HiLL PUBLISHING

Journal of Applied Mathematics and Computation, 2022, 6(2), 167-177
https://www.hillpublisher.com/journals/JAMC/
ISSN Online: 2576-0653
ISSN Print: 2576-0645

Extraction of Drug Review Polarity Using Sentimental Analysis

Raja Marappan*, S. Bhaskaran, S. Ashwadh, H. Aathi Raj

School of Computing, SASTRA Deemed University, Thanjavur, India.

How to cite this paper: Raja Marappan, S. Bhaskaran, S. Ashwadh, H. Aathi Raj. (2022) Extraction of Drug Review Polarity Using Sentimental Analysis. *Journal of Applied Mathematics and Computation*, 6(2), 167-177.
DOI: 10.26855/jamc.2022.06.001

Received: February 10, 2022
Accepted: March 8, 2022
Published: April 8, 2022

***Corresponding author:** Raja Marappan, School of Computing, SASTRA Deemed University, Thanjavur, India.
Email: raja_csmath@cse.sastra.edu

Abstract

The social networking site and user sites contain a large amount of information with users' feelings and opinions in various fields. For example, pharmaceutical companies provide users with text reviews and drug numeric ratings. Anyhow, these text-oriented reviews may not always be consistent with numerical values. In this project, we use different sentiment analysis models to differentiate drug user rating emotions and compare their accuracy. Various machine learning models including Logistic Regression, XG boost, and Naïve Bayes classifiers are being implemented by pushing it with drug reviews as inputs. In this, the XG boost model performed much better than other models with a total accuracy of 79.7%. This study has shown that these classification models can be used to separate drug reviews and identify the overall polarity or sentiment of the consumers. Since the main focus of this study was to separate the text reviews, the data were segregated according to each polarity. The division of binary classification was chosen instead of multiple classes as the purpose of this project was to identify the best and worst points/polarity. With the help of this polarity, it helps the nutritionist and the doctor to get updated with the in-market polarity of a particular drug and its public response.

Keywords

Sentimental Analysis, Drug Review, Accuracy

1. Introduction

Because the number of ailments is increasing exponentially, nations are facing a shortage of doctors, particularly in rural areas where the number of specialists is lower than in urban areas. It takes an allopathic practitioner almost half to a full decade to attain the necessary qualifications. As a result, the number of doctors cannot be increased at a rapid rate in a short period. Nowadays, clinical mistakes are quite common. Every year, about 2 lakh people in China and 1 lakh in the States get affected due to prescription mistakes. Specialists make errors while prescribing since their knowledge is limited. Choosing a Grade "A" medication is crucial for sick civilians who require the care of specialist doctors who are well-versed in nano-sized organisms, antibiotic drugs/medications, and patients. Every day, a new field of study comes out, bringing with it more medications, tests, and opportunities for clinical personnel.

Item reviews have become an irreplaceable and imminent factor for shopping of things worldwide, thanks to the rapid upscale of the internet and the online commerce industry. Individuals all across the world have become accustomed to reading reviews and visiting websites before deciding on a purchase. While the majority of previous research focused on rating expectations and making recommendations in the E-commerce area, the core region of medicare and the therapeutic clinic has been neglected. There has been an increase in the number of people concerned about their health and searching for a diagnosis online. According to a recent "American Research Center" poll conducted, over 60 percent of

adults made online searches about health-related subjects, and around 35 percent of users made searches for diagnosing health conditions. Medicine is chosen based on a unique set of circumstances, such as patient-given reviews examined with the help of sentiment analysis. Sentiment analysis is a concept or implementation of strategies and methods for recognizing and deriving emotional data from language, such as opinions and polarity.

2. Objectives

Here the proposed model constitutes several Machine Learning algorithms like Logical Regression, XgBoost, and Naïve Bayes methods that are trained on the set of medical records. These algorithms were weighed individually based on their accuracy score for different data sets. The processed dataset was trained over 3 different ML methods. Initially, the entire set of features available in the dataset is been processed, then each ML method was implemented one after next, our goal is to find a method with the best accuracy which may help in finding the polarity of the sentence.

3. Problem Statement

We live in a world where medicines have become an integral part of our daily diet. As per current estimates, it now requires at least 10 to 15 years and $500 million to $2 billion to bring a single product to market. Few people are aware that extensive assessments of hospital wards discovered that even properly prescribed medications cause about 19 lakh hospitalizations each year. Approximately a little less than a million inpatients are administered medications that can cause severe reactions, for a total of 20 lakh and 74 thousand severe ADR. Roughly 128,000 patients die as a result of medications prescribed to them. Numerous medical forums and post-trial surveys use ADR (Adverse Drug Reactions) and feelings and perceptions to discover these unforeseen side effects. These reviews assist the company, physicians, and many other patients in evaluating the drug's effectiveness. These sentiments and reviews are put through a sentiment analysis (ML) algorithms and lexicon based tools. These give away the user based satisfaction and dissatisfaction over a particular drug with their pre-existing conditions.

