Grenze International Journal of Engineering and Technology, Jan Issue

Analysis of Drugs using Machine Learning

Rushikesh Gudade[1], Pratik Jangam[2], Atharva Bongirwar[3], Abhishek Giri[4] and Dr.Ram Joshi[5]
[1-5]JSPM'S Rajarshi Shahu College of Engineering, Pune, MH-India, Department of Information Technology
Email: pratikjangam57@gmail.com

***Abstract*—Peoples are dependent on medicinal drugs for their day-to-day life. People depend on drugs for very small reasons like headaches, colds etc. Many drugs have its own side effect on a person's body. Large number of people consume drugs without physician prescription and get affected. To avoid such problems drug review is very important and useful for providing vital medical care information. This information is very useful not only for healthcare professionals but also for consumers. Customers are using online review to tell their experiences and opinion on the drugs that they have experienced. However, many times who want to buy that medicine find it very hard to go through all these comments before buying a medicine. Predicting side effects is an necessary topic to create new drugs. Many machine learning algorithm are able to predict side effects but yet there is always space for improvement. Firstly, we will adapt the multi-label learning technique as side effect prediction is a multi-label learning technique. Because of these reasons, this paper primarily helps to classify at what level drugs can effect human body by using different machine learning models.**

***Index Terms*— Classification algorithm, drugs side effects, machine learning.**

I. INTRODUCTION

Medical drugs are very helpful for curing diseases but unfortunately use of these drugs in long term can cause a negative impact on the human body and these drugs can affect their health. While creating drugs pharmaceutical companies conduct some experiments on these drugs before releasing them for sale. These companies do clinical experiments on drugs and depend on the results of that experiments for finding side effects. These experiments are in limited numbers so the negative impact may not be revealed that early sometime it takes time to come in front of all. When person use drugs in long term then we got to know about the adverse effects of the drug [7]. As per the research, drug toxicity is one of the leading cause of death in the world so Adverse effect or side effect of drugs is the most important and difficult challenge for researchers[6]. For accelerating the drug development process and improving patients' health effective ADR prediction is essential. Research and development team of farms industry rigorously working and investigating towards designing and bettering performance. The demand for new and improved drugs which are effective and efficient is always high, but the major problem is often due to the off-target effect of drugs [8]. In traditional methods, testing of drugs is done on animals and the side effects of drugs have been observed and recorded in pre-clinical trials. But the result came after testing on animals is not generalised across all the species. In this theory, we are doing experiments with different side effects that are caused by drugs that are used by a pharmaceutical company. Mainly diseases such as dizziness, headache, rashes, diarrhoea, dermatitis, vomiting, nausea, weakness, abdominal pain, and chronic fatigue. From these side effects which side effect has the highest variance selected for doing the study. CHEMBL, SIDER and Twitter are the public data repositories and the data got collected from there. These

Grenze ID: 01.GIJET.9.1.541_2

databases are free and bioinformatics resources of information. Information that are range from drug indications contained in CHEMBL which is the bioactive database and information like side-effect frequency, classification, and drug target relation is present in SIDER. And Twitter where user can share their review opinions on the side effect faced by them. By analysing these views by using machine learning algorithms we can find some inferences.

