“ Medicinal Drug Recommendation System ”

Major project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

in

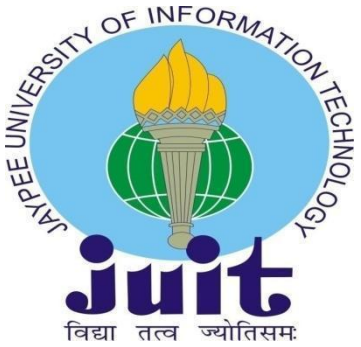
Computer Science and Engineering

By
DEEPTI AGGARWAL (181378)

UNDER THE SUPERVISION OF

DR. RUCHI VERMA

Assistant Professor (Grade - I) in CSE Department
of Jaypee University of Information Technology



Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology, Waknaghat, 173234,

Himachal Pradesh, INDIA

CERTIFICATE

This is to certify that the work which is being presented in the project report titled **"Medicinal Drug Recommendation System"** in partial fulfilment of the requirements for the award of the degree of B.Tech in Computer Science & Engineering and submitted to the Department of Computer Science & Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by **Deepti Aggarwal (181378)** during the period from January 2022 to May 2022, under the supervision of **Dr. Ruchi Verma**, Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

Deepti Aggarwal
(181378)

The above statement made is correct to the best of my knowledge.

Dr. Ruchi Verma
Assistant Professor (Grade - I)
Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology

Candidate's Declaration

I hereby declare that the work presented in this project entitled **"Medicinal Drug Recommendation System"** has been done by me under the supervision of **Dr. Ruchi Verma** (Assistant Professor ,Grade-I), Department of Computer Science & Engineering), Jaypee University of Information Technology.

I also declare that the matter embodied in this project has not been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Ruchi Verma
Assistant Professor (Grade - I)
Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology

Submitted by:

Deepti Aggarwal,
181378
CSE Department
JUIT

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

- **EDA:** Exploratory Data Analysis
- **ML:** Machine Learning
- **i.e:** that is
- **RAM:** Random Access Memory
- **CPU:** Central Processing Unit
- **GPU:** Graphics Processing Unit
- **NLTK:** Natural Language Toolkit
- **DTC:** Decision Tree classifier
- **POS:** Parts of Speech
- **LGBM:** LightGBM

LIST OF TABLES

LIST OF GRAPHS

LIST OF FIGURES

ABSTRACT

With the increase in the amount of clinical data being generated and scattered around the internet, "Health Information" has become the most concerned and searched topic on the internet.

Since there is an overload of information about medicinal practices, it has become difficult to make patient-oriented decisions for the medical professionals.

Research shows that a lot of medicine specialists usually make errors when they prescribe medicines to patients. This happens because they are not very experienced and sometimes even perform guess work for the same which is very dangerous.

These errors and mistakes made by inexperienced doctors while recommending medicinal drugs lead to multiple deaths.

To avoid such mistakes, we provide a medicine recommendation system for doctors which can be used by them while prescribing medicines.

We believe that a recommendation system, which can recommend medicines, can be really helpful to doctors and medical staff and pharmaceuticals to recommend the correct medication as per the condition of the patients. Hence, while improving the services, it will also take care about the patient's health and being. This could help in a lesser number of deaths due to errors and happier patients.

Chapter 01: INTRODUCTION

1.1 Introduction

For a very long time, vast amounts of clinical information about patients' or people's medical status, such as hospital reports, lab results, and medical operational procedures, have been stored and recorded somewhere. Now that there are a large number of records, the digital information about the same has increased dramatically, and this information is then used to make decisions about a patient's condition and medical problems. This digital information is distributed across multiple platforms rather than being confined to a single one. This makes it extremely difficult to find the information that medical personnel and patients require. Along with this, there has been a significant change in the medical field, as the number of tests, remedies, and medicines for a specific medical condition is increasing significantly with each passing day, making it very difficult to decide the correct and most appropriate remedies for the patients. Recommender systems have been integrated into the online services that provide this medical information, and these services help to make the selection process easier for the users. Previously, these systems were an important part of the healthcare domain to support the medical suggestions domain, hence the name Health Recommender Systems. These systems provide patients with a better understanding of their respective medical conditions, i.e. better personalization provided by medical data and increased details of provided recommendations. These systems aim to make patients' lives easier by providing a great experience for their medical condition suggestions, assisting them in taking better care of their health, and doing their best to make patients follow a better and fit lifestyle, as well as assisting health care workers and professionals with disease information and treatment.

We hope to depict the proposed medicine recommendation system and its operation in this project. This system employs current technologies such as machine learning, data mining, and so on. to find useful records embedded in medical data and reduce medical errors made by doctors when prescribing medications This system is made up of the following modules: database module, data preparation, data visualization, recommendation, and model evaluation module. The proposed medication recommender system employs machine learning N-Gram and Lightgbm algorithms to predict the best medicinal drug based on each patient's medical condition, achieving metrics such as good accuracy, scalability, and mode efficiency.

1.2 Project Objective

Our aim is to construct a recommendation model that can assist the doctors, be it experienced or inexperienced, people who are new to medicine fiend, medical practitioners and students as well as the patients or consumers themselves to prescribe and take ,respectively, the right medicinal drug, which won't harm the patient in any way,i.e a framework where they look up for the condition and the system would provide all the drugs listed with its accurate measure to tell which medicine is better.Our foremost goal will be to build a reliable and efficient model.

1.3 Project Motivation

A lot of medicine specialists usually make errors when they prescribe medicines to patients. This happens because they are not very experienced and sometimes even perform guess work for the same which is very dangerous. These errors and mistakes made by inexperienced doctors while recommending medicinal drugs lead to multiple deaths.

To avoid such mistakes, we provide a medicine recommendation system for doctors which can be used by them while prescribing medicines.

1.4 Language Used

Python:

Python is a decrypted, object-oriented, raised level programming with enthusiastic semantics. It is a combination of a raised level of certain data structures, and dynamic typing as well as dynamic binding which makes it engaging for various existing components to use it as a scripting language. Python reduces the cost of program maintenance as it has easy to learn syntax which is greatly helpful in increasing the readability of the code. The edit-test-debug cycle of python language is incredibly fast due to the absence of any compilation step.

Machine Learning :

Machine learning is the idea that machines can learn from the pre existing data and work in a way that they need not be explicitly programmed every time for a new data

set. It is a branch of Artificial Intelligence that aims at automating analytical models. Machine learning builds models from test inputs. Machine learning is done where masterminding and programming express computations is unthinkable. Models incorporate spam filtering.

Natural Language Processing :

Sentiment Analysis is a well-known NLP technique that is used to analyse and then classify text information or spoken human language information into specific classes. The Sentiment Analysis technique is used to categorise public opinions by classifying them as positive, neutral, or negative based on polarity.

Since python offers such a large collection of NLP tools and libraries to choose from, we generally use ***Python*** to perform NLP tasks. Other than this, there are other features of Python which makes it one of the best programming language choices for such NLP tasks and projects like the structure of syntax, which is so simple and the transparent semantics of python results in making it such a great choice of tasks that include Natural Language Processing tasks.

Python has so many amazing features that make it so versatile and overall such a great technology for working on tasks involving machines processing natural languages. We know that Python provides programmers with a vast array of tools that can aid in the performance of natural language processing tasks. One of them is POS tagging, which will be used in the project's sentiment analysis phase. It also allowed coders to create classifications of documents and models.

Since python has such a vast collection of modulus and libraries, it has been used in so many tasks and has therefore replaced many programming languages and has become one of the most popular programming languages, for performing the nlp related sentiment analysis task. Some of the Python libraries that used in performing Machine Learning tasks are :

- Numpy
- Scipy
- Pandas
- Matplotlib
- Seaborn

Keras:

In Python script, which is capable of running on TensorFlow, CNTK, or Theano allows for fast experimentation, Keras is one of the high-level API neural networks which was built to enable faster experimentation.

NumPy:

Another open source Python library is Numpy which is used for scientific computing and helps users and programmers to work with the mathematical functions , around which the concept of machine learning revolves. Not only this, numpy also allows python to work with most efficient arrays and matrices.

Pandas:

Pandas is a package which helps in data manipulation and is a very important tool when it comes to cleaning of data, exploration of data and visualization tasks as well as data manipulation.

1.5 Technical Requirements (Hardware)

1. RAM requirement will be >= 1 GB
2. Any Intel Preprocessor will work fine
3. Hard Disk is required to be >=6 GB
4. Speed need to be >=1 GHZ

- ❏ Linux Operating System/Windows
- ❏ Modern Web Browser
- ❏ Python Platform (Anaconda2, Spyder, Jupyter)

- ❑ NLTK Toolkit
- ❑ Visualization Modules
- ❑ Sklearn Module

1.6 Deliverables of the Project

➔ Medicine recommendation systems that will act as a helping hand to the doctors or medicine practitioners in selecting the right medicinal drug for the patients according to their diseases/medical conditions. Hence, a strategy that will take care of the safety of the people consuming that medicine as well as providing exceptional services.

➔ We measure the total mean predicted values for every drug listed for the diseases/medical condition and then on the basis of the order of the value, to recommend the most accurate and most useful, less harmful drug for every problem/disease .

Chapter 02: LITERATURE SURVEY

- **Garg, Satvik. (2021).” Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning”.**

Preprocessing:

Start by checking for missing records as well as redundant columns or rows or any such data, and removing all such unwanted and unusual information. It also removes all the patient conditions that have no meaning at all. The data is cleaned for further analysis.

Methodology:

Begin by cleaning the data and then visualising the data exploration process. There was a feature engineering and feature extraction scope. Perform a test train split once the data is ready for further analysis. SMOTE can be used to balance the data. Using classifiers, check the performance metrics, and then create a recommendation system.

Conclusion:

In this work,every comment left by the consumers were categorised as a positive feedback or a negative feedback, and at the same time the ratings were also considered. Depending upon the rate, it was considered good and bad by taking ratings less than equal to 5 as negative and others as positive.

Research Gaps:

taking up different values of n-gram and also comparisons between different over-sampling techniques.

- **Sridevi. U.K, Shanthi. P,” An Ontology-Based Sentiment Analysis Model**

towards Classification of Drug Reviews", Vol-2786, CEUR-WS

Preprocessing:

Different sets of keywords related to mental disorders have been used to categorize the reviews by several patients.

Methodology:

Long Short Term Memory Network which is an RNN network was used to categorize opinions. Precision, recall and F-score were used to calculate the accuracy and also comparison in between the models.

Conclusion:

After combining the values of F-score, Precision and Recall for all the three polarities, the accuracy of the model came out to be 63%.

Research Gaps:

The possible drug review interactions should have been concentrated upon more so that the comprehensive drug review data would have better served an individual's healthier lives.

❖ **M. D. Hossain, M. S. Azam, M. J. Ali and H. Sabit, "Drugs Rating Generation and Recommendation from Sentiment Analysis of Drug Reviews using Machine Learning," 2020 Emerging Technology in Computing, Communication and Electronics (ETCCE), 2020, pp. 1-6, doi: 10.1109/ETCCE51779.2020.9350868.**

Preprocessing:

Start by looking at the null data, and the data which is not required and is unwanted, any unrelated fields that are beside the point of the data provided. The moving onto the handling of reviews by breaking them down to tokens, taking care of stop words, formatting the words to their root forms, finding for scope of feature extraction and looking into n grams.

Methodology:

The methodology starts from cleaning the data by using techniques discussed in

preprocessing. Then considering the rating and reviews left by consumers and generating a model for example KNN, Decision Tree Classifiers, SVC and performing sentimental analysis and then finally generating a recommendation system.

Conclusion:

The models except KNN were showing good accuracy.The KNN model was found to be least accurate. The decision tree model was recommending drugs with accuracy of about 76% while on the other hand, the support vector model was about 83% accurate.

Research Gaps:

Their long-term goal was to improve the efficiency and reliability of their system by performing phrase-level sentiment analysis. They would also like to work on tensor factorization techniques.

- **Varun A.Goyal , Dilip J. Parmar , Namaskar I. Joshi , Prof. Komal Champanerkar,"Medicine Recommendation System",International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 03 | Mar 2020**

Preprocessing:

It uses data mining techniques.

Methodology:

(i) Data extraction
(ii) Pre-processing
(iii) Random Forest, GBDT
(iv) Recommends medicine

Conclusion:

The user gives the details of his condition and a medicine is predicted according to it.

Research Gaps:

The ability of preprocessing big diagnosis data needs to be enlarged.

- **T. Venkat Narayana Rao, Anjum Unnisa, Kotha Sreni,"Medicine Recommendation System Based On Patient Reviews" ,International Journal of Scientific & Technology Research Volume 9, Issue 02,February 2020**

Preprocessing:

It started by understanding and investigating the data and then preprocessing it.The information is cleaned for further analysis. Since it consists of multiple missing values and correlated data. The tasks were to find such data and remove all defective, unusual information and remove data redundancy.

Methodology:

Started by analysing the review dataset, by performing exploratory analysis and visualizations. Then preprocessing the data according to the defective, unusual values present. Then once the data was read for further modelling, the next was the model building stage for which the N-gram model and LightGBM model were used. And finally the model was predicting the proper medicine for a particular disease.