4. Literature Survey & Limitations

With a Dynamic increase in the advancement of AI, there is exertion in applying ML or DL strategies to sentiment analysis frameworks. Little less fortunate to know that few studies are happening in the field of drug polarity or proposal framework utilizing sentiment analysis. This grounds that the medication reviews are substantially more difficult to analyze. The aspect level based method is presented in compiling drug review data from three annotators to be and be pushed for processing over multiple rounds. After getting reasonable agreement from these annotators, the collected data will be sentimentally analyzed over multiple layers, namely Embedded layer, DBIGRU layer, PM-DBIGRU layer, Attention layer, Softmax layer to determine its polarity among its consumers [1]. DRAW—Drug Review Analysis Work, this research paper is primarily focused on converting raw text data into useful numeric representation since ML algorithms cannot process raw text data. It is done with BOW (Bag of Words) model and Feature Extraction methods for pre-processing. Multiple Vectorizer's such as TF-IDF etc and Multiple methods such as Logistic Regression etc are being implemented to determine polarity over large variants [2]. Over 100 thousand people were affected every year because of prescription mistakes. Over 40% medicine, specialists make mistakes while prescribing. Thus, getting reviews from patients may help doctors to choose which treatment or medications to give to a patient based on indications, past clinical history. Various methodologies were described by the author such as visualization, feature extraction, train test split, smote, classifiers etc. While analysis the reviews given by patients it'll be classified as postive or negative, depending on the user's star rating [3]. Extract reviews by age and sex which have been commented on by patients, as well as the recognition of sentiment words in each review. After establishing the opinion word, the next aim is to decide the orientation of the opinion word to ascertain if the reviews are pleasant or unpleasant and generate a synopsis [4-6].

5. Proposed Method—Data Flow Diagram

The overall flow diagram is shown in Figure 1.
The Entity Relationship Diagram is depicted in Figure 2.

6. Results & Analysis

6.1 Scikit

Scikit-learn is an open source Python Library. It is a simple and efficient tool for data mining and data analysis that is

accessible to everybody and can be reused in various context. It provides an array of tools for classification, clustering, regression, cross-validation, model fitting, data pre-processing, loading and splitting data, and evaluation of model using different metrics. The only drawback of this package is that many packages need to be installed before installing this package. The list of packages that should be installed in order are:

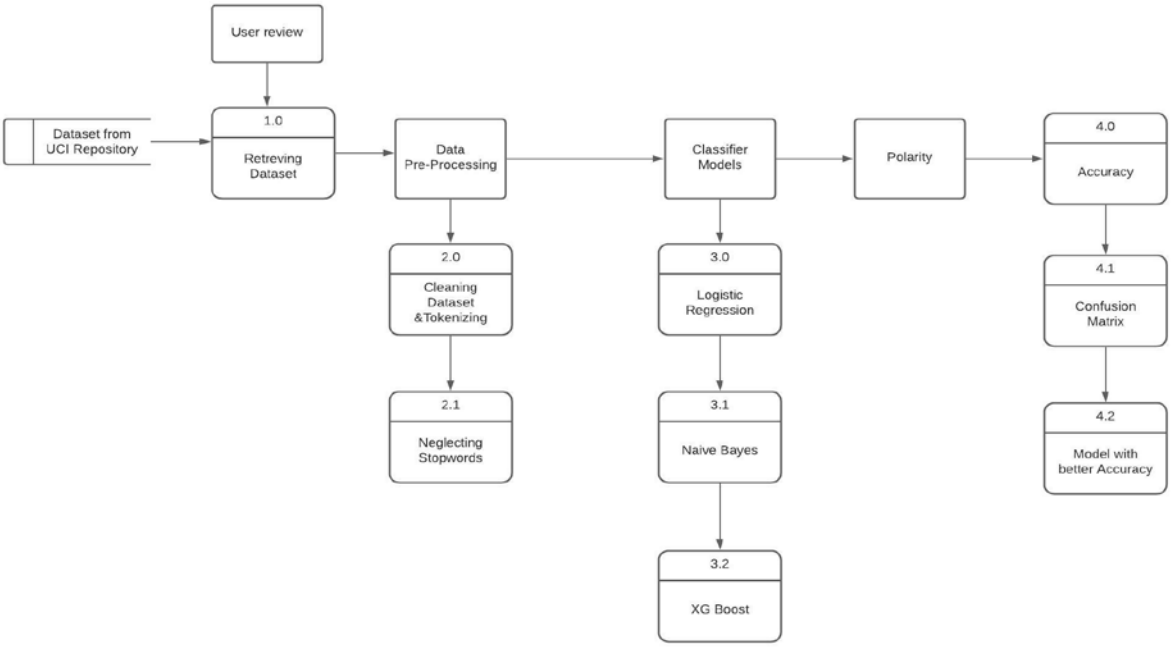

Figure 1. The overall data flow diagram.