II. Review of Literature

M. Moh , Y. Peng and T. S. Moh proposed a model in which they extracted the data then after transformation they load the data . In next step they extracted tweets from sentences and merge them to single file for next step that is pre-processing. Using regex and natural language toolkit used for data pre-processing .
Classification of drug and sentiment analysis from tweets done with WEKA. Melody Moh, Fan Yu and TengSheng Moh model mainly focused on to find tweets which are included with side effect. Tweepy is used to collect tweet then these tweets are pre-processed and filtered. Then supervised algorithm are used for classification . This classification is with output of tweets included with side effect. Teng-Sheng Moh, Liang Wu and Natlia khuri proposed a model in which using twitter API , Tweepy tweets are collected.
tweets are collected by drug name as a keyword to collect tweet. MongoDB database is used to store collected tweets the main reason behind this is tweets were is JSON format. PyMongo API is used to access the database then tweets are preprocessed using python techniques regex then classification is done by WEKA.
Metamap is used to extract side effect of drug. Grisoni ,G Karaboga, A. S. Souchet are proposed a system in which they had collected data from SIDER and drug database. The side effect which are reported , clustering is performed on them .More frequently occurring item are extracted which result in side effect of drug. Most frequently drug are find by different machine learning algorithm.
The flow chart is shown in the following fig 1.
Many features which are related to drugs can be get from databases which is available on many platforms associated with side effects. In this project collected review of different drug for different condition as shown in table. Four datasets were collected and used to evaluate the performance of the proposed approach for predicting adverse drug reactions. The first (named-Record-A) had FDA approval obtained from the drug bank using the chemical sub structure defined in PubChem and the corresponding adverse drug reaction obtained from sider. Of small molecule drug were included [2-4].
In another model collecting tweets from twitter, feature engineering, classification of drugs text extraction is the four main sections of the following study (proposed system) twitter4j open library is used and tweets are extracted by drug name as keyword. keyword extractions using RAKE which is present in python. pre-processing performed on collected tweets tweet with side effect is labelled as 1 and without side effects labelled as 0(Bhat G et al, 2017).
In a next stage classification of tweets with or without side-effect are performed. classification of tweets is done by support vector machine (SVM) output of classification with label 1 given as input to text extraction stage SVM provides accuracy of 75% compared to other classification algorithm. As seen from figure 2 and figure 3 SVM provides more accuracy of 75% opposed to that of MultinomialNB [1].
In different model dataset contain reviews of patient about drug applied machine learning model using linear SVM for classification task. They have performed pre-processing by lemmatization all words and lower-casing. Methods are tested on 5-folds cross validation, and they have improved the results on computed macro-averaged recall (R), F1-measures (F) and precision (P). table 1 presents classification result for 4 classes for different sets of features. 1.5% for binary and 2.1% for multi-class classification tasks.
In a developed model of prediction drug side effect data is collected from Chembl database which contains marketed medicines and their bad drug reactions through which multi-label KNN method and FS-MLKNN prediction mode are the main steps involve in this proposed model. Design the feature which is selection-based and multi-label using KNN method to find the optimal feature and build multi-label prediction models. Now combine two algorithms, MLKNN and genetic algorithm use to select the subset of optimal dimension. Now develop multi-label classification model[13]. FS-MLKNN produced high-accuracy prediction and the results which can help to reveal reasons of side effects of drugs.

III. Conclusion

The proposed system focuses on drug side effects using machine learning concepts. In this paper, we took the first steps towards the extraction of drug responses and diseases. we have proposed a web application which will

