Drug Sentiment Analysis

Samin Ahsan Tausif	**180104053**
Rownok Jahan Nishat	**180104062**
Md.Nafisuzzaman Ayon	**180104065**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Fall 2021

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

4th September, 2022

Drug Sentiment Analysis

Submitted by

Samin Ahsan Tausif	**180104053**
Rownok Jahan Nishat	**180104062**
Md.Nafisuzzaman Ayon	**180104065**

Submitted To

Farzad Ahmed

Sajib Kumar Saha Joy,

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

4th September, 2022

ABSTRACT

As the number of viruses rises, so does the inaccessibility of genuine therapeutic resources, such as a scarcity of experts and healthcare workers, a lack of adequate equipment and medications, and so on. The whole medical community is in crisis, which has resulted in the deaths of many people. Individuals began taking medicine without necessary consultation due to the lack of availability, worsening their health condition. Machine learning has recently shown useful in a variety of applications, and there is a growth in new work for automation. In this research, a medication recommendation system is presented that might significantly lessen the workload of experts.

In this study, we have proposed a medicine recommendation system. that thoroughly utilizes patient reviews and users' ratings to classify positive and negative sentiment about any medicine using multiple Machine learning classification based algorithm like LGBM Classifier, Gaussian Naive Bayes, Random Forest Classifier, Decision Tree Classifier, which can help recommend the top drug for a specific disease in an efficient manner.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Clinical errors often occur today. Every year, medication errors have an impact on over 200 000 people in China and 100,000 people in the USA. Since experts only have a limited amount of information, they frequently prescribe the wrong medication (more than 40 percent of the time). Choosing the best prescription is essential for patients who require medical professionals with extensive knowledge of microscopic organisms, antibacterial drugs, and patients. Every day, new research is published along with more medications and diagnostic tools that are made available to healthcare professionals. As a result, selecting a treatment or medicine for a patient based on indications and prior clinical history becomes increasingly difficult for clinicians.

Item reviews have grown in importance due to the internet's rapid expansion and the growth of the web-based company sector. People worldwide have gotten used to reading reviews and browsing websites before making a purchase decision. While the majority of prior research focused on evaluating expectations and suggestions for the E-Commerce industry, the area of healthcare or therapeutic medicines has only sometimes been covered. The number of people searching online for a diagnosis because they are concerned about their health has increased. According to a Pew American Research Center poll conducted in 2013, 35 percent of users searched for diagnosing health disorders, while 60 percent of adults sought information on health-related topics online.

The drug recommendation system uses sentiment analysis and feature engineering to provide medications based on patient evaluations conditionally. Sentiment analysis is a development of techniques for identifying and extracting emotional information from text, such as opinions and attitudes. In contrast, the process of "feathering engineering" involves adding new features to the ones that already exist in order to enhance model performance. This analysis was divided into five parts: The introduction section presents a brief explanation of the necessity for this research, the related works section provides a succinct overview of prior studies on this topic, the methodology section covers the research methodologies, and the conclusion section summarizes the findings. The applied model results are evaluated using a variety of metrics in the Result phase, and the Discussion segment lists the framework's limitations before moving on to the conclusion.

Chapter 2

Literature Reviews

With the rapid growth of AI, there has been a push to apply machine learning and deep learning methodologies to recommendation systems. Nowadays, recommendation frameworks are used in the travel sector, e-commerce, restaurants, and so forth. Unfortunately, there are a limited number of studies available in the field of drug proposal framework utilizing sentiment analysis because medication reviews are significantly more difficult to analyze because they incorporate clinical wordings like infection names, reactions, and synthetic names that are used in the production of the drug.In this research, multilingual sentiment analysis was performed using Naive Bayes and Recurrent Neural Network (RNN). Google translator API was used to convert multilingual tweets into the English language. The results exhibit that RNN with 95.34 percent outperformed Naive Bayes, 77.21 percent.SVM was chosen for the drug recommendation module. since it performed admirably in all three distinct bounds - model exactness, model proficiency, and model versatility In addition, an error check mechanism was developed to assure analysis, accuracy, and administration quality. This proposed system was at first established on collaborative filtering strategies in which the solutions are initially bunched into clusters as demonstrated by the useful description data. However, after considering its shortcomings like computationally exorbitant, cold begin, and data sparsity, the model is moved to a cloud-helped approach utilizing tensor decomposition for progressing the quality