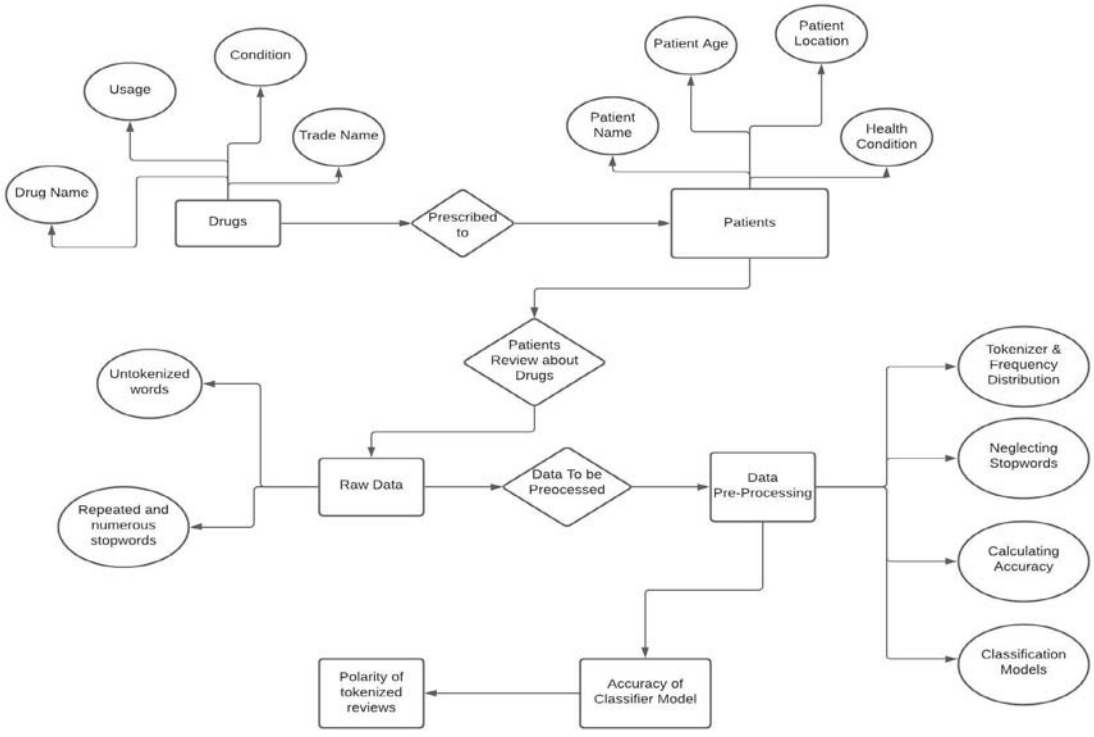

Figure 2. Entity Relationship Diagram.

Numpy 1.6.1+ - Library to handle advanced numerical capabilities
Pandas- Library for importing and analyzing data
Matplotlib – Library for data visualization
Sci-Py 0.9+ - Fundamental library for scientific calculations
Scikit-learn – Library for modeling data

Steps involved in coding the ML model using sklearn in python are:

1) Loading the data: The stored data is fetched using the pandas package with read_excel command.

```
>>>import pandas as pd
>>>data = pd.read_excel('file path')
```

2) Training and test data : The dataset is split into training and test dataset using train_test_split command available in sklearn.model_selection package.

```
>>> from sklearn.model_selection import train_test_split
>>> X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25)
```

3) Pre-processing the data : The dataset is preprocessed using different techniques like standardization, Normalization, Binarization, Inputing Missing Values, etc,.

```
>>> from sklearn.preprocessing import StandardScaler
>>> scaler=StandardScaler().fit(X_train)
>>> standardized_X=scaler.transform(X_train)
```

4) Creating the model : The model to be used for prediction or classification is created.

```
>>> from sklearn.linear_model import LinearRegression
>>> lr=LinearRegression(normalize=True)
```

5) Model fitting: Fir the model to the trained data.

```
>>>lr.fit(X,y)
```

6) Prediction: Predict the labels with predict command using the model created

```
>>>y_pred=lr.predict(X_test)
```

7) Evaluating model's performance : Evaluate the performance of the model selected using various benchmark tools such as MSE, MAE, R2 score etc.,

```
>>>from sklearn.metrics import mean_squared_error
>>>m=mean_squared_error(y_test,y_pred)
```

The dataset import is sketched in Figure 3.

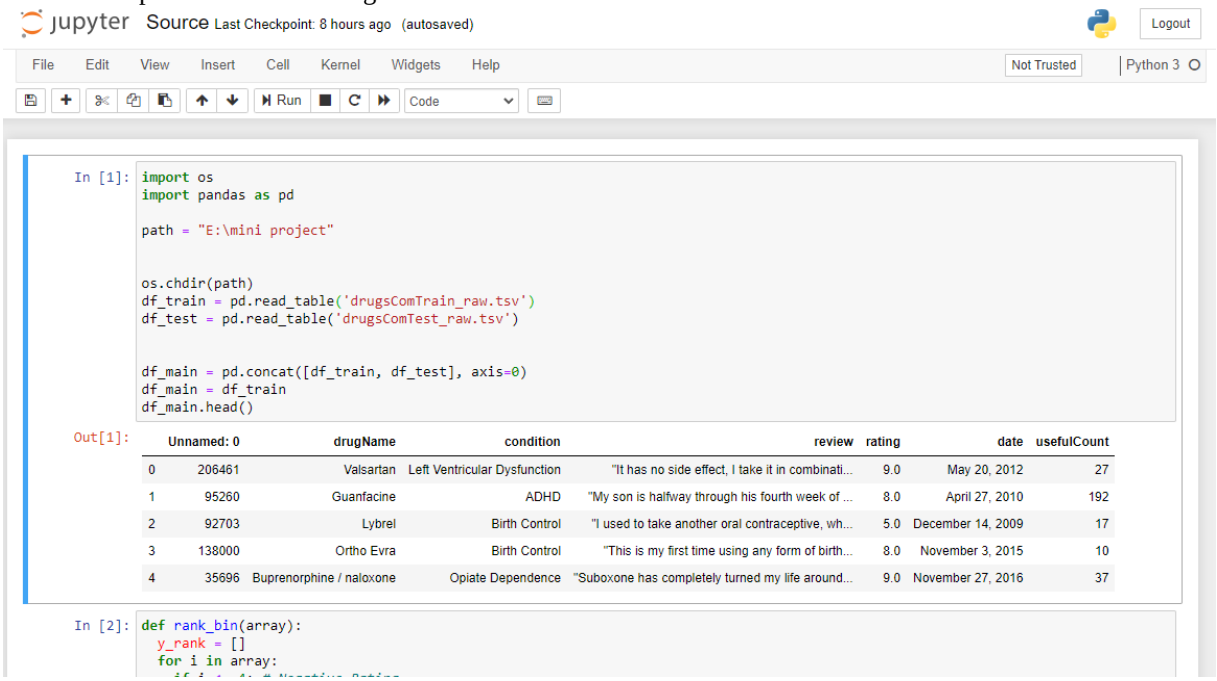

Figure 3. Importing Dataset.