Conclusion:

The system recommended all medicinal drugs based on their mean predicted value. The medicine was deemed accurate based on these predicted values, and if the mean value was higher, it was deemed more accurate, and thus that medicine was recommended. The N-gram model was found to be 80% accurate at recommending appropriate medications, while the Lightgbm model was found to be 90% accurate.

Research Gaps:

Since the ages and demographic information was not considered during the training phase, efficiency would have improved by considering the same . Apart from this, the products and drugs that were given were not provided with their brands or the chemical contents which could've helped to recommend better medicines.

- **Y. Bao and X. Jiang, "An intelligent medicine recommender system framework," 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), 2016, pp. 1383-1388, doi: 10.1109/ICIEA.2016.7603801.**

Preprocessing:

The preprocessing started by processing the missing values in the data and then performing a chi-square test for analysis of correlation followed by normalization of data..

Methodology:

The modules in the research paper were - the database system, followed by a data preparation stage, then there was a recommendation model part and finally, evaluation of model and data visualization was done . Neural networks, SVM and DT classifiers are some of the models that were used to train the model.

Conclusion:

Accuracy for the ID3 decision tree was found to be 89%, whereas,The SVM model was considered best. It was found to give out the best predictions with 95% accurate values. Although BPNN was found to be the most accurate with 97% accuracy, its running time and data understanding was not good.

Research Gaps:

The authors want to increase the accuracy and efficiency of their model as well as want to compare the existing model with the one where they would use a technology called MapReduce parallel.

- **Na, Jin-Cheon & Kyaing, Wai. (2015). "Sentiment Analysis of User-Generated Content on Drug Review Websites". Journal of Information Science Theory and Practice. 3. 6-23. 10.1633/JISTaP.2015.3.1.1.**

Preprocessing:

The grammatical relationship of words in a clause was processed using the Stanford NLP library.

Methodology:

Sentiment Lexicons were created to collect the negative and positive phrases from the data and then the dependency tree was applied to the data for finding the grammatical relationship between the words i.e. if it is the governor or the dependent. SVM and Linguistic Approach was also used to check the accuracy.

Conclusion:

The accuracy of the model came out to be 62% according to the first SVM and 66%

according to the second SVM but the highest accuracy that we came across came from the Linguistic Approach i.e. 69%.

Research Gaps:

To reduce the possibility of error, the tagged aspects were manually checked, which is a time-consuming process that should be changed. Furthermore, the evaluation was conducted on relatively smaller datasets, which must be improved in conjunction with the use of specialised rules using machine learning for a large set of sentiment analysis rules.

Chapter 03: SYSTEM DEVELOPMENT

3.1 Computational Model Development

For this project work we used the machine with the following specs at the time of training.

CPU: The computer we used had the following specs:

Table 3.1 CPU specifications

Parameter	Specifications
CPU Model name	Intel(R)Core(TM)
CPU frequency	2.3 Ghz
No of CPU cores	4
Available Ram	7.86 GB
Disk Space	400 GB

These are the CPU specs of the machine we used to do computations. Most of the computational work mostly happens on the GPU. But the CPU takes care of most of the preprocessing. The large amount of RAM did not put loads of pressure and made it easier for the whole dataset to be loaded in time and we did not have to worry about any system crashes occurring. The clock speed of the CPU mentioned 2.3 Ghz is the basic clock speed which if needed can go upto 5 Ghz. But no overclocking was needed as the system was able to do the work in its normal 4 cores.

Table 3.2: GPU Specifications

Parameter	Specifications
GPU	NVIDIA GeForce GTX 1060
GPU Memory	10 GB
GPU Memory Clock	1.40 Ghz
GPU Release Year	2016
Cores	2
Available RAM	6 GB
Disk Space	400 GB

Tools Used: We used the following tools in making our model.

- Python
- Matplotlib
- Seaborn
- Jupyter Notebook
- Pandas
- Plotly
- Numpy
- Scikit Learn
- Flask Framework

These packages mentioned above were used in their latest upto date editions. The code works properly and would not cause any issue until any further updates in them.

3.2 Date Set Used in the Major Project

- The drug dataset that is being used for the project is taken from UCI Machine Learning Repository , provided in portals for such pharmaceutical databases known as Drug.com and Druglib.com. These portals are open to medicine practitioners as well as patients and have a large visit count by the people. largest and the most widely visited. This

contains **over 200,000 patient drug reviews.**

	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	3-Nov-15	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27-Nov-16	37
...	...	...	...	...	...	...	...
161292	191035	Campral	Alcohol Dependence	"I wrote my first report in Mid-October of 201...	10	31-May-15	125
161293	127085	Metoclopramide	Nausea/Vomiting	"I was given this in IV before surgey. I immed...	1	1-Nov-11	34
161294	187382	Orencia	Rheumatoid Arthritis	"Limited improvement after 4 months, developed...	2	15-Mar-14	35
161295	47128	Thyroid desiccated	Underactive Thyroid	"I've been on thyroid medication 49 years...	10	19-Sep-15	79
161296	215220	Lubiprostone	Constipation, Chronic	"I've had chronic constipation all my adu...	9	13-Dec-14	116

Fig 3.1(a) Dataset

3.2.1 Data Format & Features

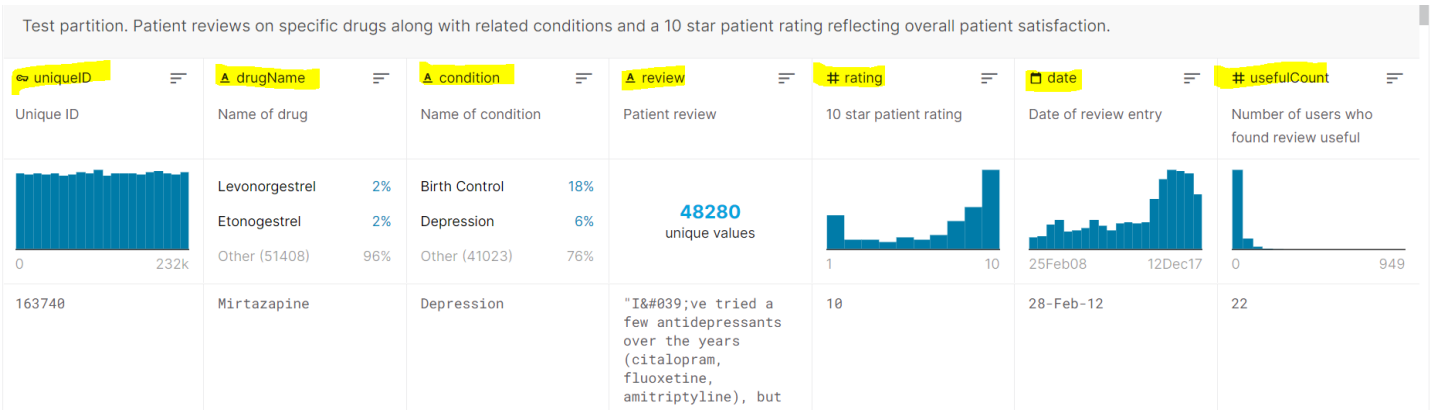


Fig 3.1(b) Features of dataset

The data is arranged in such a way that every patient who takes some medicine according to his medical condition has a unique ID. After the prescription of the medicine, each patient has to write a review on how the medicine was useful or helpful in his respective condition. After that, whatever patient takes the same medicine and finds the review useful clicks on the usefulCount which adds 1 to the variable.

3.3 Design of Problem Statement

Our goal is to build a recommendation model that can help doctors, whether experienced or inexperienced, people new to medicine, medical practitioners and students, and patients or consumers themselves prescribe and take the right medicinal drug. When treating an infectious disease, it is critical for doctors to choose a first-line drug. So the goal is to make it a little easier for doctors to prescribe a medicine by simply using the framework and searching about the diseases of the patient that they may not have had prior experience with.

Medicine recommendation systems that will act as a helping hand to the doctors or medicine practitioners in selecting the right medicinal drug for the patients according to their diseases/medical conditions. Hence, a strategy that will take care of the safety of the people consuming that medicine as well as providing exceptional services.

Task 1: Prepare the data and find n gram
Task 2: Perform Sentiment Analysis on Reviews and usefulcount
Task 3: Generate a recommendation system

3.4 Algorithm / Pseudo code of the Project Problem

3.4.1 Exploring the data/ Data Exploration

It is a process of inspecting or understanding the data and extracting useful insights or main characteristics of the data that would be used in the training of our model.
In the data exploration part, we will look at data types with various **visualization techniques and statistical techniques**. The aim of this process is to set the topic, preprocess the data so that it fits the objective of our project and also create various variables to fit into the model.

- ★ Exploring variables, comparing to check whether a person has provided multiple reviews for the condition, and making sure that there exists no such case where one consumer writes more than one review.
- ★ Checking for errors or unwanted fields present in the data.
- ★ Considering the fact that it is a recommendation system based on the patient reviews, it is not feasible to recommend when there is only one medicine for a particular condition. Therefore, we will analyze only the conditions that have at least 2 drugs per condition for safe and better recommendation of medicines.
- ★ Visualizing the data, Word Count Plots, Graphical Analysis
- ★ Correlation Analysis

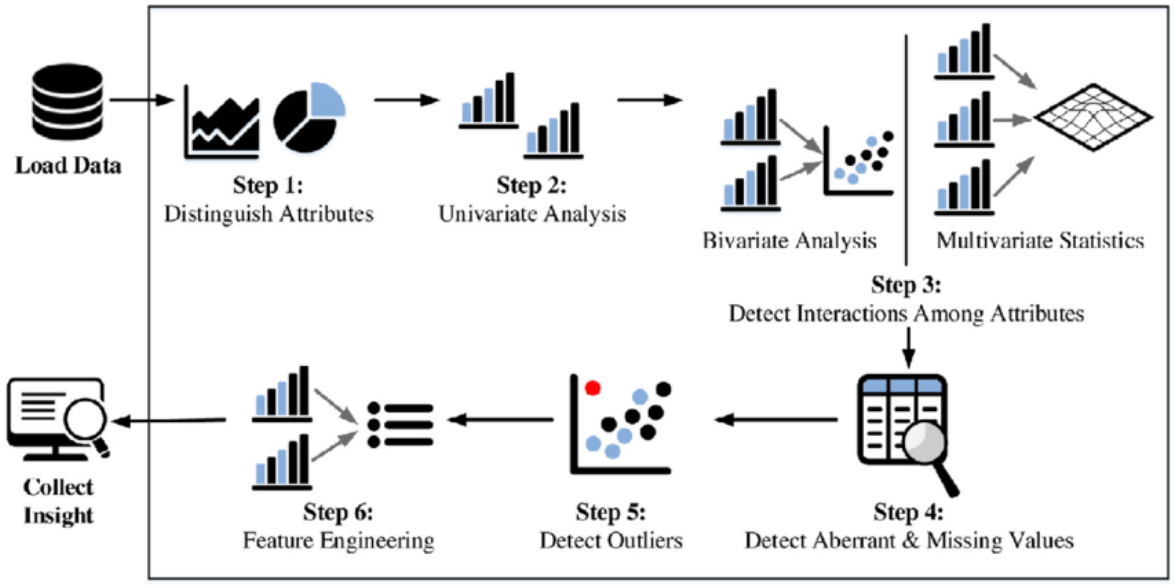


Fig 3.2 Exploratory Data Analysis

3.4.1.1 Using N Gram in our project

It is a proximate succession of n words gathered through a certain sample of subjective information where is not continuous and occurs within a range starting from 1 till infine. It is widely used in probability theory and data compression.
For example,
The word “information” is called a unigram .
When we start combining the words, “gathered information”, it becomes a bigram, consisting of two words.
Similarly, “preparation techniques like” is a 3-gram (trigram).
And “preparation techniques like checking” will be 4 gram because it considers four words and this can go on and on.
We will look for the perfect N gram which could give us the best and most useful information from reviews.
We will check through 1 ~ 4 grams which corpus best classifies emotions so that the model makes the best sense from the reviews given by various patients.

3.4.2 Data preprocessing

As we saw in may papers, the preprocessing stage started by looking into the data and finding whether their exists null or empty values in the records, or whether there are values that are repeating and are redundant, or if the data consists of unusual and out of context data fields and records, and then remove all such irregularities to make our data ready for statistical analysis. Now in our data, the null values were only seen in the

conditions column, so we need to remove them. Apart from this, since the unique ID must be unique for every patient we make sure that there are no duplicate values.

