International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Analysis of Patient Condition Classification Methods Utilizing Drug Reviews

Mrs. Deepali Sanjay Chavan[1], Kiran Nitin Babar[2]

[1]Artificial Intelligence & Data Science & DMCE, Airoli.
[2]Artificial Intelligence & Data Science & DMCE, Airoli.

Abstract - Nowadays, a new area of marketing and communication has emerged thanks to internet reviews, bridging the gap between conventional word-of-mouth and a viral feedback loop that can sway consumers' perceptions. Reviews of specific medications, however, are much more important in the medical industry because they may be used to track side effects and determine how consumers feel about a medication overall. This paper's main goal is to use medication reviews to categorize patient conditions, specifically Type 2 Diabetes, High Blood Pressure, and Depression. The goal is to study and understand the effectiveness of drugs for specific conditions and their potential side effects by analysing patient reviews, ratings, and useful counts. The insights gained from this analysis can be used to recommend suitable drugs for patients based on their condition and the experiences of other patients with similar conditions.

***Key Words*:** TF-IDF, BOW, Passive Aggressive Classifier, stop word.

1. INTRODUCTION

The internet has now evolved to be a forum where consumers evaluate different products and services based on impressions and feedback from other like-minded consumers. In recent times, online reviews have created a new field in marketing and communication that bridges the gap between traditional word-of-mouth and a viral form of feedback that can influence consumer's opinions. A medication review can explain how and when your medications should be taken to get the most benefit out of them. However, in the field of medicine, reviews made for particular drugs play an even more vital role as they can help in monitoring its adverse reactions and identify an overall impression of the drug among its users. Reviews are very important to get the overview of product whether it is service, offerings or products. Reviews also plays a very important role in healthcare domain especially in terms of drugs. By analyzing the reviews, we can get the understanding of the drug effectiveness and its side effects. But in this project, we will classify the condition of patient based on his review so that we can recommend him a suitable drug. One of the key benefits of drug recommendation using AI is that it can help identify potential drug interactions and adverse reactions. Such systems analyze patient data, including medical history and current medications, in order to identify potential issues and recommend appropriate alternatives. Another benefit of AI-powered drug recommendation systems is that they can help reduce the risk of medication errors. By automating the process of medication selection, dosing, and administration, these systems can ensure that patients receive the right medication in the right dose at the right time. This can result in better patient outcomes and a decrease in healthcare costs. In proposed method a drug recommendation system using NLP and machine learning algorithms that will not only predict medical conditions but also recommend the top 3 drugs based on predicted medical conditions and top reviews and useful reviews count.

2.DATASET AND ITS FEATURES

The Drug Review Dataset is taken from the UCI Machine Learning Repository. This Dataset provides patient reviews on specific drugs along with related conditions and a 10-star patient rating reflecting the overall patient satisfaction. The data was obtained by crawling online pharmaceutical review sites. The Drug Review Data Set is of shape (215063, 7) i.e. It has 7 features including the review and 215,063 Data Points or entries. Features are

- DrugName (categorical): The name of the drug that the patient is reviewing. This feature will be used to group reviews by drug and analyze the effectiveness of each drug for specific conditions.
- Condition (categorical): The name of the condition that the patient is reviewing the drug for. This feature will be used to identify reviews related to Depression, High Blood Pressure, and Type 2 Diabetes.
- Review (text): The patient's review of the drug. This feature will be used to extract insights on the effectiveness and potential side effects of drugs for specific conditions.
- Rating (numerical): A 10-star patient rating reflecting overall patient satisfaction with the drug. This feature will be used to understand the level of patient satisfaction with different drugs for specific conditions.
- Date (date): The date on which the review was entered. This feature will be used to analyze trends over time in patient reviews and ratings.
- UsefulCount (numerical): The number of users who found the review useful. This feature will be used to identify reviews that are likely to be helpful in understanding the effectiveness and potential side effects of drugs for specific conditions.

3. EDA - EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is done to get the insight about the data and summarize the main characteristics. To understand dependency or correlation of the features. The Figure 1 is a bar graph which shows the top 10 drugs given in the data set with a rating of 10/10.