6.2 Models used in Scikit

6.2.1 Logistic Regression

This is an efficient ML functional algorithm under supervised learning implemented where the objective is categorical. In simpler terms, LR can be defined as Linear Regression for classification problems. To predict and fit a categorical output variable, LR essentially deploys a LR function. When the distribution is not balanced and the effect of false +ve and false -ve are not equal, the resultant accuracy may not be of appropriate valuation metric in all classification.

Precision - TP/(TP+FP), Hence, the proportion of points model classify as positives are actually positives.

Recall - TP/(TP+FN), meaning the proportion of actual positives that are correctly classified by the model.

F1 score - the harmonic method of precision and recall.

Train-Test Split

Allow the data to be the test set, as defined by the class label.train, test = train_test_split(df_reviews, test_size = 0.3, stratify = df_reviews['labels'], random_state = 42)

Text Preprocessing

Cleaning the text in the data is a crucial step in most sentiment analysis or natural language processing tasks.

- Clean out symbols.
- Removal of stop word.
- Removing additional stop words. I.e. additional drug-based stop words.

Vectorization

To get out sensible information about all the data for our ML model, we ought to convert every review to a numeric value, which we'll term vectorization.

Confusion Matrix

After vectorizing the test data through classifier algorithm for a particular data set, a confusion matrix is constructed to characterize the performance of a classification model (Logistic Regression) on the set of test data. The confusion matrix is sketched in Figure 4.

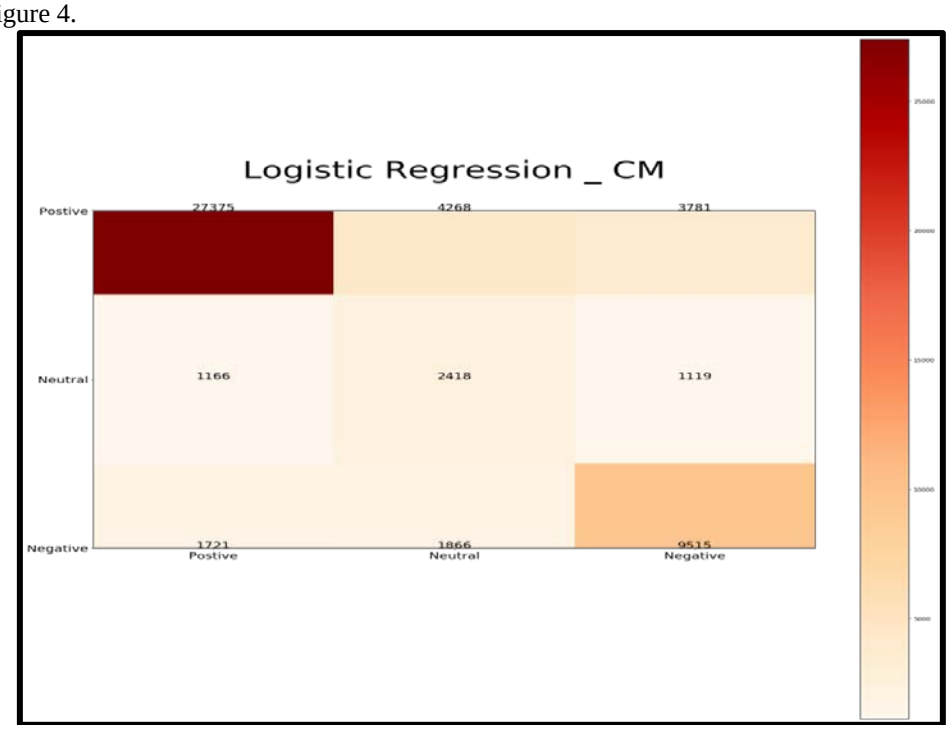

Figure 4. Confusion Matrix of Logistic Regression.

Accuracy: The accuracy of Logistic Regression is 73.84696312.

6.2.2 Naïve Bayes

This Naive Bayes classifier is a probabilistic machine learning model that would be used for classification methods.

The crux of said classifiers is dependent on the Naïve bayes theorem.

A theorem of Bayes:

$$P(A|B)=(P(B \mid A)P(A))/(P(B))$$

We can predict or find the probability of anything occurring using the Bayes theorem, assuming that B has happened. B is the evidence, while hypothesis is denoted by A. The assumption now is that the predictors or features are conscious. That is, the existence of one feature has no bearing on the presence of another. It is referred to as naive.

Data Preparation

Cleaning the text in the data is a crucial step in most sentiment analysis or natural language processing tasks.