The reviews in our data is subjective information and requires cleaning for which we took care of the following -

- The process of **Tokenization** helps us to break down the continuous text/sentence thanks makes sense into chunks of single words known as tokens. This results in providing a list of tokens of the sentence and removing special characters like dots at the same time.
- Then we look for the **Stop words** in our token list.This process helps to remove such frequently used words from our subjective information because if a word is used so frequently it provides zero to no meaning to our target value since the word is no more distinct but very common. The stop words are words that are very generi, for example 'have','a','the', etc,which become insignificant.
- Another technique is **Negation Handling**. There are good chances of us removing important meaningful and insightful words that act like stop words, but actually have the power to change the meaning of the whole sentence, for example 'not','not so','too bad', etc. So to avoid such problems we take care of these conditions and provide a separate list of words that should not be considered as stop words for our preprocessing. This will be used in our project as well.this will help us to not remove important words from the reviews.
- One of the processes is **Stemming,** which helps to break and bring down the word to its root form by removing the affixes from that lexicon. The idea of stemming is to bring the word to its base form from where it cannot be broken down further. It increases the retrieval accuracy along with reducing the size of the index.
- Just Like Stemming, there is a process called **Lemmatization** in which we again break the words down to their atomic form properly by using vocabulary analysis of words. It also groups the various forms of the same words together. The base form returned is known as the lemma.

Fig 3.3 Data Preprocessing

3.4.3 Sentiment Analysis

Because there is no sentiment assigned to any of the reviews in the dataset, we must assign sentiment to the patient reviews and ratings. Its primary function is to determine the emotional tone of any text's body. This proposed sentiment analysis, or what is commonly known as opinion mining, for the given reviews is very helpful and important as it makes the given data more valuable in terms of the use of various parties involved seeking the recommended medicinal drug through our model. This is a popular way to categorize the medicinal drug as useful or not taken by the respective patient.
Emotion analysis will be done using word dictionary sentiment analysis or using SentiWordnet of NLTK module, to label our pre-processed reviews data as positive or negative.
The formula that we will be using is :

Positiv_ratio = the number of positive words / (the number of positive words+the number of negative words)

Now analysing the Positiv_ratio :

1. Less than 0.5 then it is classified as negative
2. Values having ratio more than 0.5 are considered positive
3. If it comes out to be exactly 0.5 then it is classified as neutral

This process will help us identify the emotions each patient meant while writing the review for the medicine he/she took for their medical condition. This will easily categorize the drug as per the patient's review who consumed the respective medicine in their medical condition.

	uniqueID	drugName	condition	review	rating	date	usefulCount	review_clean	sentiment	day	year	month	sentiment self 1	sentiment self 2	count_sent	count_word	count_unique_word	c
53433	76567	Lorcaserin	Obesity	"First let me tell you if you go on belviq.com...	9	2017-05-10	3	first let tell go belviq com site print coupon...	1	10	2017	5	-1.875	-3.375	1	66	56	
196595	122920	Ethinyl estradiol / norgestrel	Birth Control	"This pill is honestly so terrible. The first ...	1	2016-06-30	4	pill honest terribl first week began take star...	0	30	2016	6	1.375	-0.750	1	50	43	
71709	29804	Topiramate	Bipolar Disorde	"I started taking this medicine several weeks ...	4	2009-07-23	56	start take medicin sever week ago combin celex...	0	23	2009	7	0.250	-0.875	1	50	46	
56697	114314	Nicotine	Smoking Cessation	"Have been an on and off smoker for more than ...	10	2014-01-19	57	smoker year start lozeng work fantast take fee...	1	19	2014	1	-0.125	0.125	1	41	37	
46291	137603	Ocular lubricant	Eye Redness	"I make no tears. If it were not for this ins...	10	2013-11-06	15	make no tear not insert would put disabl list ...	1	6	2013	11	-1.125	-1.000	1	30	28	

Fig 3.4(a) Sentiment Analysis

count_word	count_unique_word	count_letters	count_punctuations	count_words_upper	count_words_title	count_stopwords	mean_word_len	season
66	56	368	40	4	9	60	4.590909	1
50	43	311	19	7	11	52	5.240000	2
50	46	288	37	5	9	46	4.780000	2
41	37	240	25	7	13	49	4.878049	4
30	28	160	8	4	8	28	4.366667	3

Fig 3.4(b) Sentiment Analysis

3.4.4 Modelling

At the modelling part, emotion or sentiment analysis using NLTK's Wordnet and SentiWordnet is done to make the most sense out of the reviews written by various patients. To make the best sense out of the reviews we used n-gram 1~ 4 grams and then applied deep learning, etc. To compensate for the natural language processing limitation we used LightGBM which is a distributed and fast machine learning model which is based on decision trees, that eventually results in the increase in efficiency of the model while reducing the memory usage. The reliability of our review data was further secured through a usefulcount feature of the dataset that allows users to like the review which was helpful to them also which in turn lets us know if the written review was actually of any use to them or not. This feature helps us identify the reviews which are not that reliable.

3.5.5 Performance Metrics and Results

All these steps include the sentiment analysis of our data and then using n-gram to make the most sense out of the reviews, followed by a LightGBM model that allows us to measure the total mean value that is predicted for each drug under certain medical disease/condition and according to the order of these values, we check which medicine is the most appropriate to consume. Hence making the process of recommending first grade medicine not only for doctors or any health care workers but also for the patients themselves easier and effective. We find the accuracies and the confusion matrices for the models for better understanding of the results

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{all predictions}}$$

Classification metrics

When performing classification predictions, there's four types of outcomes that could occur.

- **True positives** are when you predict an observation belongs to a class and it actually does belong to that class.
- **True negatives** are when you predict an observation does not belong to a class and it actually does not belong to that class.
- **False positives** occur when you predict an observation belongs to a class when in reality it does not.
- **False negatives** occur when you predict an observation does not belong to a class when in fact it does.

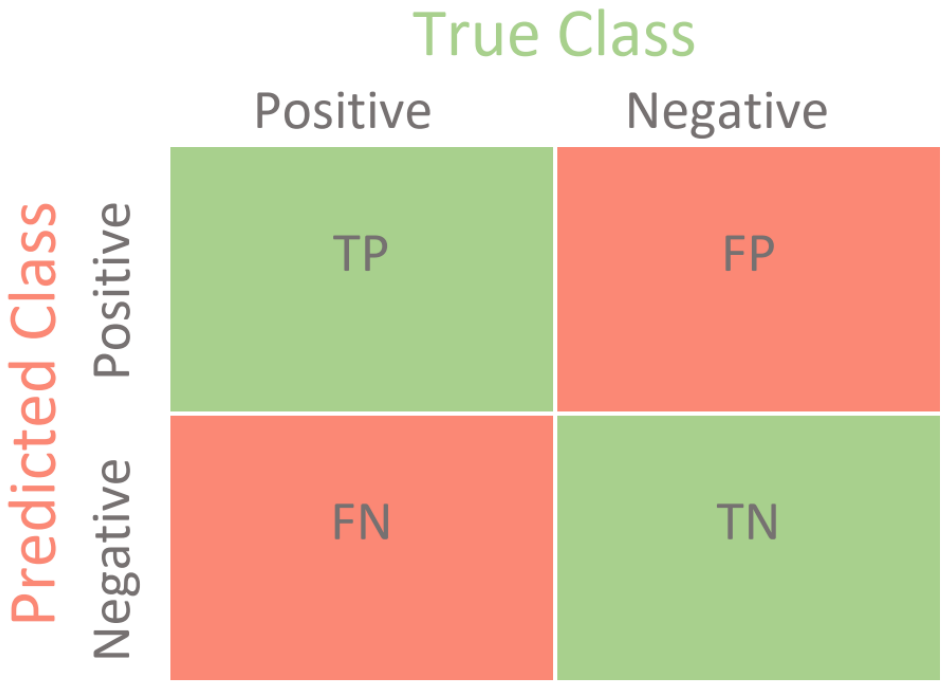


Fig 3.5 Format of Confusion Matrix

3.5 Models

3.5.1 LightGBM

Light GBM is characterized to be a gradient boosting framework and uses a tree based learning algorithm. In particular, it uses two techniques called GOSS i.e. Gradient-based One Sided Sampling and EFB i.e. Exclusive Feature Bundling, which basically makes the algorithm more accurate and also reduces memory usage in our algorithm and overcomes the shortcoming of histogram-based algorithms.

Light GBM spreads trees vertically while other algorithms grow trees horizontally so that Light GBM grows tree leaf-wise while other algorithms grow level-wise. A Leaf-wise algorithm can reduce more losses than a level-wise algorithm. It selects a leaf with maximum delta loss to grow. It may increase the complexity of the model and may lead to overfitting in small databases.

Light GBM is named 'Light' because of its high speed and can handle large data sizes and takes lower memory to execute. Another reason Light GBM is popular, because it mainly concerns the accuracy of results.