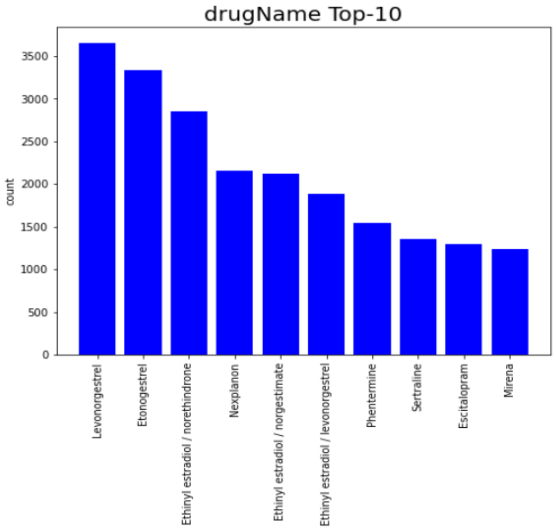

Figure 1. Top 10 drugName

'Levonorgestrel' is the drug with the highest number of 10/10 ratings, about 1883 Ratings in the data set for 'Levonorgestrel'. It's followed by 'Phentermine' with 1079 ratings. Levonorgestrel (LNG) is a synthetic progestogen similar to Progesterone used in contraception and hormone therapy. Also known as Plan B, it is used as a single agent in emergency contraception, and as a hormonal contraceptive released from an intrauterine device, commonly referred to as an IUD.

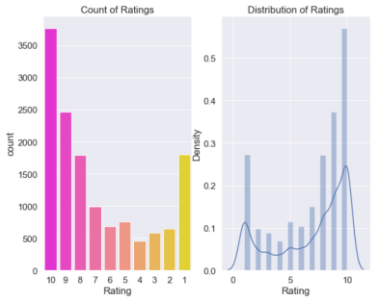

Figure 2. Count of Ratings

The Figure 2 shows a distribution plot on the right hand side and a bar graph of the same on the left hand side. This shows the distribution of the ratings from 1 to 10 in the data set. It can be inferred that mostly it's 10/10 rating and after that 9 and 1. The data points with rating of the drugs from 2 to 7 is pretty low.

The Figure 3 is a bar graph which exhibits the top 10 conditions the people are suffering from. In this data set 'Birth Control' is the most prominent condition by a very big margin followed by Depression and pain.

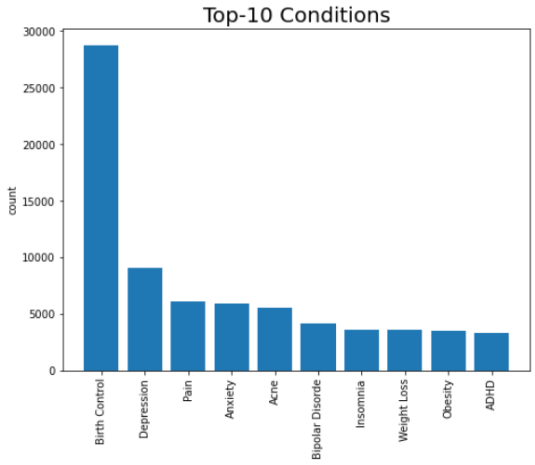

Figure 3. Top 10 Conditions

The 'Birth Control' condition has occurred about 28,000 and the depressions has occurred about 9,500. It can easily be noticed that the 'Birth Control' is more than 3 time the depression in the whole data set. In top 10 conditions ADHD is on the 10th Position, ADHD stands for Attention deficit hyperactivity disorder.

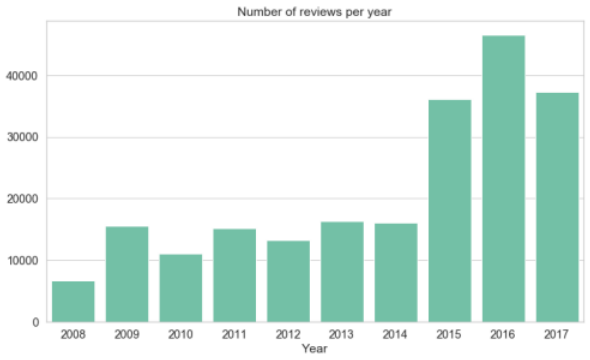

Figure 4. Number of reviews per year

The Figure 4. is a Bar graph that shows the number of reviews in the data set per year. It can be inferred that most ratings are given in 2016 and 2008 has the least number of reviews. 2016 have 46507 reviews whereas 2008 have 6700 reviews. The Figure 5. shows the mean rating of the drugs per year. 2008 have a mean rating of 8.92 which is the highest but it can also infer from the above bar graph Figure 3 that it has the lowest number of reviews by the patients. 2017 has a mean rating of 6.04 which is the lowest in the graph. The average rating per year is not below 5 in any given year from 2008 to 2017.

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analysing data from social network websites. The Figure 6. is a word cloud that shows most frequent word in big size.