- Clean out symbols.
- Removal of stop word.
- Removing additional stop words. I.e. additional drug-based stop words.

```
def preprocess_data(data):
# Remove package name as it's not relevant
data = data.drop('package_name', axis=1)
# Convert text to lowercase
data['review'] = data['review'].str.strip().str.lower()
return data
data = preprocess_data(data)
```

Splitting Data

To begin, divide the columns into dependent and independent variables (or features and labels). Then you divide those variables into train and test sets.

```
# Split into training and testing data
x = data['review']
y = data['polarity']
x, x_test, y, y_test = train_test_split(x,y, stratify=y, test_size=0.25, random_state=42)
```

Vectorization

To get out sensible information about all the data for our ML - model, we ought to convert every review to a numeric value, which we'll term vectorization.

```
# Vectorize text reviews to numbers
vec = CountVectorizer(stop_words='english')
x = vec.fit_transform(x).toarray()
x_test = vec.transform(x_test).toarray()
```

Confusion Matrix

After vectorizing the test data through classifier algorithm for a particular data set, a confusion matrix is constructed to characterize the performance of our classification model (Naive Bayes) on the set of drug test data. The resultant confusion matrix is shown in Figure 5.

Accuracy

Accuracy of Naïve Bayes is 75.22027466

6.2.3 XG Boost

XGBoost is a technique that has reached the top of applied “ML” & Kaggle tournaments for tabulated data. XG-Boost is a gradient-based decision tree approach that is optimized for speed and performance.

It is an implementation of Tianqi hen’s gradient boosting machines, now along with improvements from a plethora of developers. It relates to a larger group of tools known as the Distributed Machine Learing community or DMLC.

Model Features

The model’s implementation supports the functionality of the scikit-learn and R implementations, as well as additional features such as regularization. Three main types of gradient boosting are recommended:

- GB - algorithm, also known as the gradient boosting machine, which includes the learning rate.

- Stochastic Gradient Boosting with sub-samples there at row, column per split levels.
- Gradient Boosting Regularized with L1 and L2 Regularization

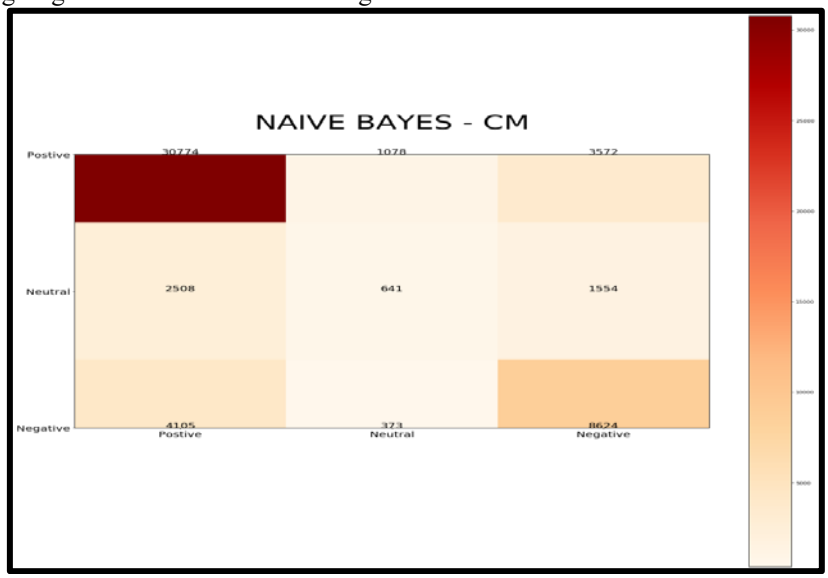

Figure 5. Confusion Matrix of Naïve Bayes.

Data Preprocessing

Cleaning text is a crucial step in most text mining or natural language processing tasks.

- Clean out symbols.
- Removal of stop word.
- Removing additional stop words. I.e. additional drug-based stop words.

Vectorising

To get out sensible information about all the data for our ML - model, we ought to convert every review to a numeric value, which we'll term vectorization.

Confusion Matrix

After vectorizing the test data through classifier algorithm for a particular data set, a confusion matrix is constructed to characterize the performance of a classification model (XG-Boost) on the set of test data. The resultant matrix is shown in Figure 6.

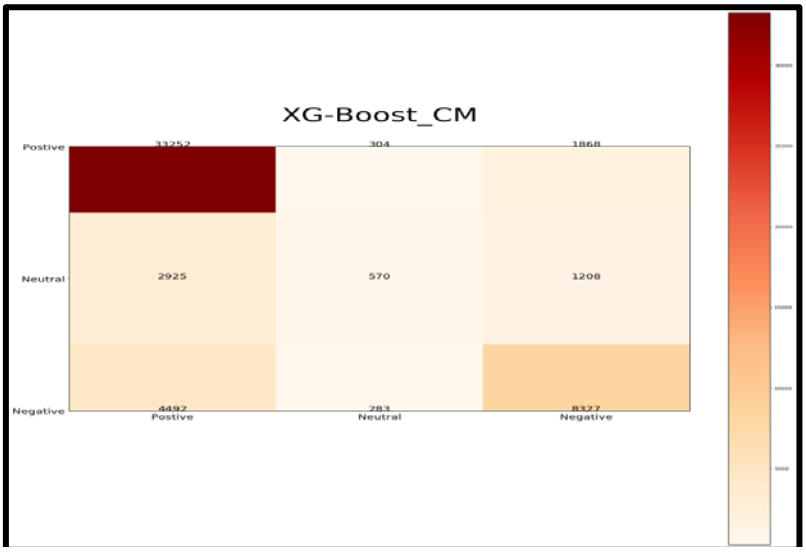

Figure 6. Confusion Matrix of XG-Boost.