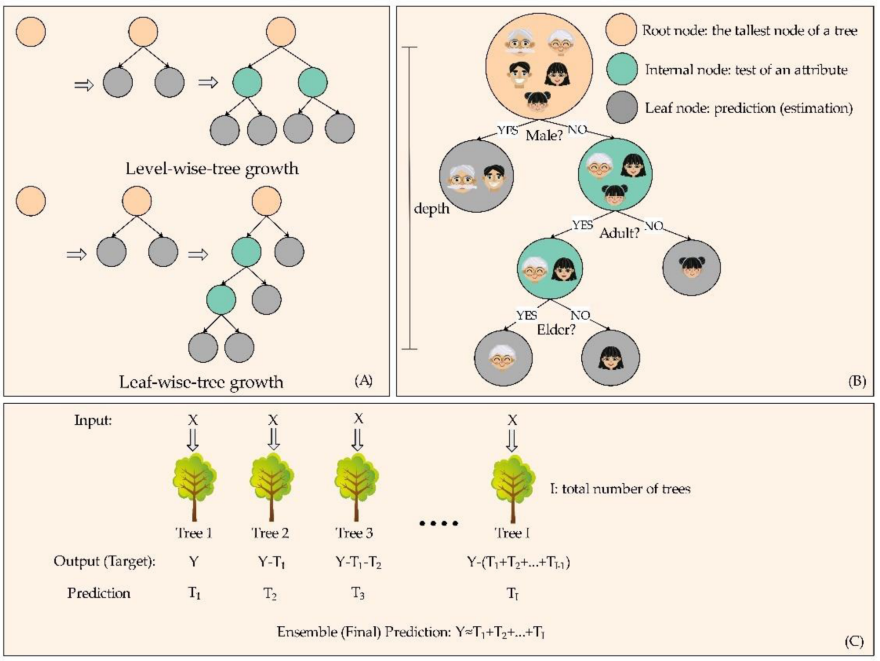


Fig 3.6 Machine Learning LightGBM Model

3.6 Flow Graph of the Project

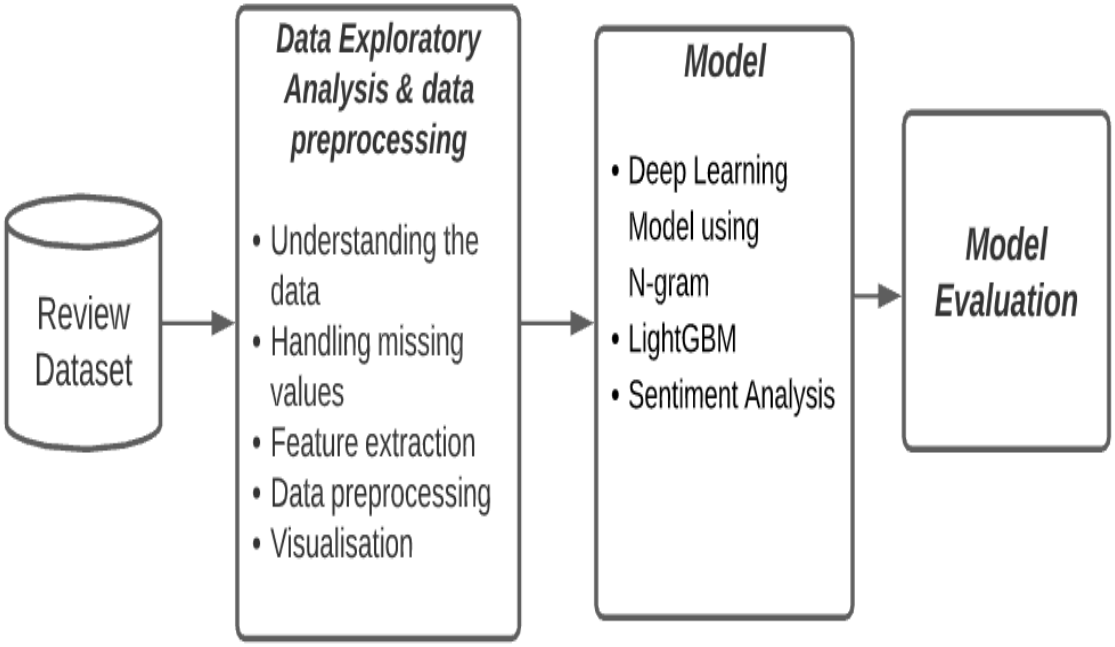


Fig 3.7 Major Project flow chart

1. **Data Exploration Analysis & Data Preprocessing:**

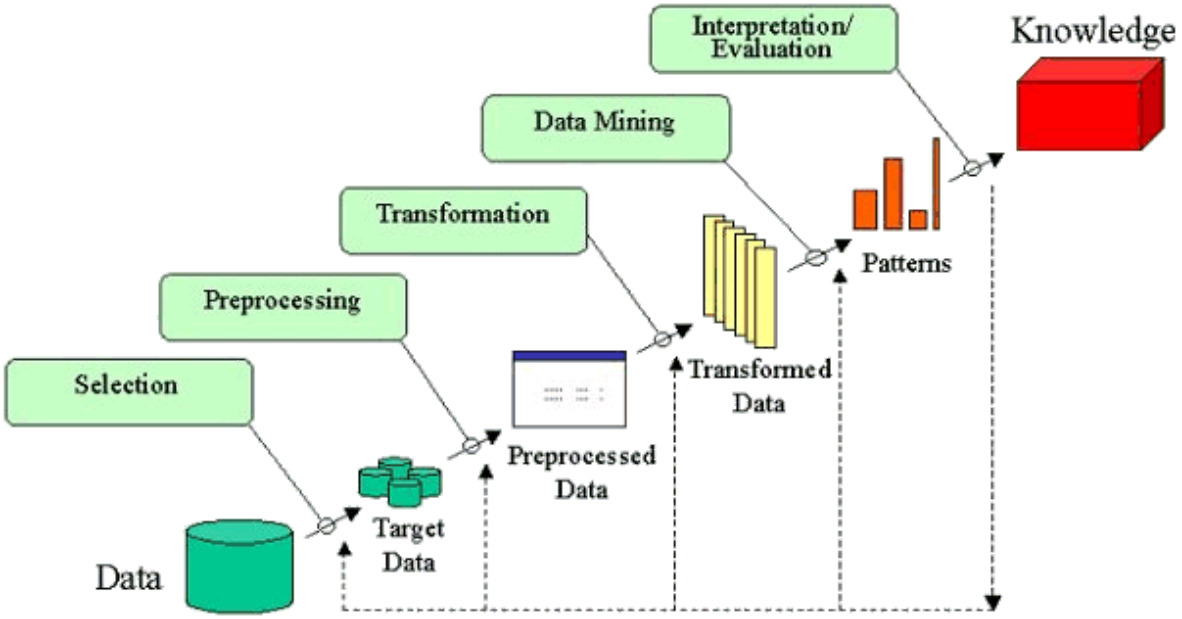


Fig 3.8 The data exploration part is the first part of data analysis that is used to explore data to gather insights from the start.

2. **Model:**

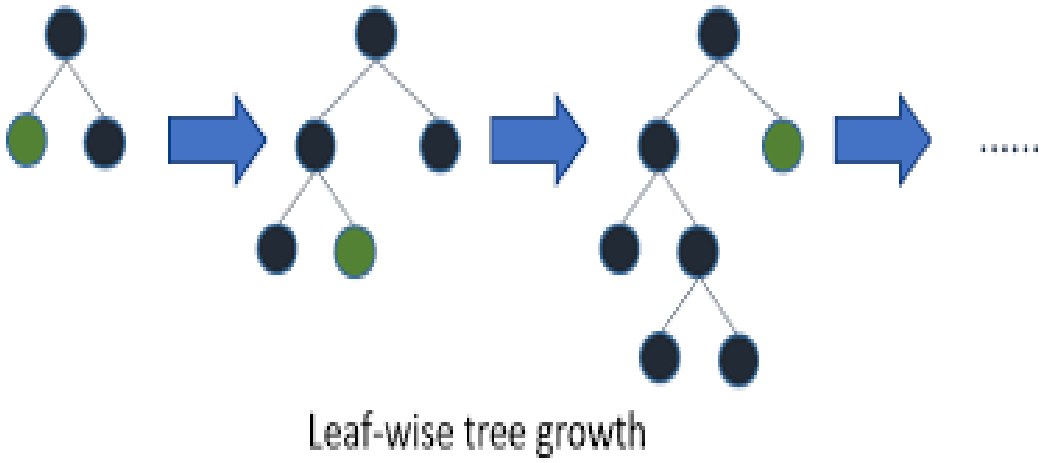


Fig 3.9 The Leaf-wise growth in a LightGBM Model

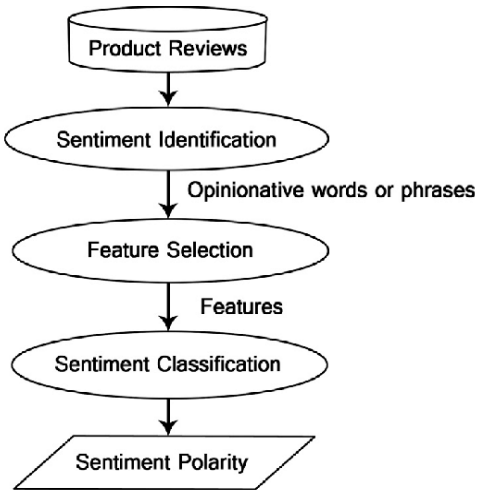
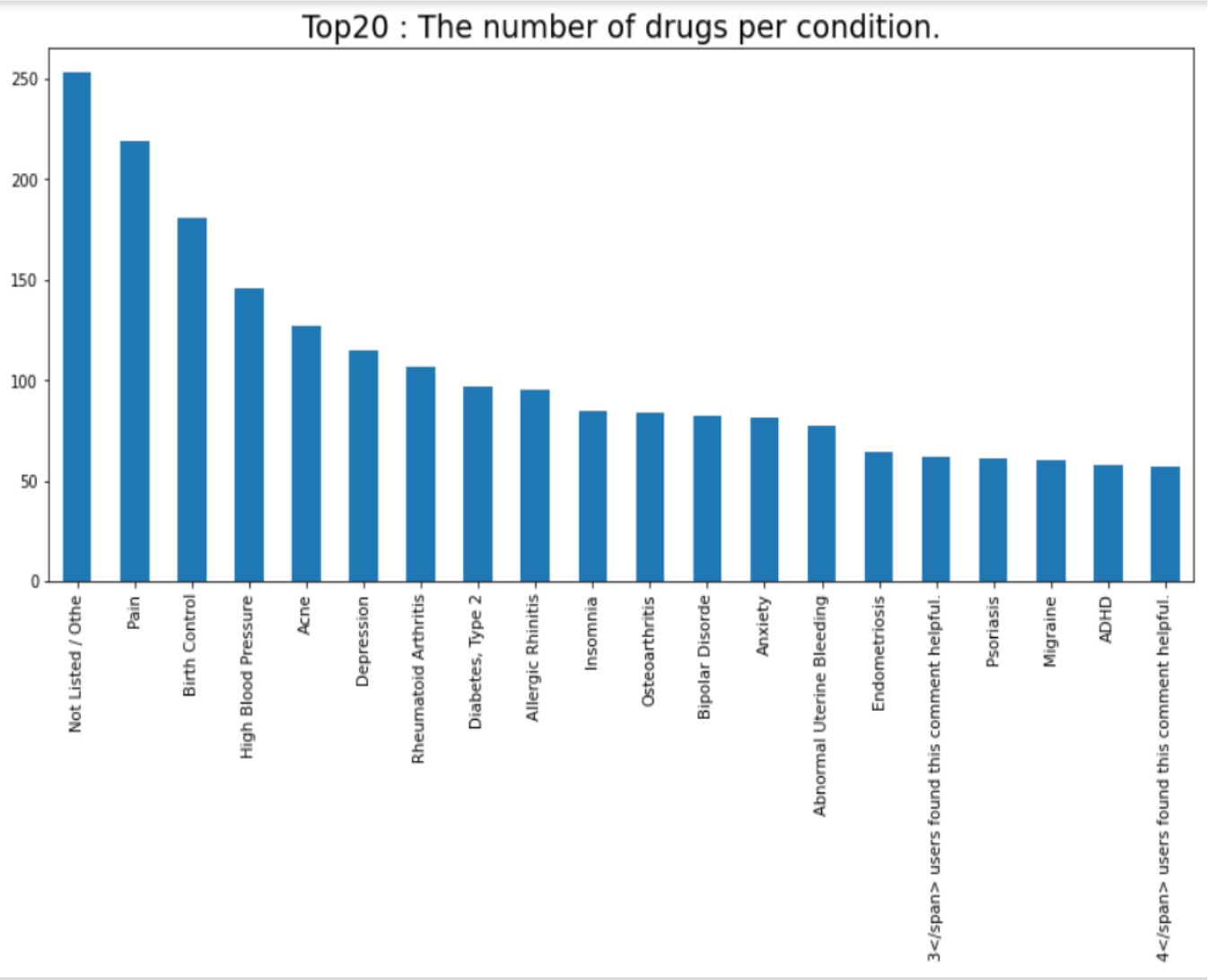
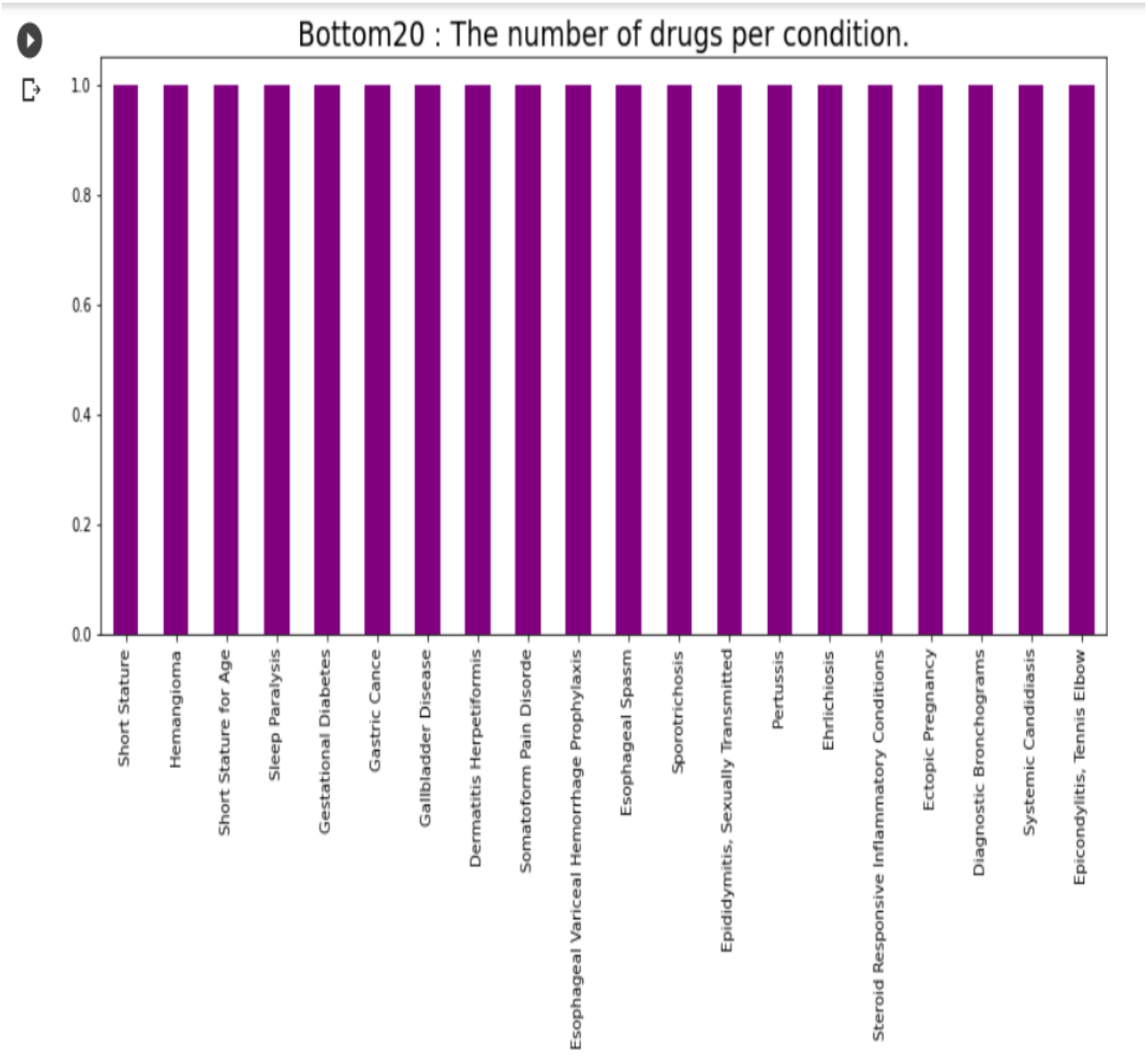


Fig 3.10 Flow chart for Sentiment Analysis

3.7 Screenshots of the various stages of the Project



Graph 3.1 We begin by grouping the data according to the condition of the patients and finding out the maximum number of drugs present for a particular medicinal condition.