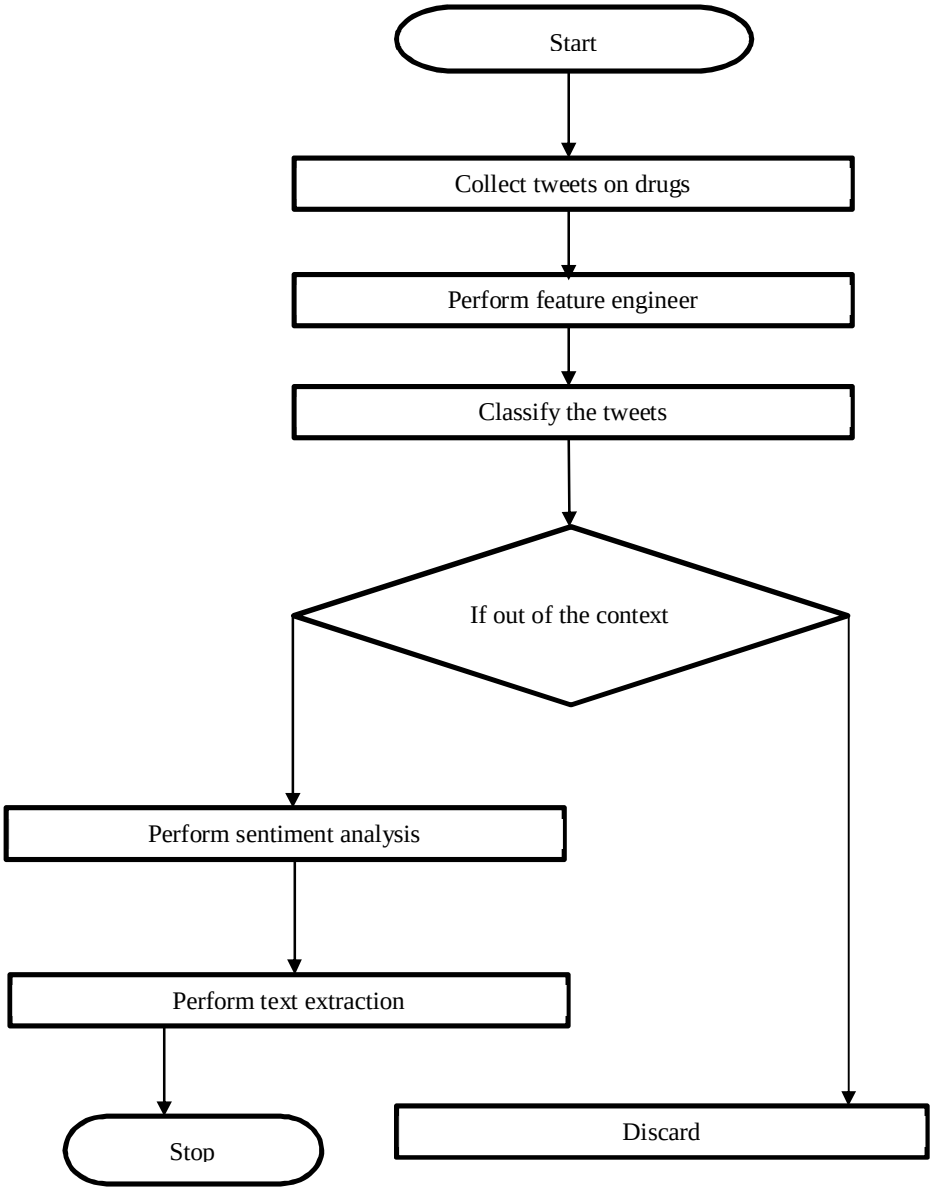

Fig 1: Proposed System flow chart

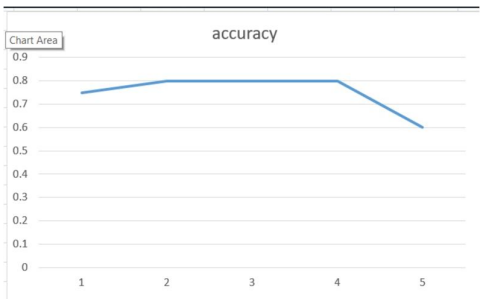

Fig 2: SVM accuracy graph

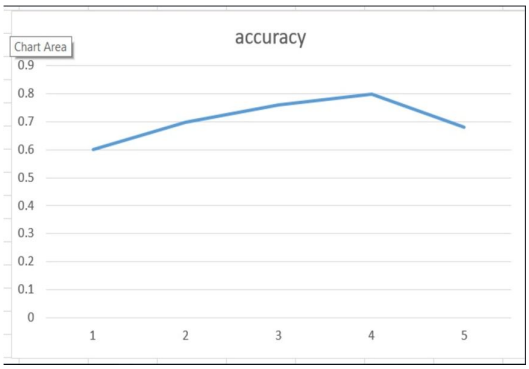

Fig 3: MultinomialNB accuracy graph

TABLE I: 5-FOLD CROSS VALIDATION

Features	Folds (F_1-measure)					Av. F
	1	2	3	4	5	
bow	.568	.543	.537	.569	.502	
Bow, pos	.564	.550	.542	.560	.509	.544
dis, Bow	.556	.550	.550	.586	.524	.550
w2v, Bow	.574	.553	.551	.551	**.551**	.559
Pol, Bow	.566	.545	.538	.538	.511	.547
Sent, Bow	.567	.546	.537	.569	.502	.545
Drug, Bow	.565	.543	.538	.570	.501	.544
Pmi, Bow	.555	.532	.535	.564	.508	.538
Bow, w2v, dis,	.556	.553	.542	.586	.541	.556
Bow, pos, dis, w2v	.581	.545	.551	.589	.522	.558
Bow, pos, w2v	**.588**	.553	.547	.566	.530	.557
Bow, dis, w2v, pos	.564	.557	.547	.590	.540	.564
Bow, w2v, pmi, pos	.562	**.558**	**.558**	**.594**	**.551**	**.565**

detect a side effect of drug. side effect of a particular drug modelled by machine learning algorithm with this web application we able to find out from which particular age group consumed a particular drug more. presented dataset is of user reviews about the classification method based on machine learning and medications. We have improved the results for multi-class and binary classification tasks.