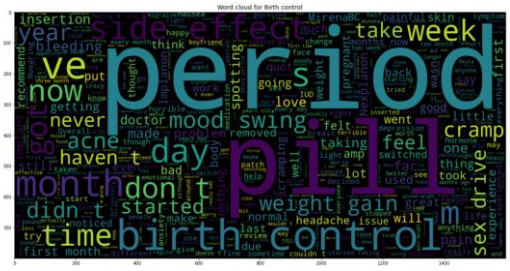

Figure 5. Word cloud for Birth control.

4. DATA PRE-PROCESSING

Data pre-processing is an important step in NLP because it can help to improve the accuracy and performance of NLP models. By cleaning and transforming the text data, data pre-processing can help to remove noise, standardize the data, and group together words that have similar meanings. Figure 7 shows data pre-processing of review feature to review_clean new feature.

Stop Word:

A stop word is a commonly used word (such as “the”, “a”, “an”, or “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

Lemmatization:

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item. Lemmatization is similar to stemming but it brings context to the words. So, it links words with similar meanings to one word.

Text pre-processing includes both Stemming as well as lemmatization. Many times, people find these two terms confusing. Some treat these two as the same. Lemmatization is preferred over Stemming because lemmatization does morphological analysis of the words.

	condition	review	review_clean
2	Birth Control	I used to take another oral contraceptive, which had 21 pill cycle, and was very happy- very light periods, max 5 days, no other side effects. But it contained hormone gestodene, which is not available in US, so I switched to Lybrel, because the ingredients are similar. When my other pills ended, I started Lybrel immediately, on my first day of period, as the instructions said. And the period lasted for two weeks. When taking the second pack- same two weeks. And now, with third pack things got even worse- my third period lasted for two weeks and now it's the end of the third week- I still have daily brown discharge.\r\nThe positive side is that I didn't have any other side effects. The idea of being period free was so tempting... Alas.	used take another oral contraceptive pill cycle happy light period max day side effect contained hormone gestodene available u switched lybrel ingredient similar pill ended started lybrel immediately first day period instruction said period lasted two week taking second pack two week third pack thing got even worse third period lasted two week end third week still daily brown discharge positive side side effect idea period free tempting ala
3	Birth Control	This is my first time using any form of birth control. I'm glad I went with the patch, I have been on it for 8 months. At first it decreased my libido but that subsided. The only downside is that it made my periods longer (5-6 days to be exact) I used to only have periods for 3-4 days max also made my cramps intense for the first two days of my period. I never had cramps before using birth control. Other than that in happy with the patch	first time using form birth control glad went patch month first decreased libido subsided downside made period longer day exact used period day max also made cramp intense first two day period never cramp using birth control happy patch

Figure 6. Data preprocessing

5. SPLIT THE DATASET

Data splitting is a crucial process in machine learning, involving the partitioning of a dataset into different subsets, such as training, validation, and test sets. This is essential for training models, tuning parameters, and ultimately assessing their performance. Split the dataset with 80% of training set and 20% of test set.

6. MODEL SELECTION

TF-IDF:

TF-IDF is a very popular feature extraction technique. Text needs to converted into vector or matrix before fed them to the Machine Learning model.

Term Frequency: TF of a term or word is the number of times the term appears in a document compared to the total number of words in the document.

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}}$$

Inverse Document Frequency: IDF of a term reflects the proportion of documents in the corpus that contain the term. Words unique to a small percentage of documents (e.g., technical jargon terms) receive higher importance values than words common across all documents (e.g., a, the, and).

$$IDF = log(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}})$$

$$TF\text{-}IDF = TF * IDF$$

The TF-IDF of a term is calculated by multiplying TF and IDF scores. Figure 8 shows confusion matrix of TF-IDF.

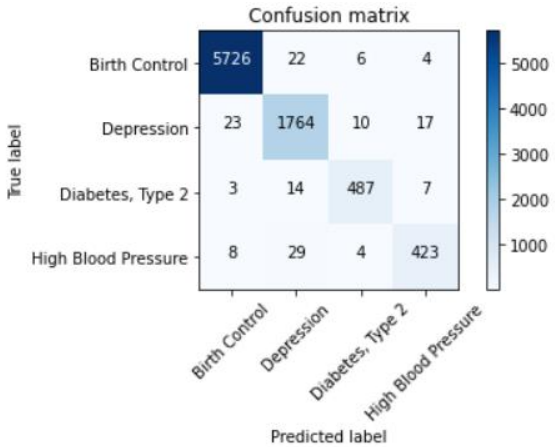

Figure 7. Confusion matrix of TF-IDF

Bag of Words

Bag-of-words(BoW) is a statistical language model used to analyse text and documents based on word count. The model does not account for word order within a document. BoW can be implemented as a Python dictionary with each key set to a

word and each value set to the number of times that word appears in a text.