Graph 3.2 The next step was to again group the medicines related to different conditions and then plot them in a way that we can find the bottom 20 conditions with the least number of medicines per condition as prescribed by the patients.

```
plot_wordcloud(df_all["review"], title="Word Cloud of review")
```

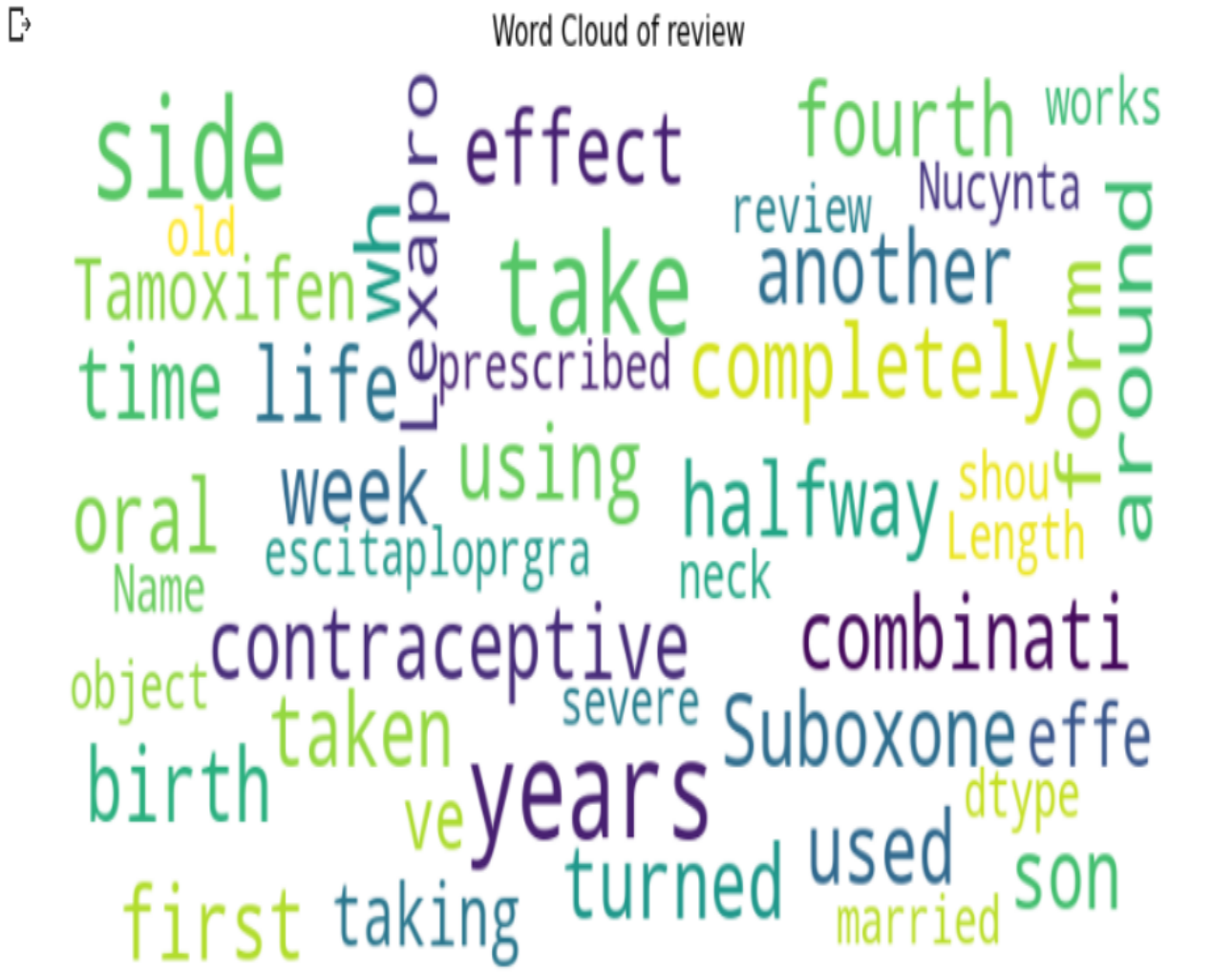
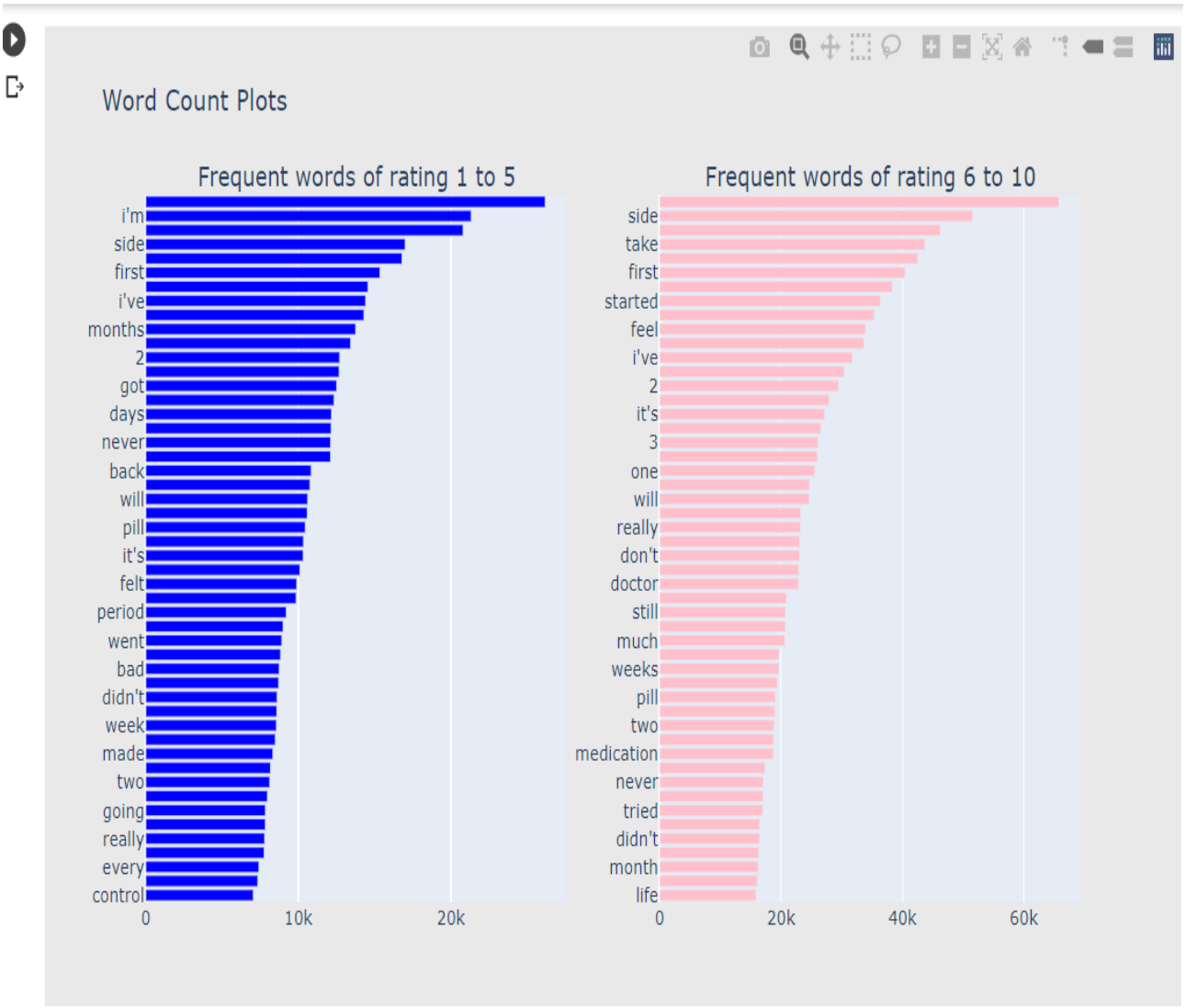
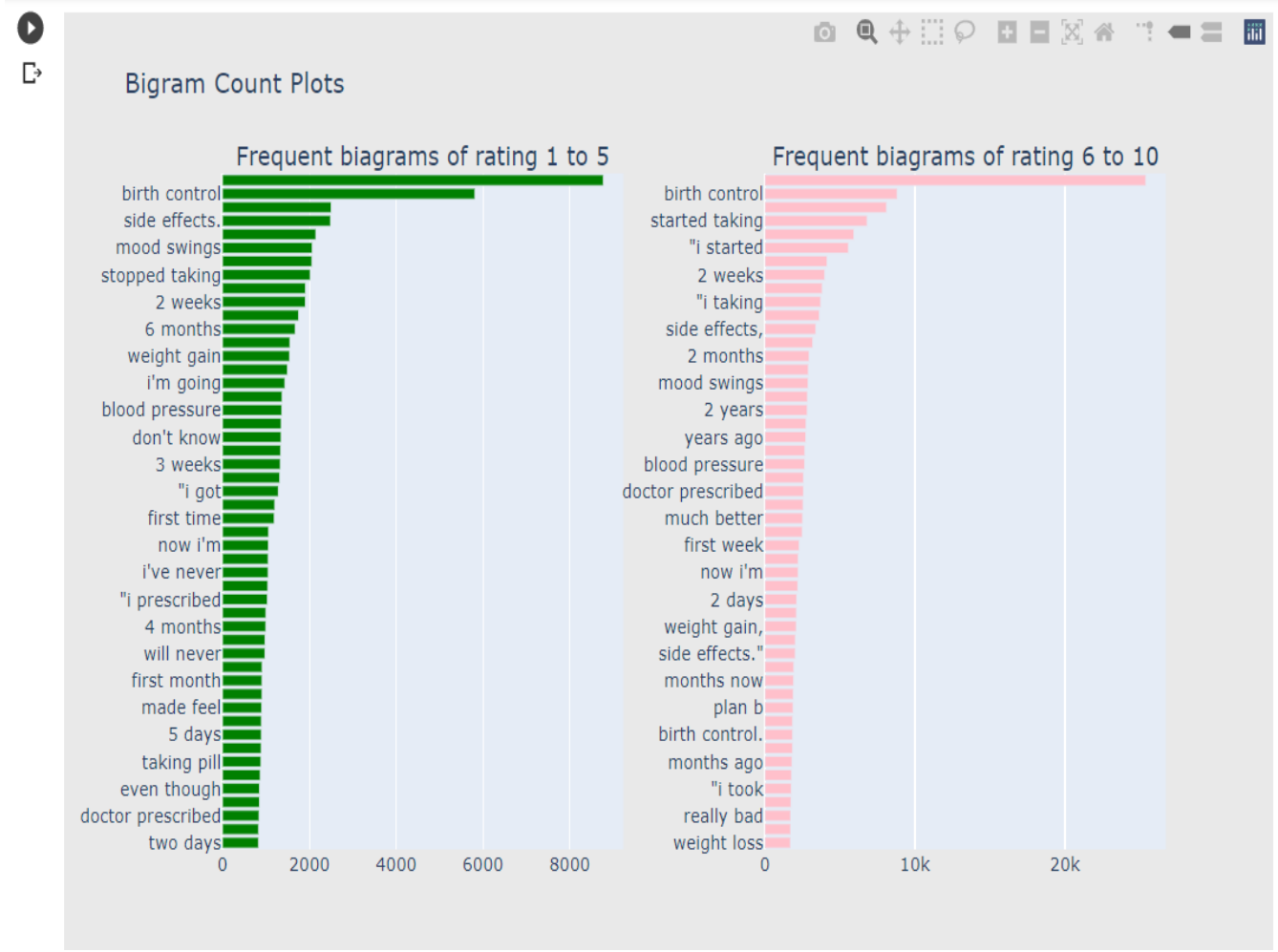


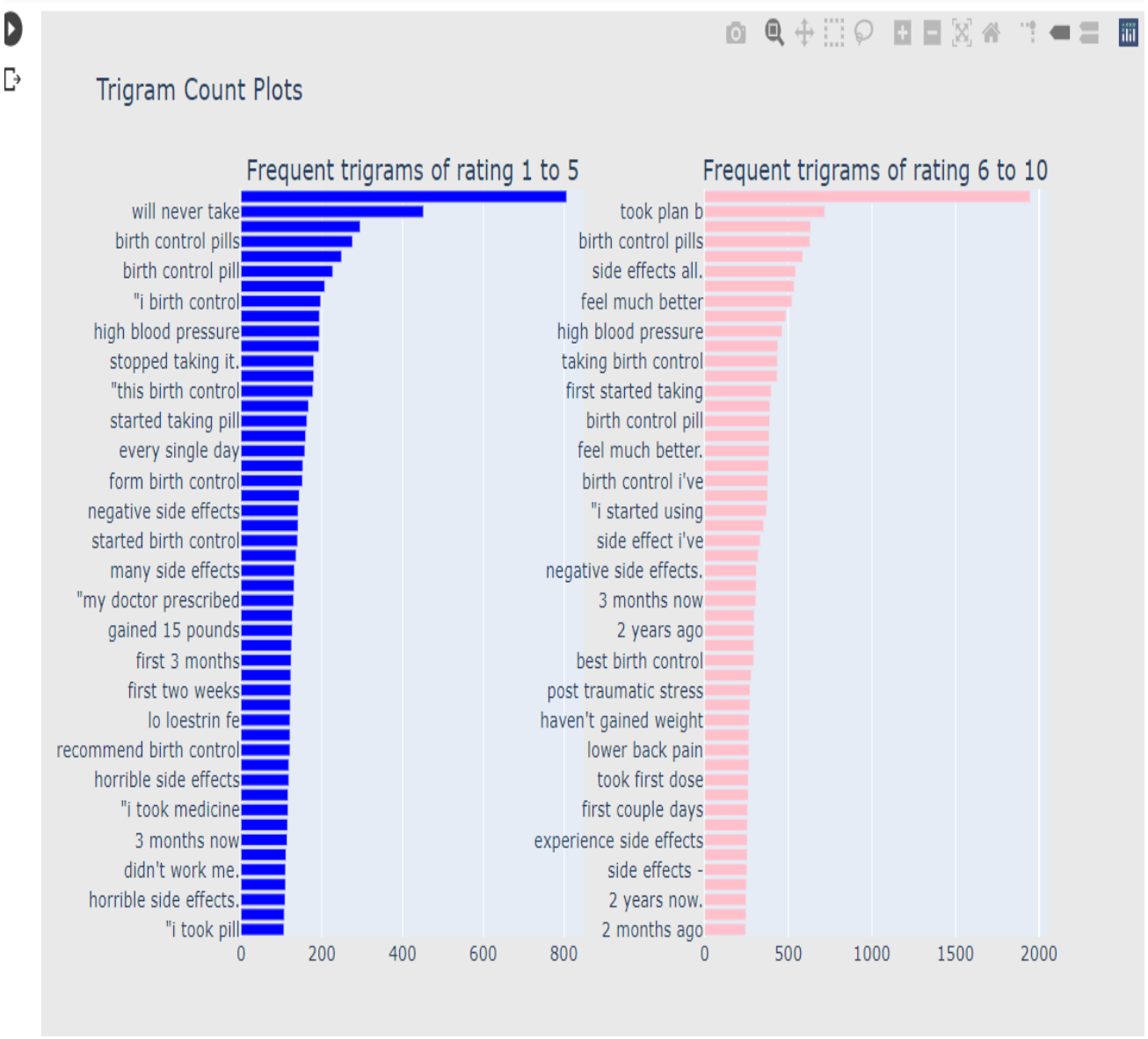
Fig 3.11 Next we used the wordcloud library which is a data visualization technique. In this the size of each word indicates its importance as well as its frequency.