of involvement of medication recommendation.Five different metrics, precision, recall, f1score, accuracy, and AUC score were used which reveal that the Linear SVC on TF-IDF outperforms all other models with 93 percent accuracy. On the other hand, the Decision tree classifier on Word2Vec showed the worst performance by achieving only 78 percent accuracy. In the paper it was added best-predicted emotion values from each method, Perceptron on Bow (91 percent), LinearSVC on TF-IDF (93 percent), LGBM on Word2Vec (91 percent), Random Forest on manual features (88 percent), and multiply them by the normalized usefulCount to get the overall score of the drug by condition to build a recommendation system.

Chapter 3

Data Collection & Processing

3.1 Data Collection

3.2 Derived Features

Seven specific features have been derived from the dataset for the imperative purpose of sentimental analysis of users regarding medicine. **UniqueID** represents the unique ID of the users who registered to provide review regarding a specific nedicine. **Date** means the date a user has purchased the medicine. **Drugname** represents the name of the medicine a user bought for his specific disease and **condition** emphasises that disease of the particular user was suffering from. **Review** means the sentiment or review a user has provided after purchasing and using that medicine for a certain period of time, while **Ratings** represents the rating that user has provided for that medicine. Lastly, the feature **usefulCount** portrays how many medicine users have found that review helpful.

3.3 Data Visualization

Data visualization was performed on to show top 20 medicines according to conditions, medicines consisting 1 rating out of 10 and most importantly

word cloud of negative and positive reviews.