REFERENCES

[1] Adarsh Bhat G, Abhilash Prajwal Dsouza, Rajgopal K.T, Jayadeva Bhat K, Mahesh, "Drug Side Effect Analyser Using Machine Learning", Adarsh Bhat G et al, International Journal of Advanced Research in Computer Science, 8 (3), March-April 2017,835-838

[2] Md. Deloar Hossain, Md. Shafiul Azam, Md Jahan Ali, Hakilo Sabit, "Drugs Rating Generation and Recommendation from Sentiment Analysis of Drug Reviews using Machine Learning ", 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE) | 978-1-6654-1962-8 /20/$31.00 ©2020 IEEE | DOI: 10.1109/ETCCE51779.2020.9350868

[3] Sathien Hunta, Nattapol Aunsri, Thongchai Yooyativong, "Drug-Drug Interactions Prediction from Enzyme Action Crossing Through Machine Learning Approaches", 978-1-4799-7961-5/15/$31.00 ©2015 IEEE

[4] JIN LI , XIANGMIN JI, LIYAN HUA, "Improving the Prediction of Adverse Drug Events Using Feature Fusion-Based Predictive Network Models", March 2, 2020, Digital Object Identifier 10.1109/ACCESS.2020.2979452

[5] Priyanka Nath, Sumran Kilam, Aleena Swetapadma, "A Machine Learning Approach to Predict Volatile Substance Abuse for Drug Risk Analysis", 978-1-5386-1931-5/17/$31.00 ©2017 IEEE

[6] Behnaz Pourebrahim, Mohammadreza Keyvanpour, "Adverse Drug Reaction Detection Using Data Mining Techniques: A Review Article", (ICCKE2020) October 29-30, 2020 978-1-7281-8566-8/20/$31.00 ©2020 IEEE

[7] Abdelrahman I. Saad , Yasser M. K. Omar, And Fahima A. MaghRABY, "Predicting Drug Interaction With Adenosine Receptors Using Machine Learning and SMOTE Techniques ", Received September 20, 2019, accepted October 5, 2019, date of publication October 9, 2019, date of current version October 22, 2019. Digital Object Identifier 10.1109/ACCESS.2019.2946314

[8] JIANXIANG WEI, ZHIQIANG LU, KAI QIU, PENGYANG LI, AND HAOFEI SUN, “Predicting Drug Risk Level from Adverse Drug Reactions Using SMOTE and Machine Learning Approaches”, Received September 17, 2020, accepted October 5, 2020, date of publication October 7, 2020, date of current version October 21, 2020. Digital Object Identifier 10.1109/ACCESS.2020.3029446

[9] WEI-PO LEE, JHIH-YUAN HUANG, HSUAN-HAO CHANG, KING-TEH LEE, AND CHAO-TI LAI, “Predicting Drug Side Effects Using Data Analytics and the Integration of Multiple Data Sources”, Received July 31, 2017, accepted September 2, 2017, date of publication September 21, 2017, date of current version October 25, 2017. Digital Object Identifier 10.1109/ACCESS.2017.2755045

[10] ABDELRAHMAN I. SAAD , YASSER M. K. OMAR, AND FAHIMA A. MAGHRABY, “Predicting Drug Interaction With Adenosine Receptors Using Machine Learning and SMOTE Techniques”, Received September 20, 2019, accepted October 5, 2019, date of publication October 9, 2019, date of current version October 22, 2019. Digital Object Identifier 10.1109/ACCESS.2019.2946314

[11] G. Shobana, Dr. S. Nikkath Bushra, “Drug Administration Route Classification using Machine Learning Models”, 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) | 978-1-7281-7089-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICISS49785.2020.93159

[12] Muhammad Irfan Khan, “Drug side-effect prediction using machine learning methods”, Received July 31,2017 , accepted November 24, 2017

[13] Wen Zhang, Feng Liu , Longqiang Luo , Jingxia Zhang,” Predicting drug side effects by multi-label learning and ensemble learning”, BMC Bioinformatics (2015) 16:365

[14] Wei-Po Lee , Jhih-Yuan Huang , Hsuan-Hao Chang , King-Teh Lee , And Chao-Ti Lai,” Predicting Drug Side Effects Using Data Analytics and the Integration of Multiple Data Sources”, Received July 31, 2017, accepted September 2, 2017, date of publication September 21, 2017, date of current version October 25, 2017. Digital Object Identifier 10.1109/ACCESS.2017.2755045