets.
d) Sentiment Analysis.
e) Extraction of side effects.

II. LITERATURE SURVEY

In the proposed method by Y. Peng, M. Moh and T. S. Moh [1], Hive is used to extract, transform and load (ETL) data. Further HiveOL sentences are used to extract tweets. All the tweets are merged into a single file for the data processing step. Data processing is done with the help of natural language toolkits and regex of python. Drug related classification and sentiment analysis is done with the help of WEKA. Finally adverse drug events are extracted with the help of Metamap.

The proposed method by Fan Yu, Melody Moh and Teng-Sheng Moh [2] aims at finding the tweets that contain side effects. The collection of tweets is done using tweepy. Further, the tweets are filtered and preprocessed. Finally classification is implemented using Supervised Machine Learning algorithms. The result is the tweets that contain side effects

In the system proposed by Liang Wu,Teng-Sheng Moh and Natlia khuri [3], the tweets are collected from the twitter using Twitter API and Tweepy.While collecting the tweets, drug names are used as search keyword in order to get drug

related tweets. The collected tweets are stored in the database MongoDB. Because the tweets are in JSON format which is supported by MongoDB, reformatting of tweets is not required. The Database is accessed by the PyMongo API. The Collected tweets are preprocessed using python's built in regular expression to remove the hashtags, emoticons, Uniform Resource Locator. Classifier is built by using features like Textual features, syntactic features, and sentimental features by the WEKA software. Side effects of the drugs are extracted by using Metamap.

In the proposed system by Bresso, E., Grisoni, R., Marchetti, G., Karaboga, A. S., Souchet, Devignes, M.-D., and Smaïl-Tabbone [4], the annotations are collected from drug databases and SIDER. Clustering is performed on the individual side effects reported in the SIDER ,with a semantic similarity measure into term clusters(TC).More frequent items are extracted from the resulting drug*TC binary table, which results in Side Effect Profiles(SEPs).SEP is the combination of TCs which are shared by the significant number of drugs. Frequent drugs are explored using machine learning algorithms.

The current proposed system is not only capable of classifying tweets into drug related and drug unrelated tweets, it is also capable of extracting the side effects from tweets which contains side effects. The system can also handle conjunctions or negations which may change the meaning of the tweets.

III. PROPOSED SYSTEM

The flow chart of entire proposed system is shown in the figure 1.

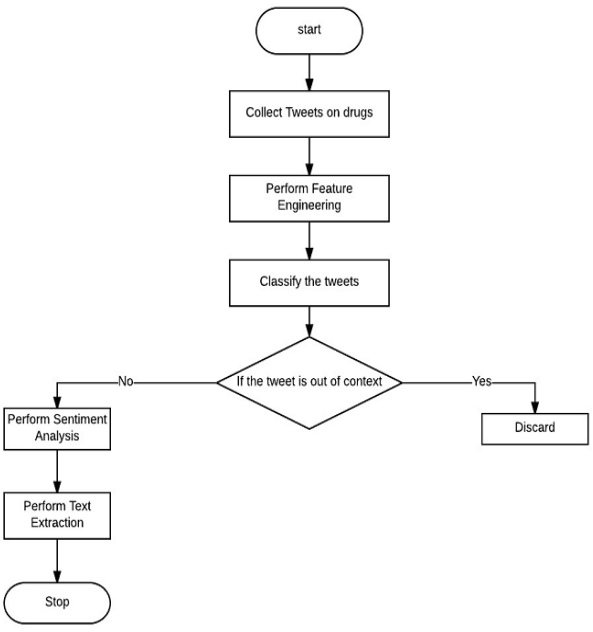

Figure 1: Proposed system flow chart

A. *Collect tweets from twitter*

In this project, twitter4j library is used along with twitter API to collect tweets. Twitter4j is an open library for twitter API. Tweets are searched by using drug names as keyword shown in table 1. These collected tweets are stored in mongoDB. Twitter does not allow users to access data which are older than 2 weeks. For this purpose the data harvesting was performed for 1 month.