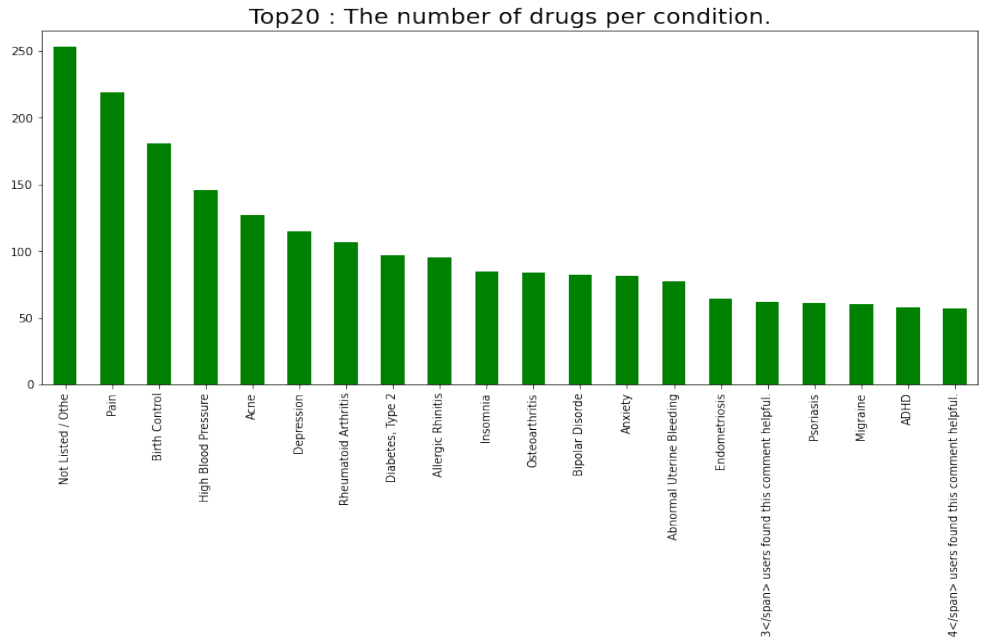

Figure 3.1: Top 20 medicines per health condition

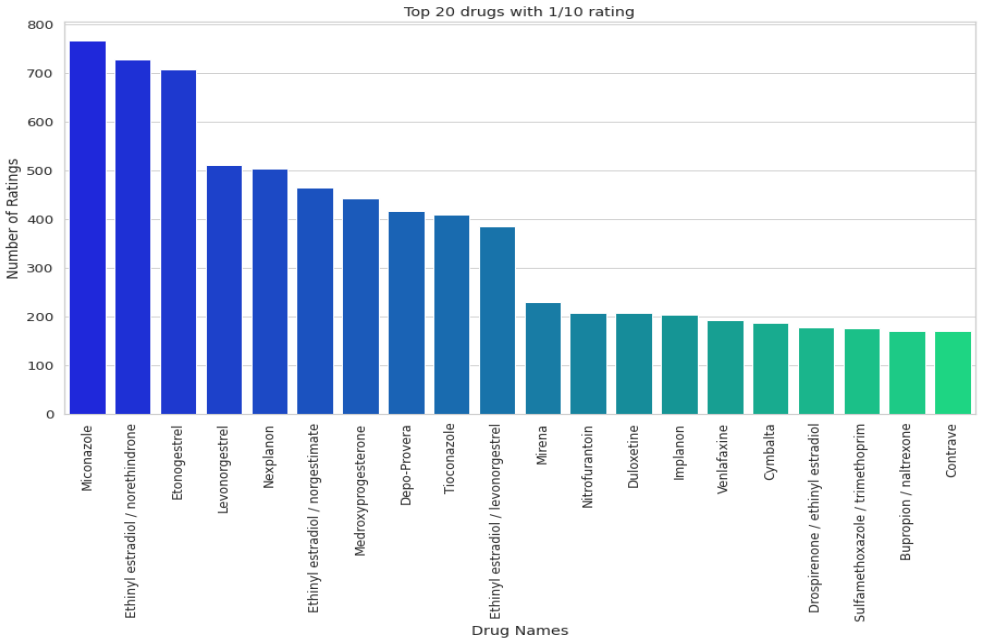

Figure 3.2: Top 20 medicines containing 1/10 rating

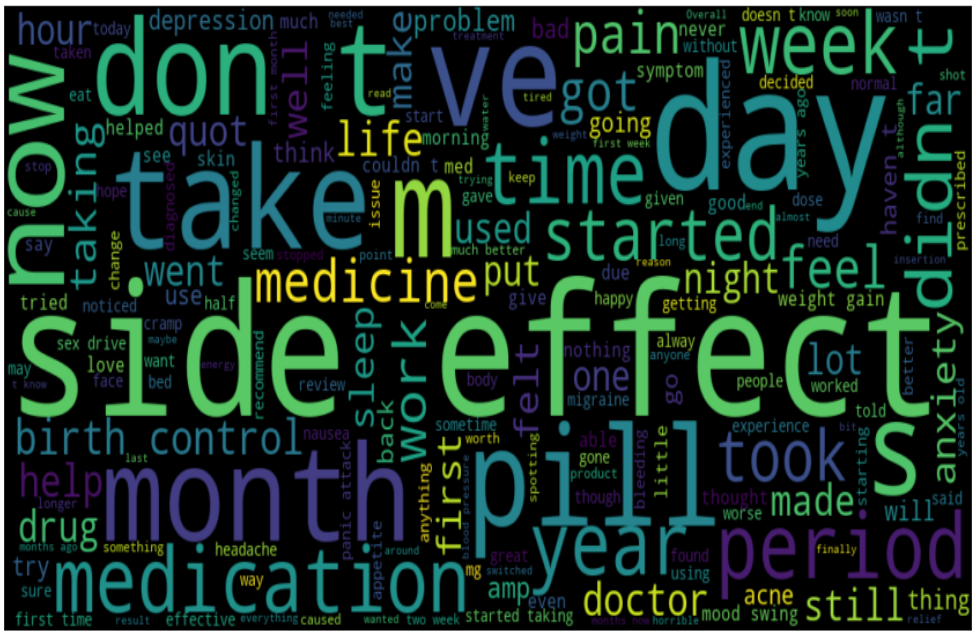

Figure 3.3: Word Cloud of Users' Reviews

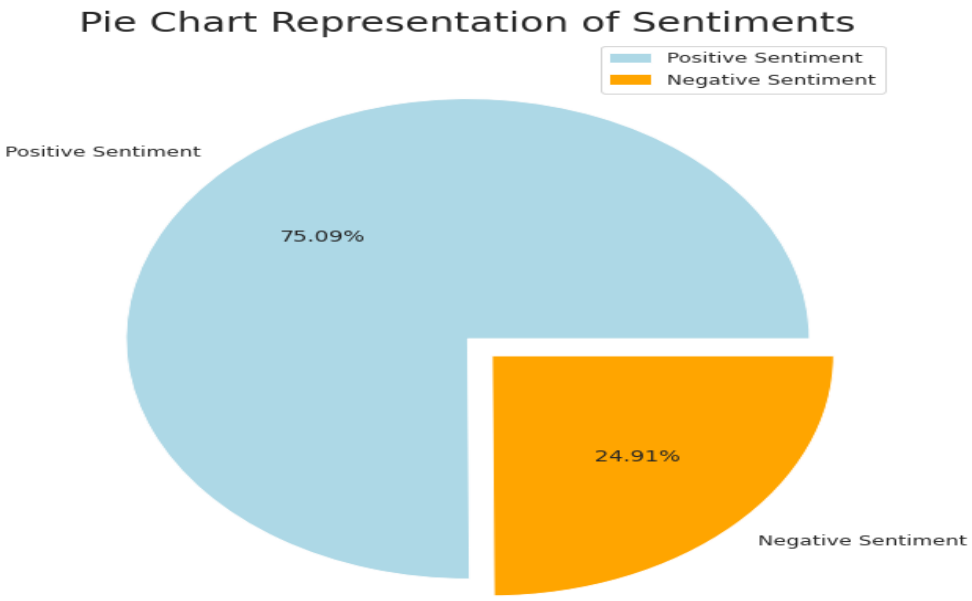

Figure 3.4: Feature Engineering

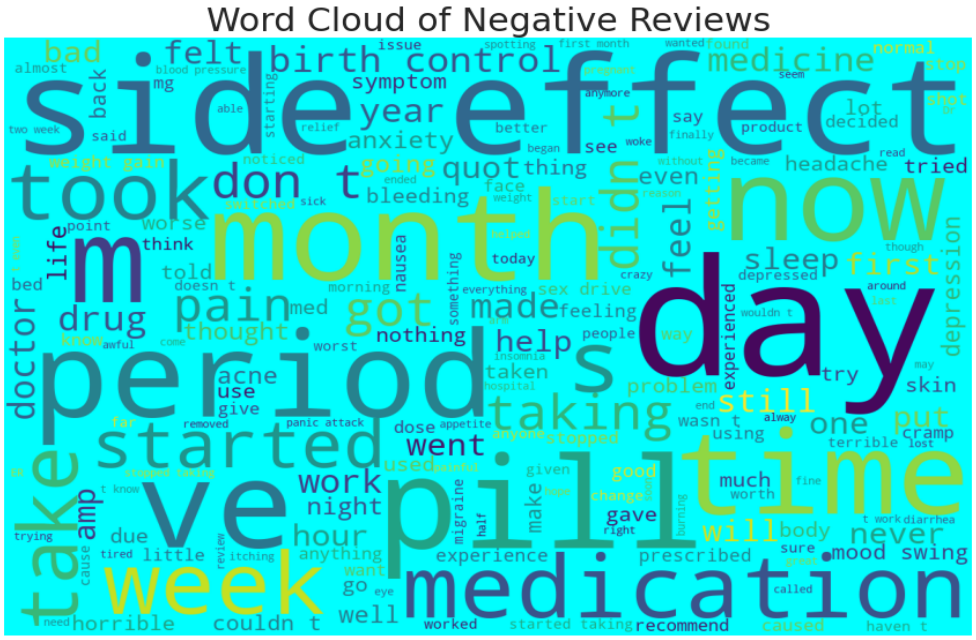

Figure 3.5: Word Cloud of Negative Users' Reviews

Chapter 4

Methodology

4.1 Working Procedure

Our workflow is defined using the following diagram. Firstly, we collected dataset based upon users' review on specific medicines according to potential diseases. Data visualization was performed on to show top 20 medicines according to conditions, medicines consisting 1 rating out of 10 and most importantly word cloud of negative and positive reviews. We derived many features and a total of around 10 features for the creation of our proposed model. We have removed null and redundant data during data preprocessing steps. Afterwards, we have performed feature engineering [Fig. 3.4] for our intrinsic purpose. In our research, users' medicine ratings above 5 has been considered as positive sentiment and rating below 5 has been availed as negative sentiment. Moreover, stopwords has been utilized to get rid of punctuations and redundant words. Furthermore, snowball stemming has been performed on this dataset in our research study to get more accurate users' reviews. Lastly, before model creation and hyper parameter tuning, label encoding was applied on drugname and users' condition. We then compared the performance of the models using these different feature selection approaches. To examine the performance of the models, we used classification metrics, a confusion matrix, and AUC ROC Curve.