Accuracy

The accuracy of XG-Boost is 79.18427924.

6.3 Visualization - Sentence length

Initially, we split our dataset into single words and then these sentences were processed and tokenized, these tokenized words further undergoes process like neglecting stopwords. For this, the sentence length of the tokenized words was taken into the account. The sentence length plot is constructed in Figure 7.

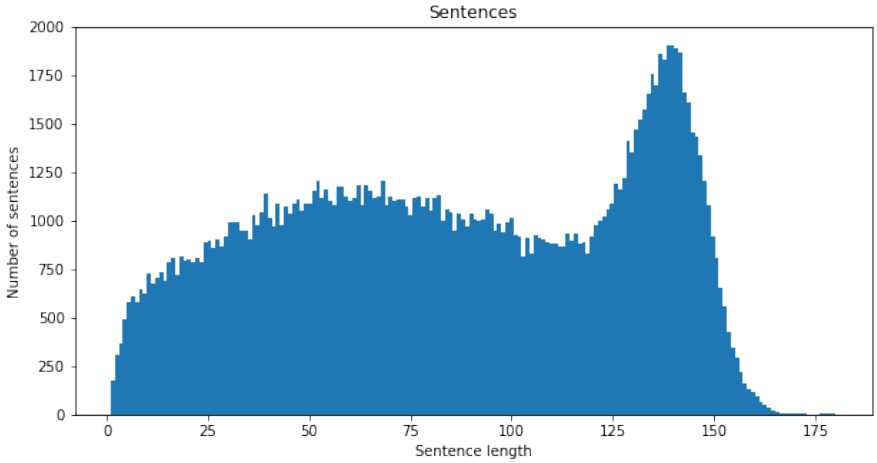

Figure 7. Plot for Sentence Length.

6.4 Polarity

Polarity describes the sentence whether it is positive or negative, once the processing phase gets over, the sentences were imported into the polarity checking phase, this helps in identifying both positive and negative reviews for the sentence. The polarity of sentences is plotted in Figure 8.

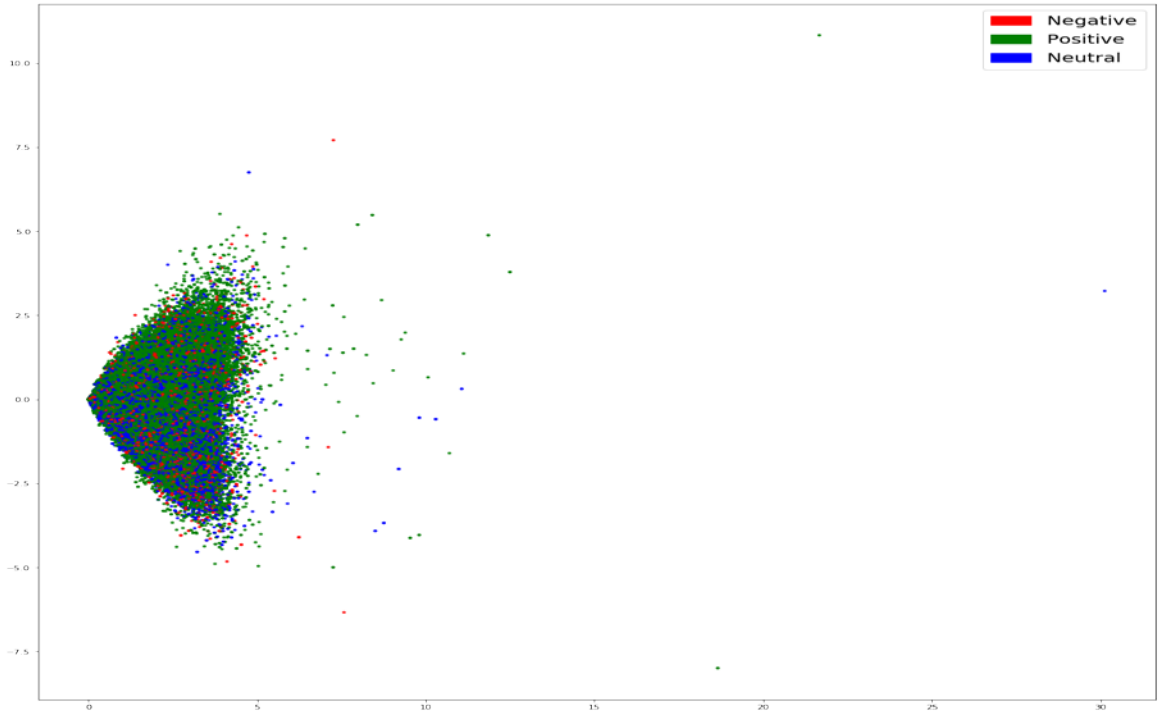

Figure 8. Polarity of Sentences.

6.5 Comparing Accuracy for Three Models

Since we used Three ML algorithms, we need to choose the best algorithm to get maximum accuracy, as a result, we plotted the accuracy attained by all three algorithms in Figure 9.