Table 1: Drug list

#	***Drug name***	***Uses.***
1	Aspirin	treat pain, fever, and inflammation
2	Paracetamol	treat pain and fever
3	Gardasil	to prevent cervical/vaginal/anal cancers caused by certain types of HPV
4	Benadryl	Used to relieve symptoms of allergy, hay fever, and the common cold.
5	Ibuprofen	to reduce fever and treat pain
6	Metformin	Controls blood sugar levels
7	Zoloft	to treat depression, obsessive-compulsive disorder, panic disorder, anxiety disorders
8	Naproxen	To treat menstrual pain, muscle and joint inflammation,
9	Abilify	Used to treat schizophrenia and bipolar I disorder
10	Amoxicillin	To treat infections caused by bacteria such as tonsillitis, bronchitis, pneumonia

B. *Feature Engineering*

It is the preprocessing of the collected tweets. Here each tweet is read manually and labelled. Tweet is labelled as 1 if it contains side effect and labelled as 0 if it does not contain side effect. Label 0 tweets can include negative, positive or sarcastic tweets. For example “Aspirin increases heart attack” is labelled as 1 where as “Aspirin is good for nothing” is labelled 0 as it does not specify side effect and it is only a negative tweet. All these preprocessed tweets are used to build a machine learning model in the next stage for classification purpose.

C. *Classification of tweets*

This stage involves classification of tweets which contain side effects and tweets which do not contain side effects. Various algorithms are used, but Support Vector Machine (SVM) gives best accuracy with an efficiency of 75%. K-fold cross validation technique is used(k=5), where the train file is subdivided into k subsets and among them one subset is used as test file and remaining k-1 subsets are used as train file. The input to this module is tweets and output will be the label predicted.

D. *Text Extraction*

The output of classification stage whose label is equal to 1 is given as input to this stage. The idea is to extract side effect from the tweet. One of the challenges faced is that the drug related classification is not capable of detecting sentiment in the tweet. That means the tweet “Aspirin causes liver damage” and “Aspirin does not cause liver damage” will have label as 1. So to overcome this problem

sentiment analysis is performed before text extraction to increase efficiency. Sentiment analysis is done with the help of VADER [5,6] (Valence Aware Dictionary for sEntiment Reasoning).When VADER is used to detect sentiment it can detect negations like “not” and conjunctions like “but” and question marks present in the tweets. Also the probability of finding side effects in negative tweets is more. So sentiment of the tweets is detected to get better results.

Keyword Extraction [7,8] in the proposed system is implemented in python language using RAKE (Rapid Automatic Keyword Extraction).RAKE is set with a path to stop word list and some parameters. The first parameter is a path to stop word list, the second parameter indicates the minimum number of characters in a word, the third parameter indicates the maximum number of words that each phrase can have and the last parameter indicates the number of times the keyword can appear in the text.

IV. RESULTS

Initially, the tweets corresponding to 10 drugs are collected and stored in mongoDb [9]. Secondly, the Feature Engineering is done on the collected tweets which acts as a train dataset. Thirdly, classification algorithm called Support Vector Machine (SVM) is used to classify the new tweets into tweets that contain side effects and tweets that do not contain side effects. The output of the classification algorithm is label 0 for tweets which does not contain side effects and label 1 for tweets which contains side effects. Classification is done with the help of scikit learn [10]. Sentiment analysis is done to detect negation like “not” and conjunctions like “but” and question marks present in the tweets. Finally, the extraction of side effects is done with the help of RAKE.

SVM gave more Accuracy when compared to other classification algorithms. The comparison of the accuracy was achieved using Cross-fold Validation technique.

As seen from figure 2, figure 3 and table 2, SVM provides more accuracy that is, an average accuracy of 75% opposed to that of MultinomialNB.