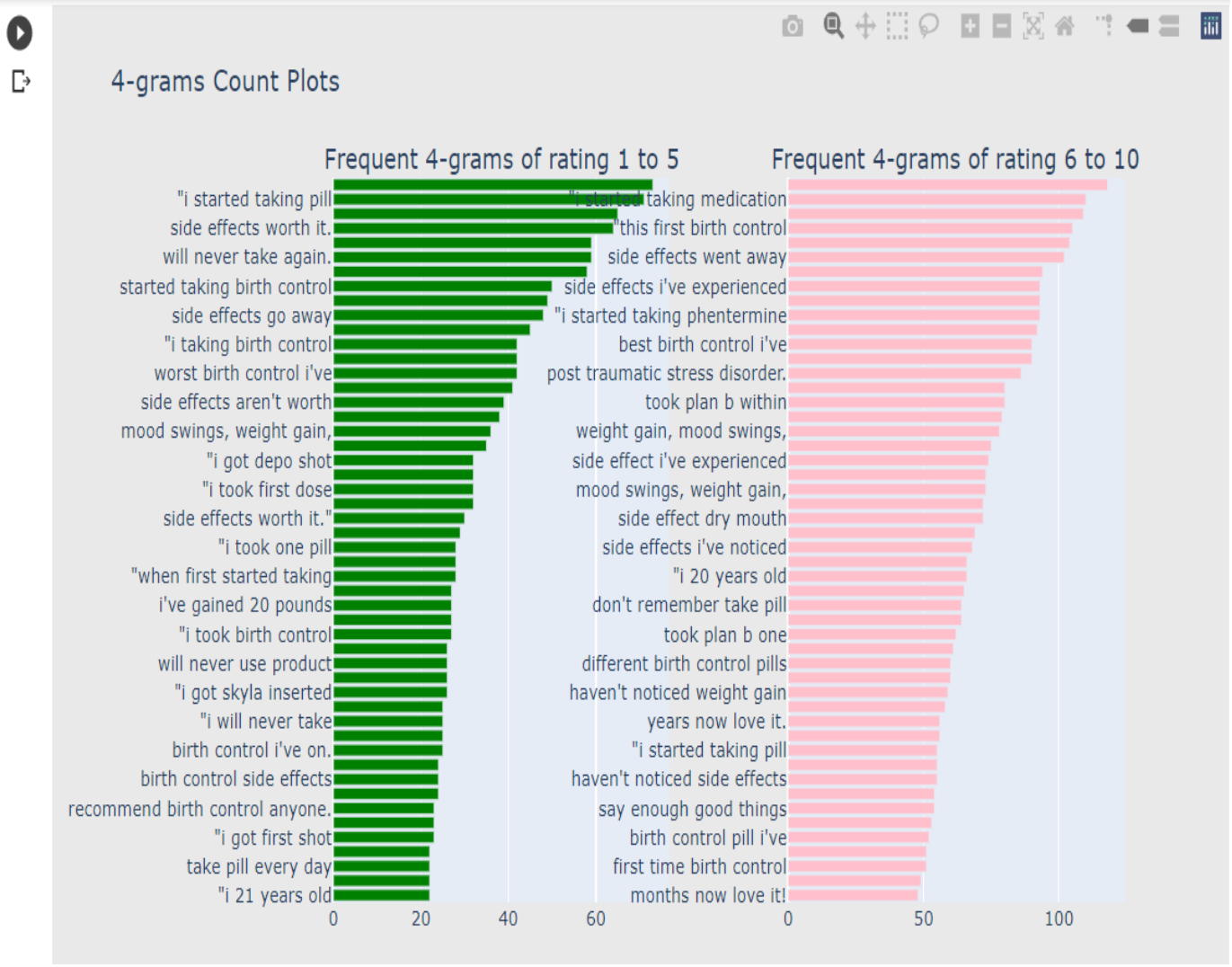
Graph 3.3 This is a word count plot which depicts the frequency of occurrence of words. The blue graph shows the occurrence between 1 to 5 and the pink shows the occurrence of the words ranging from 6 to 10.



Graph 3.4 This graph shows the frequency of bigrams i.e. two words taken at a time for our n-gram analysis. The green graph displays the words with 1 to 5 times occurrences and pink shows the words with 6 to 10 occurrences.

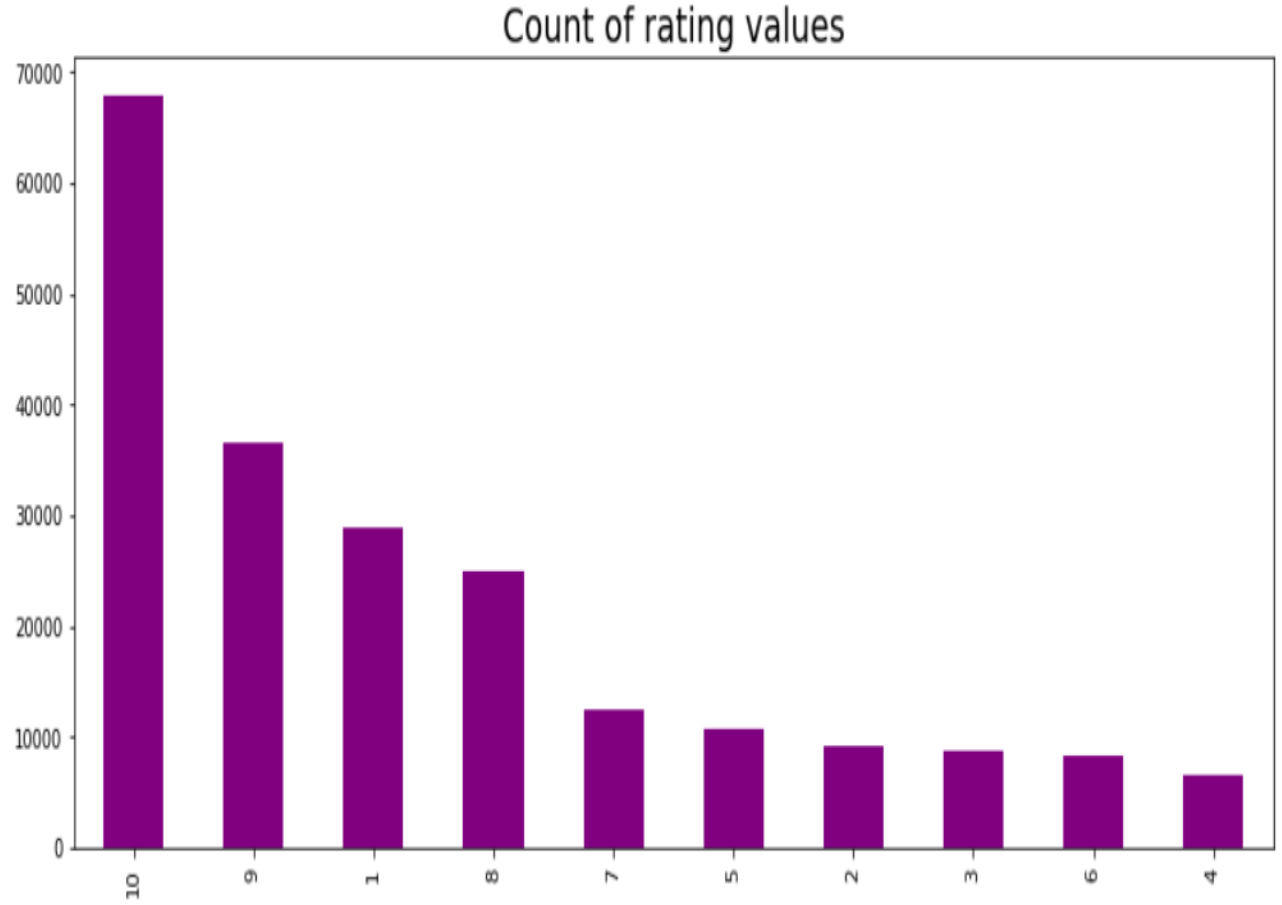


Graph 3.5 These graphs show the trigram occurrence inside the text. The blue graph shows the phrases occurring 1 to 5 times and the pink graph shows the phrases occurring 6 to 10 times.

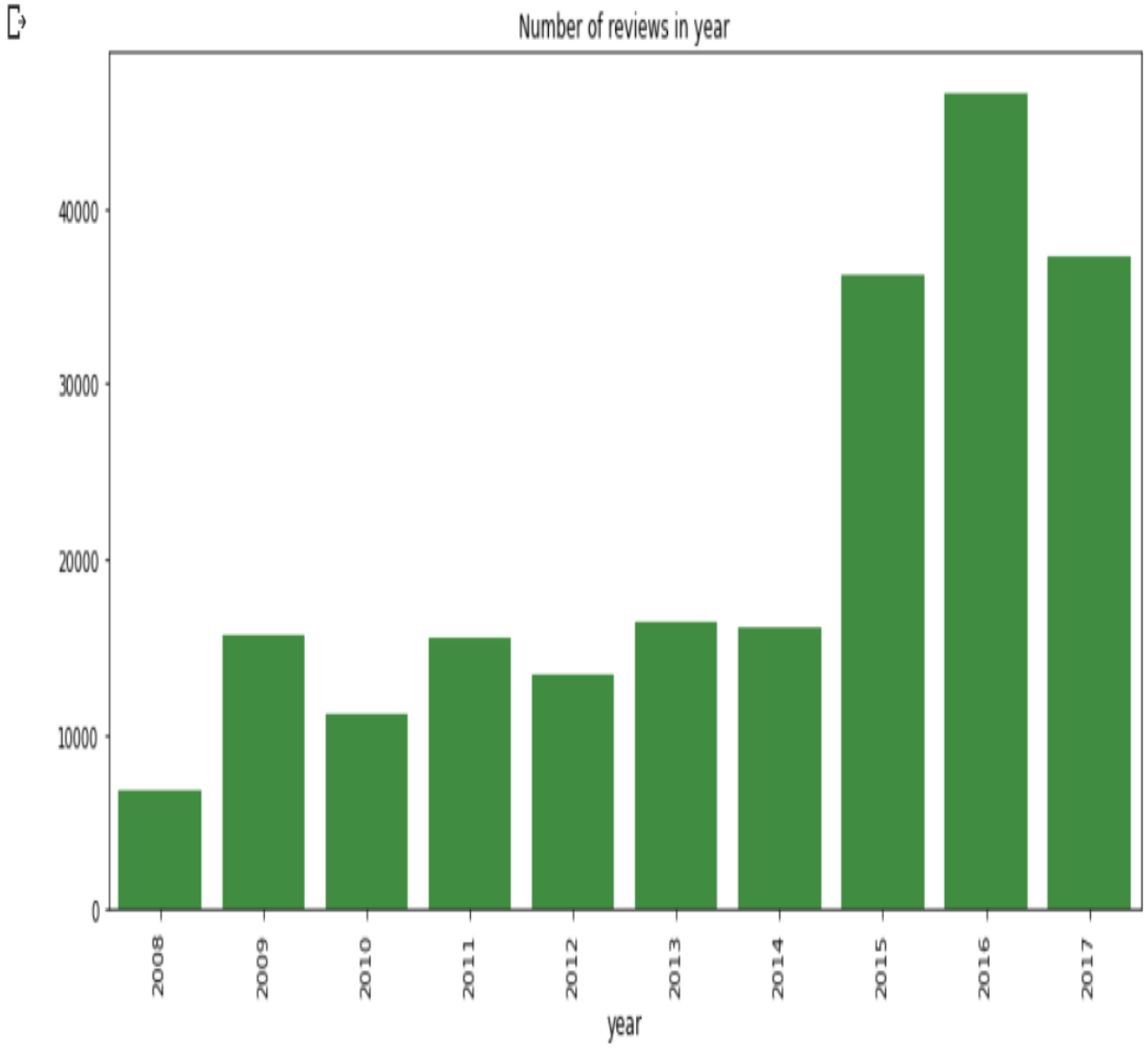


Graph 3.6 The frequency of occurrence of the qualgrams is represented in these graphs with phrases ranging 1 to 5 according to their occurrence in green and the other phrases ranging from 6 to 10 in the pink graph.