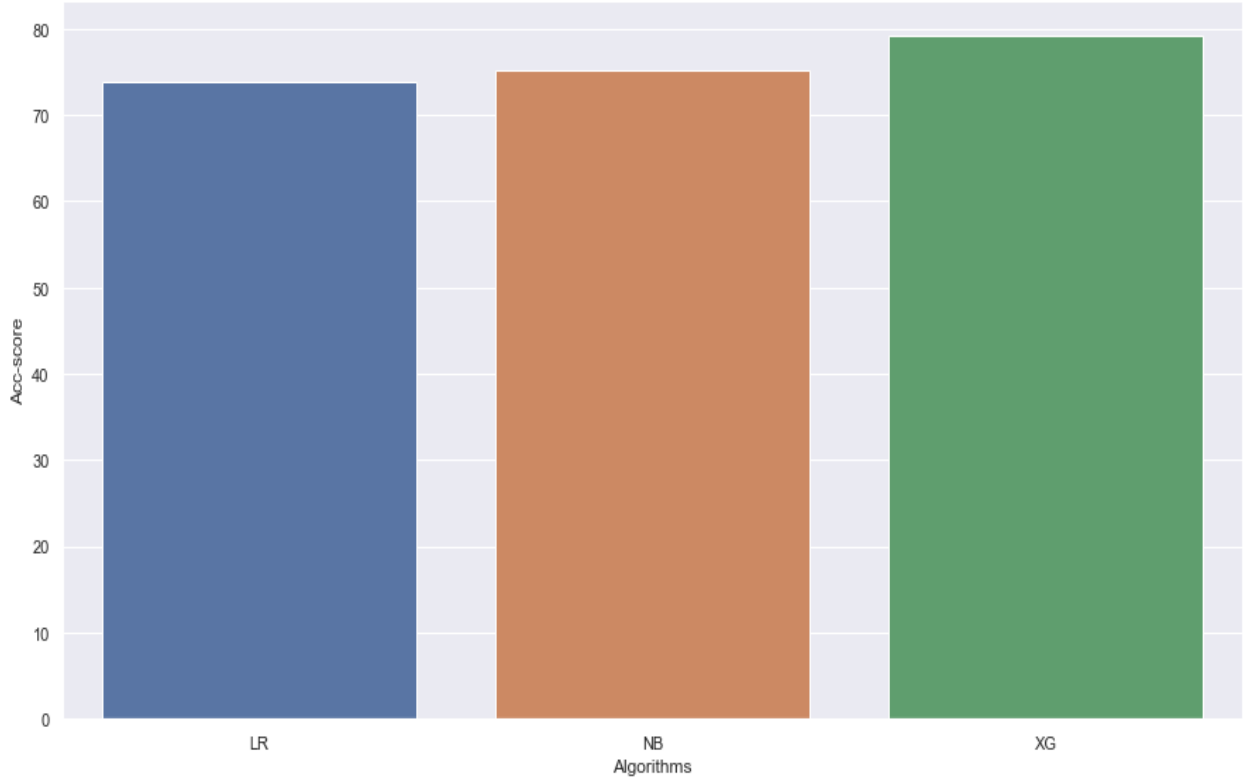

Figure 9. Comparing Accuracy for the three models.

By performance metrics, we calculated that XGBoost is the best algorithm with maximum accuracy and the accuracy of the three models is shown in Table 1.

Table 1. Accuracy of three models

Algorithmic model	Positivity	Negativity	Accuracy
Logistic Regression	27,375 Words	9,515 Words	73.84
Naïve Bayes	30,774 Words	8,624 Words	75.22
XG - Boost	33,252 Words	8,327 Words	79.18

6.6 Top & Bottom 20 Drugs on User Review

The top 20 and bottom 20 drugs based on the user reviews are shown in Figures 10 and 11 respectively.

7. Conclusions & Future Work

Using the entire data preprocessing and implementation of the above given algorithmic models, we have concluded that for this particular data set XG – Boost Algorithmic Model seems to best fit with a higher level of accuracy compared to the other two models. Therefore analyzing drugs based on user reviews helps nutritionists to prescribe better than before. We concluded that XGBoost algorithm has maximum accuracy which may be modified in the future.

The conclusion entirely depends on this data set and this method of implementation and the outcome may vary if additional processes are introduced to fine-tune any of the above-given models. In the future, we planned to host a web application based on the XG Boost algorithm through which users may update the dataset at any point in time, which would

help to provide better results instantly.