Naive Bayes

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks. Sentiment classification using Naive Bayes algorithm is done through two stages: the learning process stage and the classification stage. Figure 8 shows confusion matrix of Naïve Bayes.

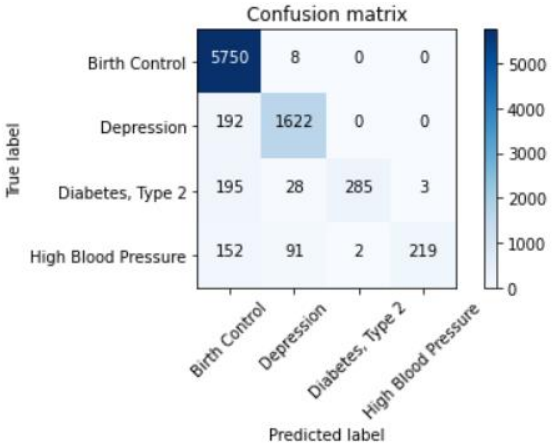

Figure 8. Confusion matrix of Naive Bayes

Passive Aggressive Classifier

The PAC algorithm responds aggressively to incorrect predictions and remains passive for the correct predictions. Passive Aggressive (PA) classifiers are a type of online learning algorithm that can be used for classification tasks. They are based on the idea of being "passive" when the current model correctly classifies a training example, but "aggressive" when it makes a mistake.

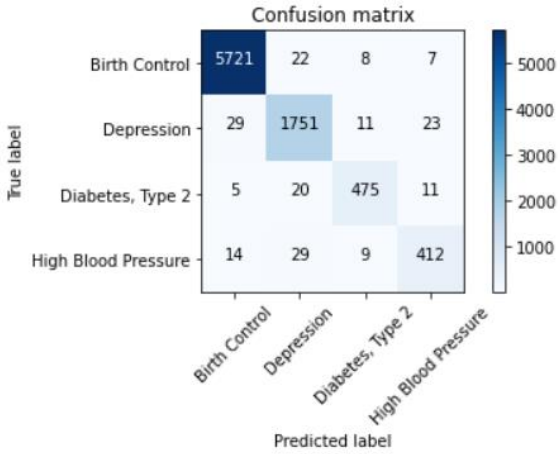

Figure 9. Confusion matrix of Passive Aggressive Classifier

7. MODEL EVALUATION

We have analysed different algorithms on the dataset to understand which one gives the better accuracy. The algorithms used are Naive Bayes, Decision tree, Random Forest, Linear SVC, Logistic Regression and Passive Aggressive Classifier. The algorithm giving the best accuracy was selected to train the dataset. The accuracy of the algorithm are given in the table 2. The best accuracy is Random forest.

Algorithm	Accuracy
Naive Bayes Algorithm	0.72
Decision Tree	0.8
Random Forest	0.85
SVC	0.74
Logistic Regression	0.74
Passive Aggressive Classifier	0.77

Table 1. Accuracy of different algorithms

8. CONCLUSION

With the advent of immense technological developments, especially the world wide web, individuals have found their ability to express opinions on a variety of products available in market. One such field is reviewing drugs for a medical condition. With many people relying on these reviews, extracting information from these reviews helps to identify whether a particular drug is proving to be beneficial as well as discover the aspect that might anger clients. In this paper we have proposed a drug recommendation system that helps to recommend the medications based on the reviews gained from their users. It is useful to understand the best possible medication for a condition and also helps in drug repurposing. We have used Random forest algorithm to recommend medicines which provides an accuracy of 85%.

REFERENCES

1. Gao, Xiaoyan, Fuli Feng, Heyan Huang, Xian-Ling Mao, Tian Lan, and Zewen Chi. "Food recommendation with graph convolutional network." Information Sciences 584 (2022): 170-183.
2. Chen, Yu-Xiu, Li-Chih Wang, and Pei-Chun Chu. "A medical dataset parameter recommendation system for an autoclave process and an empirical study." Procedia Manufacturing 51 (2020): 1046-1053.
3. Fox, Susannah and Duggan, Maeve. (2012). Implementing a Machine Learning Model to Realize an Effective IOMT Assisted Client Nutrition Recommender System. Pew Research Internet Project Report.
4. J. Ramos et al., "Using tf-idf to determine word relevance in document queries," in Proceedings of the first instructional conference on machine learning, vol. 242, pp. 133-142, Piscataway, NJ, 2013.
5. N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-Sampling Technique, 2011, Journal of Artificial Intelligence Research, Volume 16, 2020