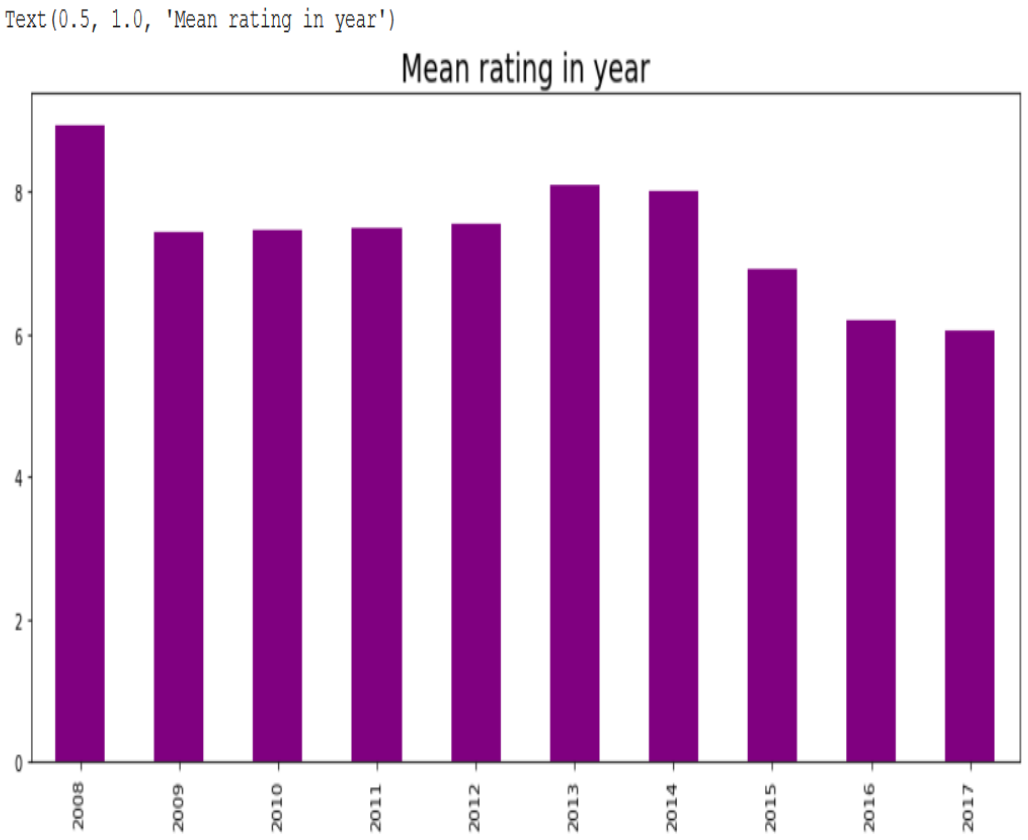
```
Text(0.5, 1.0, 'Count of rating values')
```



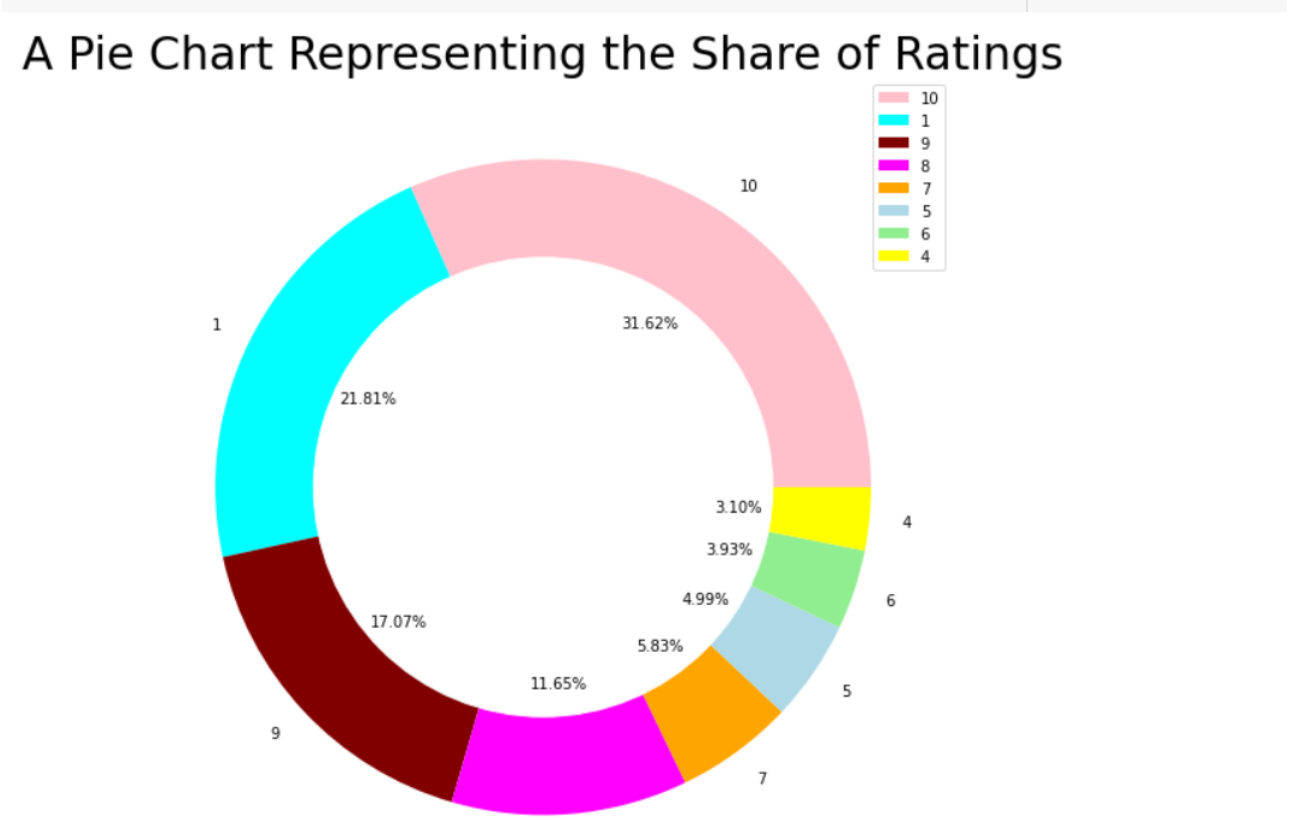
Graph 3.7 In this step we count the number of reviews that have been given 1 to 10 ratings in descending orders. This allows us to identify the number of reviews that have been helpful to others apart from the patient writing the review.



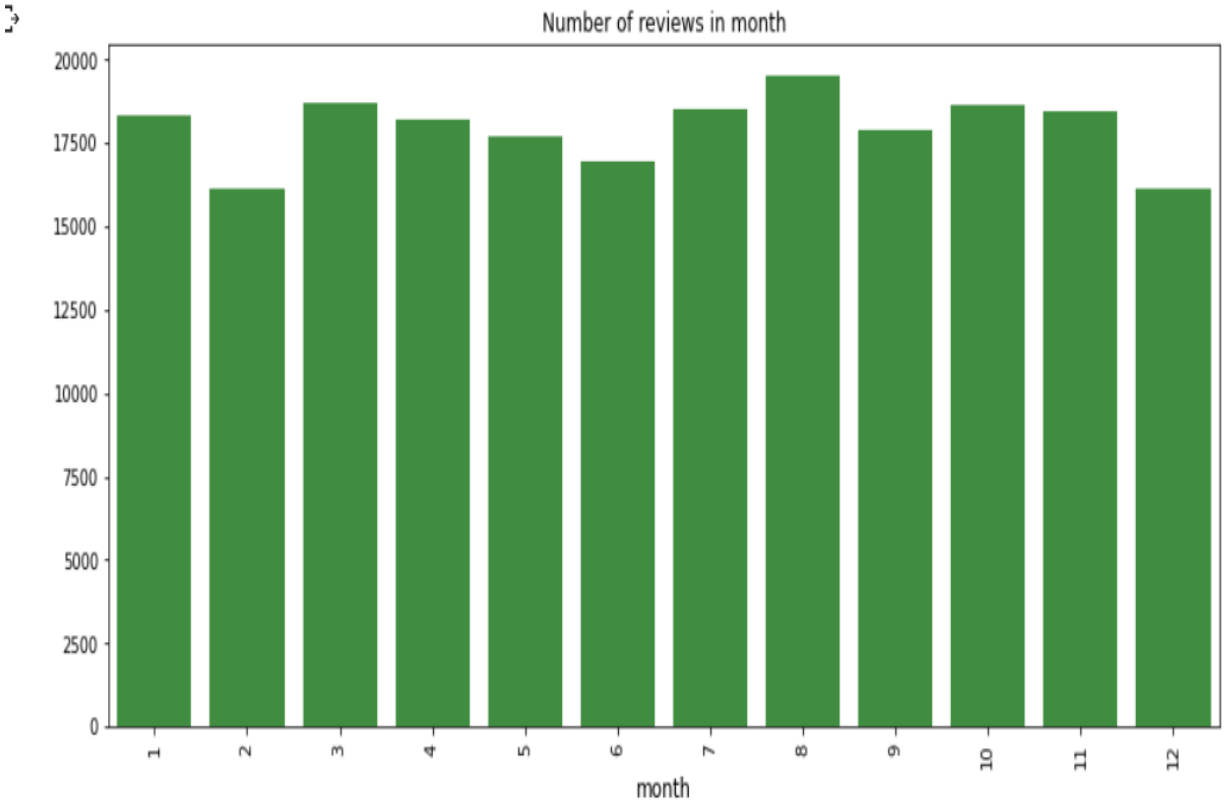
Graph 3.8 This graph displays the count of reviews that are being written each year. The highest reviews were written in 2016.



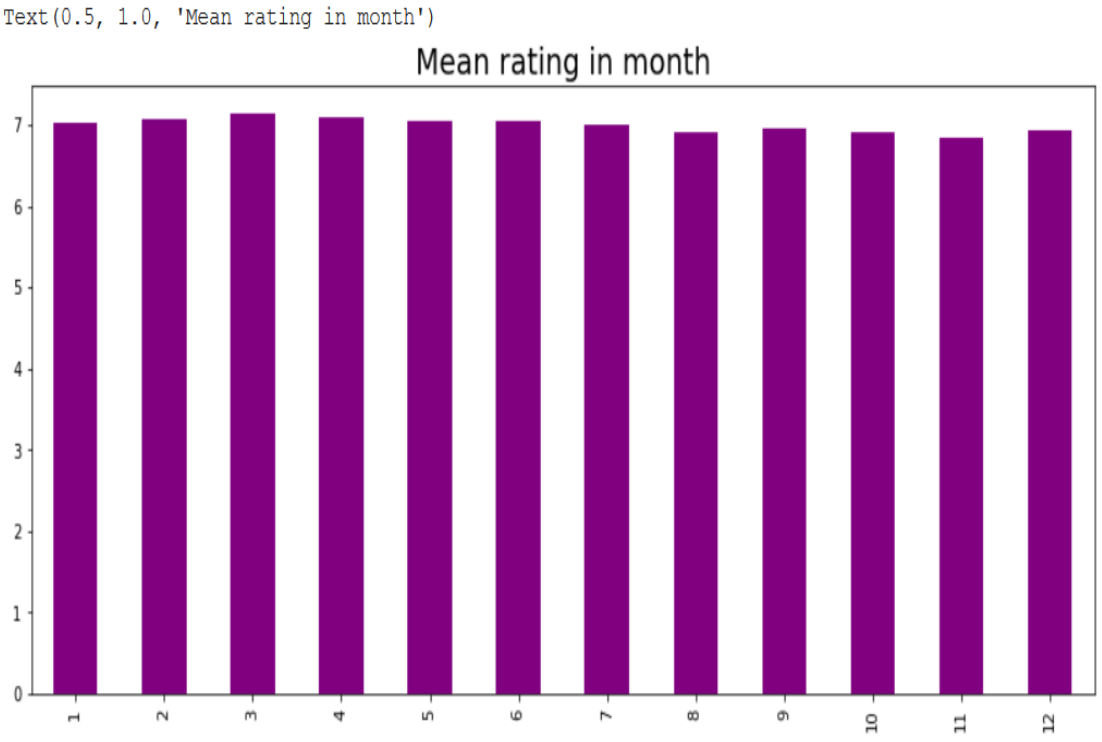
Graph 3.9 The graph displays mean ratings each year.