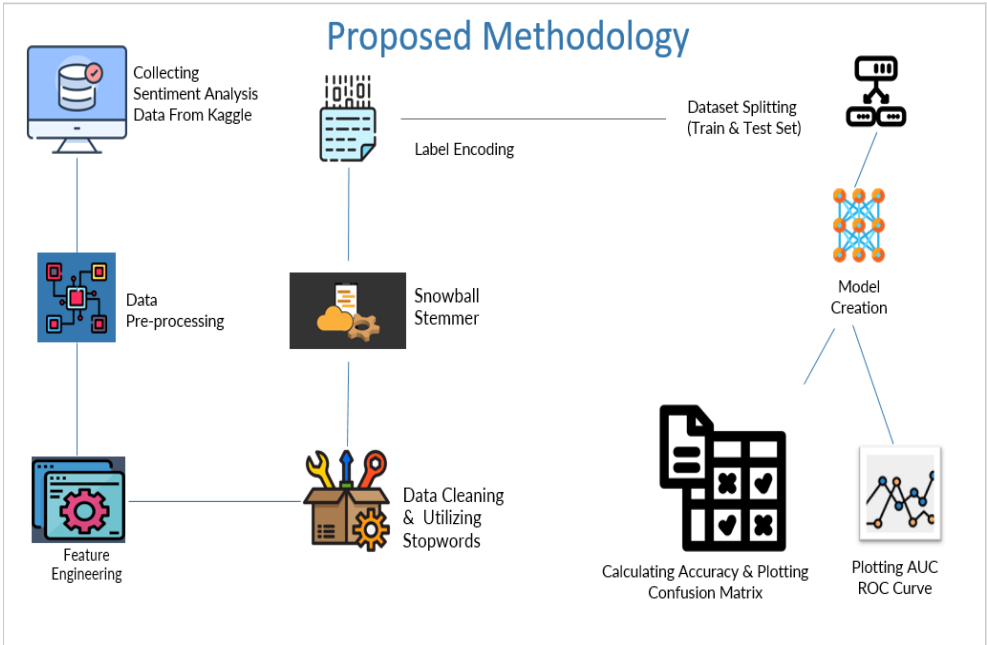

Figure 4.1: Flowchart of our proposed method

4.2 Data Cleaning & Stopwords

Stopwords are words in any language which does not add much meaning to a sentence. Stopwords can be ignored without sacrificing the purpose of the sentence. In the case of text classification or sentiment analysis, we should remove stop words as they do not provide any information to our model, i.e., keeping unwanted words out of our corpus. Stop words are often eliminated from the text before training deep learning and machine learning models since stop words occur in abundance, hence providing little to no unique information that researchers can use for classification.
We have cleaned our dataset's users review part first and afterwards utilized stopwords for our intrinsic purpose.

4.3 Stemming

Stemming is addressed as a natural language processing technique that declines inflection in words to their root forms, assisting in the preprocessing of text, words, and documents for text normalization. As a result, we utilize

stemming from reducing words to their primary form or stem, which might or might not be a legitimate word in the language.
SnowballStemmer() is actually a module in NLTK toolkit that implements the Snowball stemming technique for natural language processing tasks .

4.4 Label Encoding

In machine learning, we might need to deal with datasets containing multiple labels in multiple columns. These labels can be in the format of numbers or words. The training data is often labeled in words to construct the data understandable or in human-readable form.
Label Encoding guides converting the labels into a numeric form to convert them into a machine-readable format. ML algorithms can then determine in a better way how the models must operate those labels. It is an essential pre-processing step for the structured dataset in supervised learning. Before model creation and hyper parameter tuning, label encoding was applied on drugname and users' condition.