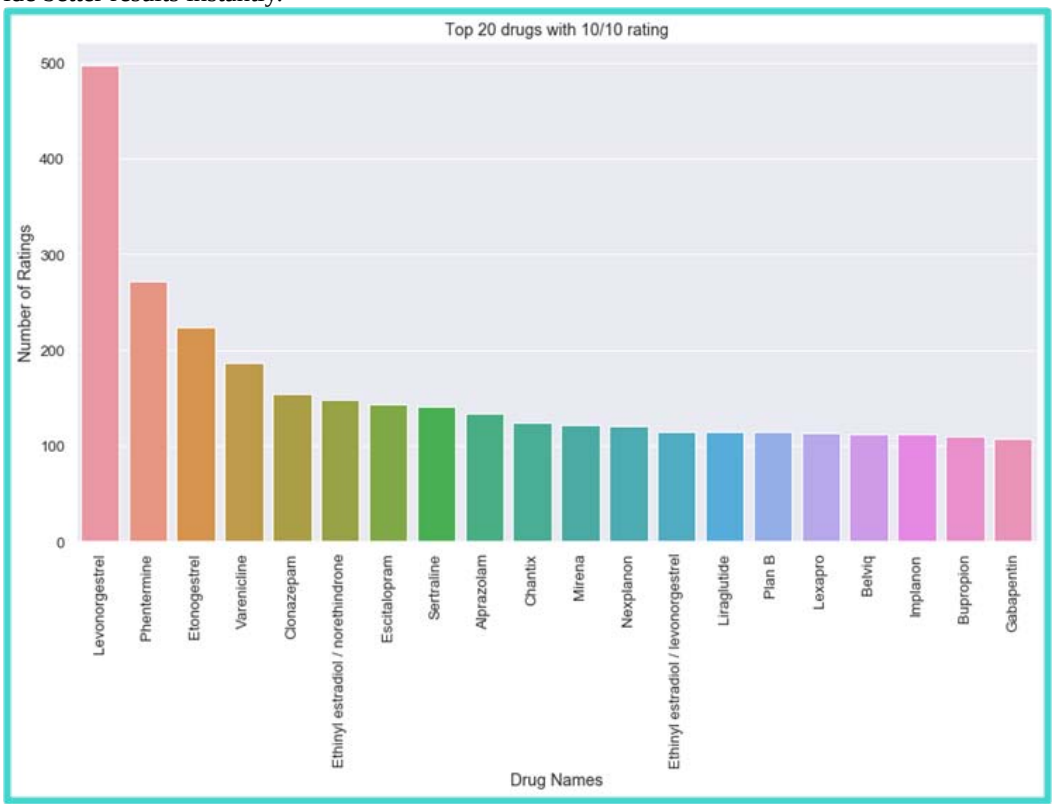

Figure 10. Top 20 drugs based on user review.

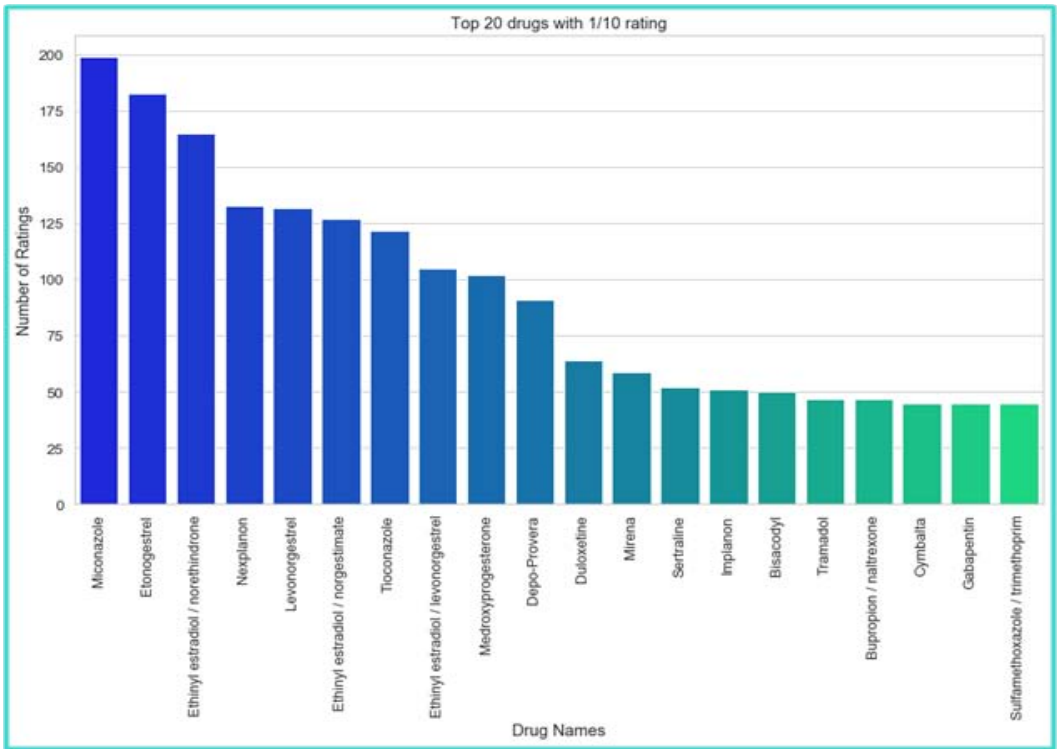

Figure 11. Bottom 20 drugs based on user review.

References

[1] S. Garg. (2021). "Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning," *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2021, pp. 175-181, doi: 10.1109/Confluence51648.2021.9377188.

[2] Y. Han, M. Liu, and W. Jing. (2020). "Aspect-Level Drug Reviews Sentiment Analysis Based on Double BiGRU and Knowledge Transfer," in *IEEE Access*, vol. 8, pp. 21314-21325, 2020, doi: 10.1109/ACCESS.2020.2969473.

[3] DRAW—Drug Review Analysis Work. Author: Akash Kunwar, Rohan Harode, Shubham Malik.

[4] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang. (2019). "SMOTETomek-Based Resampling for Personality Recognition", *IEEE Access*, vol. 7, pp. 129678-129689, 2019.

[5] J. Li, H. Xu, X. He, J. Deng, and X. Sun. (2016). "Tweet modeling with LSTM recurrent neural networks for hashtag recommendation", *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 1570-1577, 2016.

[6] V. Goel, A. K. Gupta, and N. Kumar. (2018). "Sentiment Analysis of Multilingual Twitter Data using Natural Language Processing", *2018 8th International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 208-212, 2018.