Table 2: Accuracy comparison

Classifier	***Accuracy (%)***
SVM	75
MultinomialNB	73

The output of the classification is in terms of 0's and 1's.The output is 1 if a tweet contains side effects and 0 if it doesn't. The tweets are selected only if there are side effects. If the keyword or a drug name in a tweet is a name of place or a person or if a tweet is negative and there are no side effects or if a tweet is positive, then these tweets are not selected. The output of classification is then passed to VADER which performs the sentiment analysis to exclude the tweets that contain side effects but are positive because of a prefix that precedes a side effect. The Keyword Extraction is done using RAKE which extracts only the side effects from the negative tweets that were obtained during sentiment analysis.

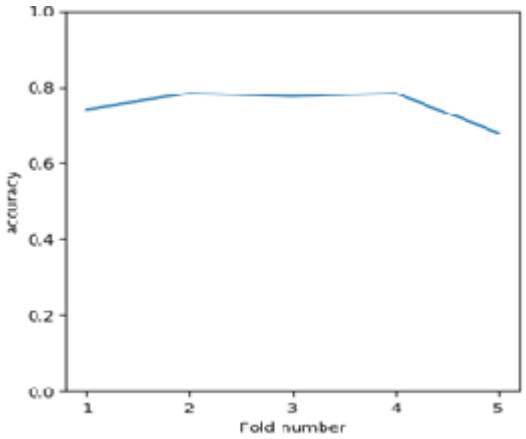

Figure 2: SVM accuracy graph

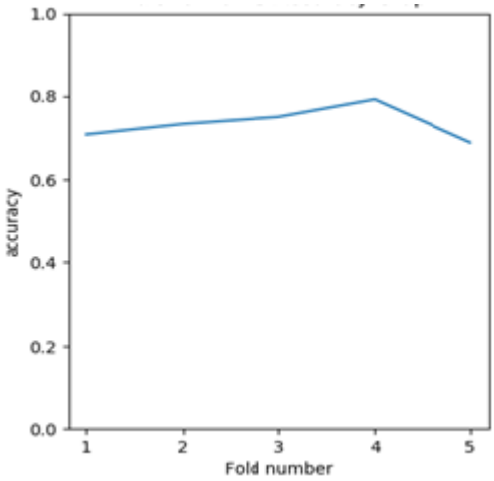

Figure 3: MultinimialNB accuracy graph

V. CONCLUSION

The proposed system extracts the side effects of 10 drugs using the concept of machine learning where SVM is used for classification and RAKE is used for keyword extraction. This approach can be used in future for any other products to analyze the reviews that are given by the users. This product works only for 10 drugs but this can be implemented for more number of drugs. This product processes only English tweets but enhancements can be done for processing tweets in other languages

VI. REFERENCES

[1] Y. Peng, M. Moh and T. S. Moh, "Efficient adverse drug event extraction using Twitter sentiment analysis," 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, 2016, pp. 1011-1018.doi: 10.1109/ASONAM.2016.77

[2] F. Yu, M. Moh and T. S. Moh, "Towards Extracting Drug-Effect Relation from Twitter: A Supervised Learning Approach," 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), New York, NY, 2016, pp. 339-344

[3] L. Wu, T. S. Moh and N. Khuri, "Twitter opinion mining for adverse drug reactions," 2015 IEEE International Conference

on Big Data (Big Data), Santa Clara, CA, 2015, pp. 1570-1574

[4] Bresso, E., Grisoni, R., Marchetti, G., Karaboga, A. S., Souchet, Devignes, M.-D., & Smaïl-Tabbone, M. (2013). Integrative relational machine-learning for understanding drug side-effect profiles. *BMC Bioinformatics*, *14*, 207.

[5] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[6] NLTK. http://www.nltk.org/

[7] Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. "Automatic keyword extraction from individual documents." *Text Mining* (2010): 1-20.

[8] Dostal, Martin, and Karel Jezek. "Automatic Keyphrase Extraction based on NLP and Statistical Methods." In *DATESO*, pp. 140-145. 2011.

[9] mongoDB. https://www.mongodb.com/

[10] scikit-learn. http://scikit-learn.org/stable/