Chapter 5

Experiments and Results

5.1 Model Performance

The summary of our findings is shown in table I. From the table, we can visualize the fact that around five models have been tuned, created and utilized for the purpose of sentimental analysis of medicine users. The table illustrates that Random Forest and Decision Tree classifier are performing eventually good as Random forest classifier is great with high dimensional data. It is also faster to train than decision trees. Meanwhile, the results in table I shows that gaussian naive bayes and support vector classifier are failing to provide better results even after necessary parameter tuning. Lastly, the table depicts that Light Gradient Boosting Model seems to perform genuinely well because of its ensemble learning methodology. LGBM model always delivers compatability with both smaller and larger datasets and avails faster training speed. This model provides better accuracy than other boosting algorithm and regulates overfitting much better while working that can be noticed from the result of this table.

Model Name	Accuracy	Precision	Recall	F1-Score	AUC
LGBM Classifier	0.864	0.875	0.956	0.913	0.881
Decision Tree Classifier	0.753	0.755	0.992	0.858	0.679
Gaussian Naive Bayes	0.692	0.767	0.844	0.804	0.622
Random Forest Classifier	0.756	0.756	0.986	0.859	0.719
Support Vector Classifier	0.641	0.76	0.763	0.761	0.65

Table 5.1: Evaluation Metrics of Different Classification Models

5.2 Confusion Matrix

The confusion matrix for LGBM Classifier and Gaussian Naive Bayes Classifier has been illustrated below depicting necessary evaluation metrics for the judgement of the model performance.

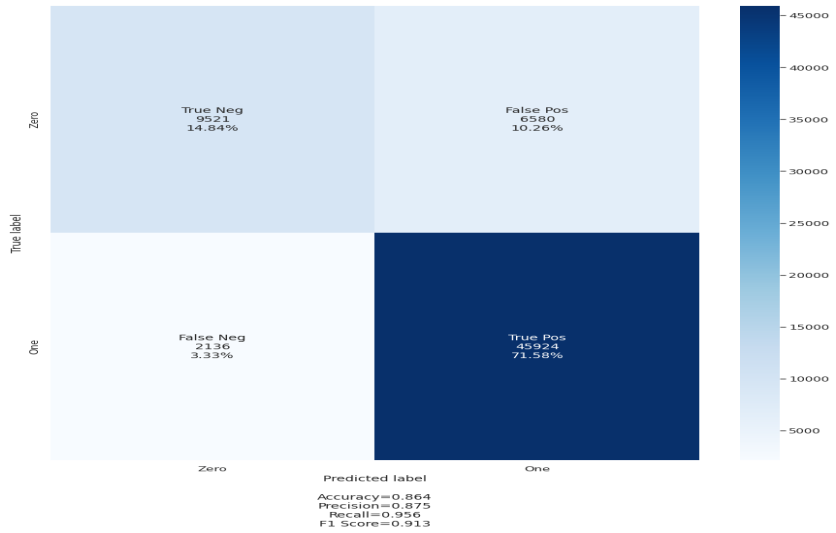

Figure 5.1: Confusion Matrix of LGBM Classifier

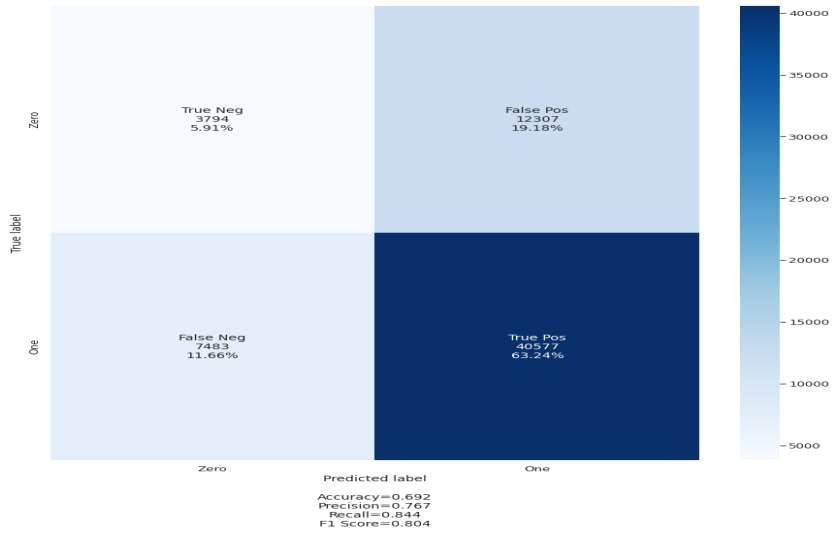

Figure 5.2: Confusion Matrix of Gaussian NB Classifier

5.3 AUC ROC Curve

THE AUC ROC curve is one of the most prominent evaluation metrics for evaluating any classification model's performance. It is also depicted as AUROC (Area Under the Receiver Operating Characteristics). AUC - ROC curve is a performance analysis tool for classification problems at various threshold settings. ROC is a probability curve, and AUC represents the degree or measure of separability. It illustrates how much the model is capable of distinguishing between classes. The higher the AUC, the better the model predicts zero classes as zero and one class as one.

From the figure, we can analyse that LGBM Classifier has higher auc score than the remaining classifiers. Random Forest Classifier tends to show second best accuracy and higher auc score than the other three classifiers. Henceforth, we can come to an outcome that LGBM and Random Forest classifiers are more precisely able to differentiate between two classes.

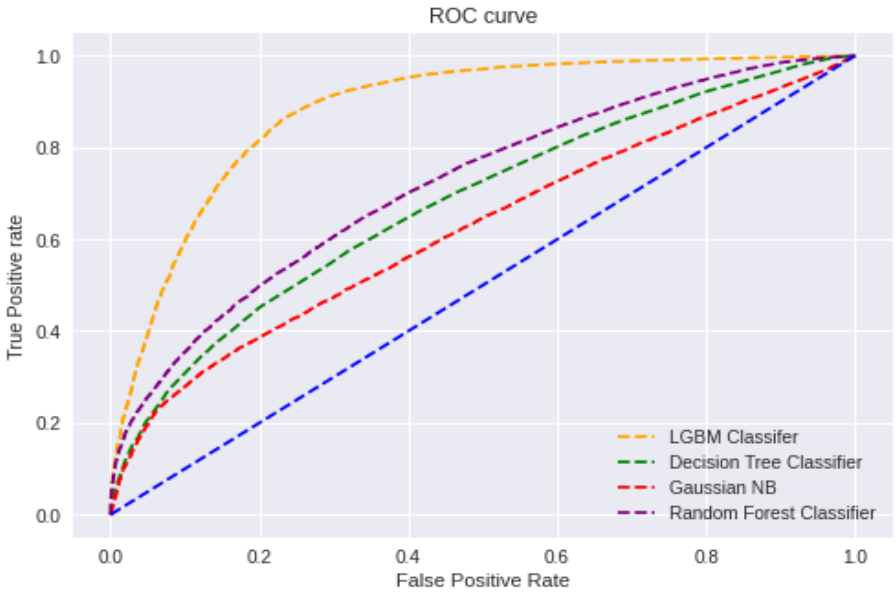

Figure 5.3: AUC ROC Curve

Chapter 6

Future Work and Conclusion

6.1 Future work

Future work involves comparison of different over-sampling techniques, using different values of n-grams, and optimization of algorithms to improve the performance of the recommendation system.Hyperparameter tuning for better model performance of the recommendation system. To reduce redundancy Usage of less features and only necessary features for classification model creation will be used for future research works.Undersampling of imbalanced dataset to reduce overfitting need to be applied. Applying TF-IDF, Word2Vec etc different approaches to find more precise evaluation.

6.2 Conclusion

Machine learning has recently shown useful in a variety of applications, and there has been a surge in new work for automation. This research attempts to offer a medicine recommendation system that can significantly minimize the workload of experts.In this study, medication evaluations were analyzed for sentiment in order to develop a recommendation system utilizing several machine learning classifiers such as LGBM, Decision Tree, Gaussian NB, Random Forrest , SVM. We evaluated them using five different metrics, precision, recall, f1score, accuracy, and AUC score.

References

[1] Cristóbal Colón-Ruiz, Isabel Segura-Bedmar (2020, October). Comparing deep learning architectures for sentiment analysis on drug reviews.
[2] Ioannis Korkontzelosa,Azadeh Nikfarjamb,Matthew Shardlowa, Abeed Sarkerb, Sophia Ananiadoua1 Graciela H.Gonzalezb (2016, August). Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts.
[3] Satvik Garg. (2021). Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning.
[4] Sairamvinay Vijayaraghavan, Debraj Basu. (2020, March). Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Sunday 4th September, 2022 at 5:07am.