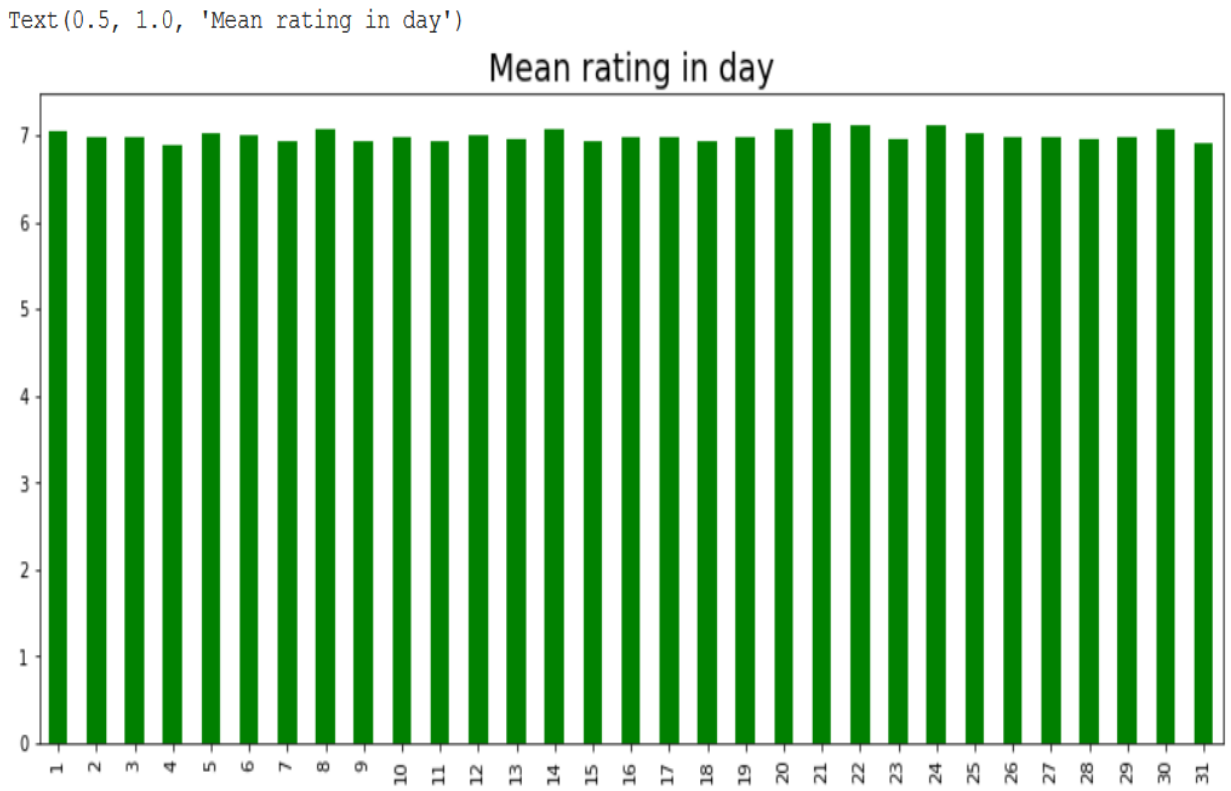
Graph 3.10 This is the pie chart corresponding to the share of each rating.



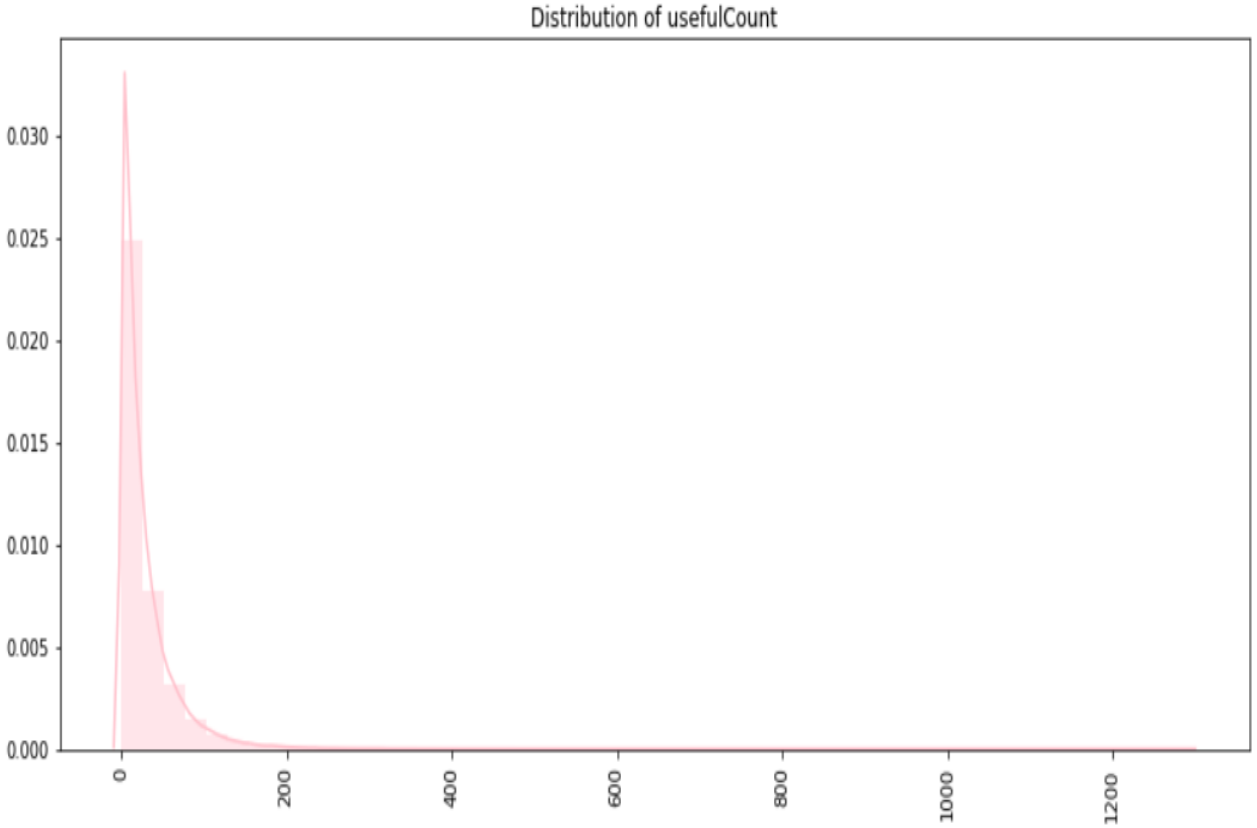
Graph 3.11 The graph represents the number of reviews written each month in the dataset.



Graph 3.12 This is the graph representing the mean ratings of the reviews every month.



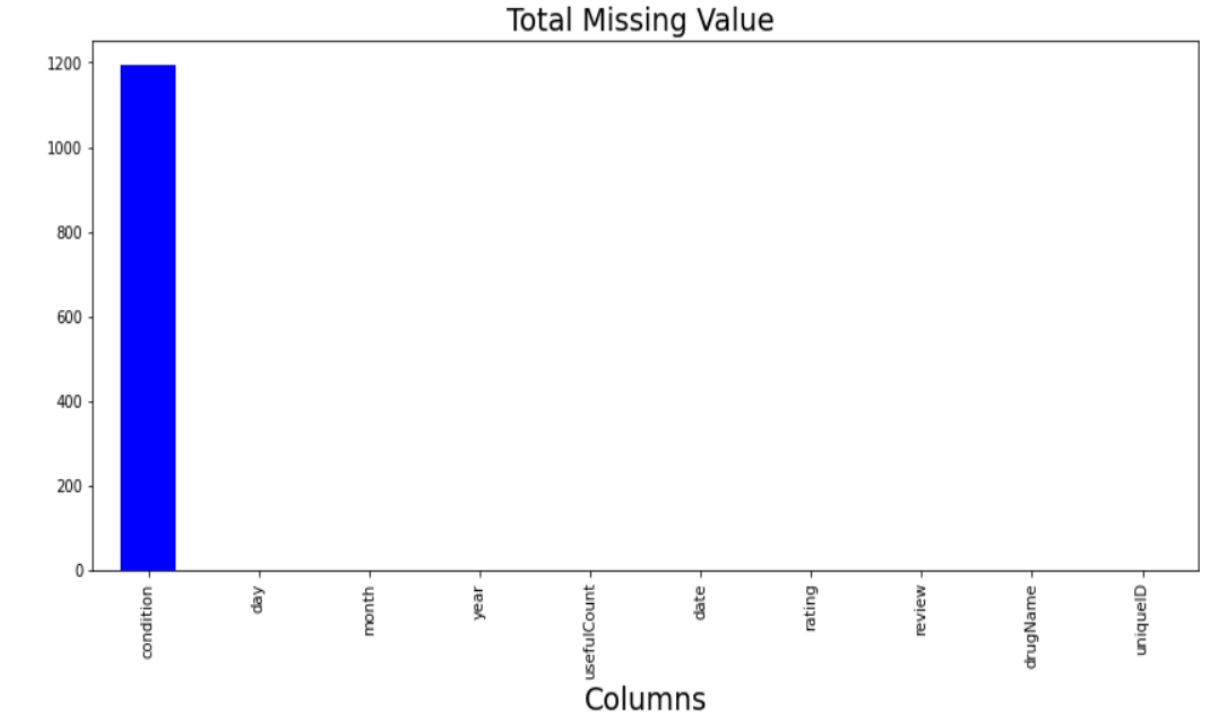
Graph 3.13 This graph displays the mean ratings in a day.



Graph 3.14 The distribution of the variable usefulCount, the count of the number of people who found the review given by a particular patient useful to increase the reliability of the

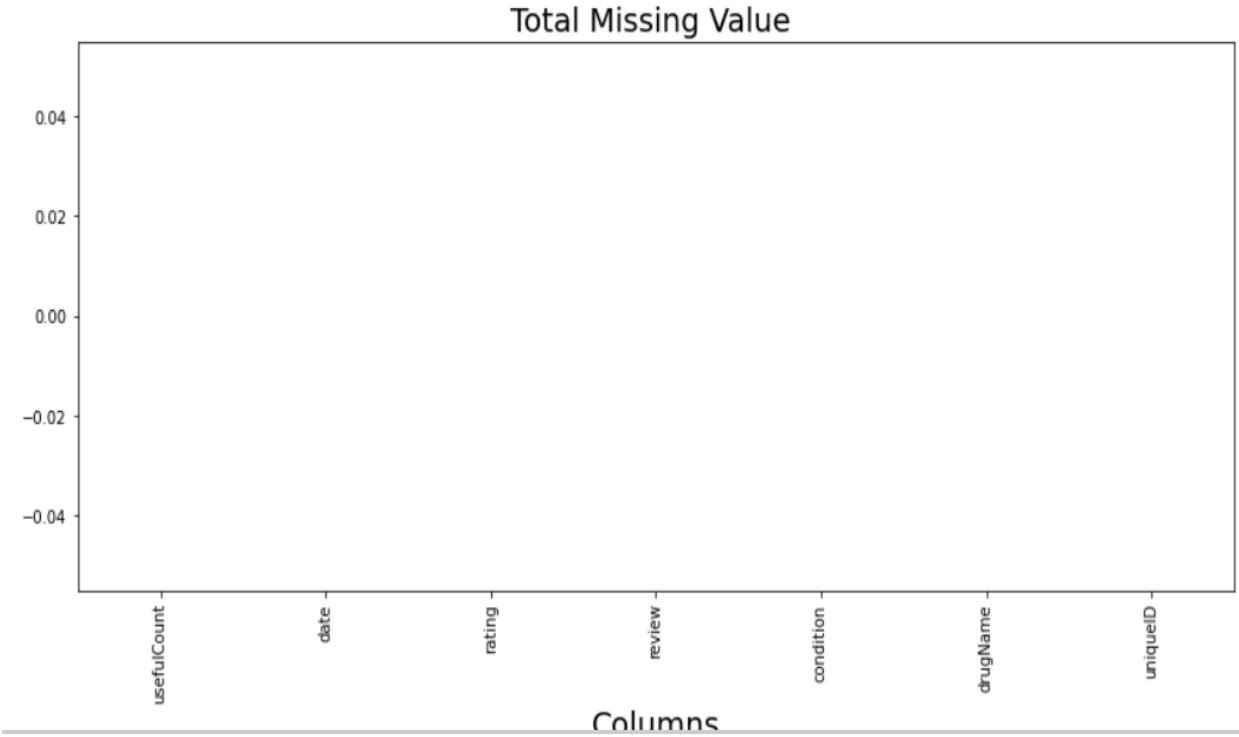
data.

```
Text(0.5, 1.0, 'Total Missing Value ')
```



Graph 3.15 The total missing values are represented using this graph in our dataset and the percentage comes out to be 0.55%.

```
Text(0.5, 1.0, 'Total Missing Value ')
```



Graph 3.16 The graph represents the number of missing terms in each column.

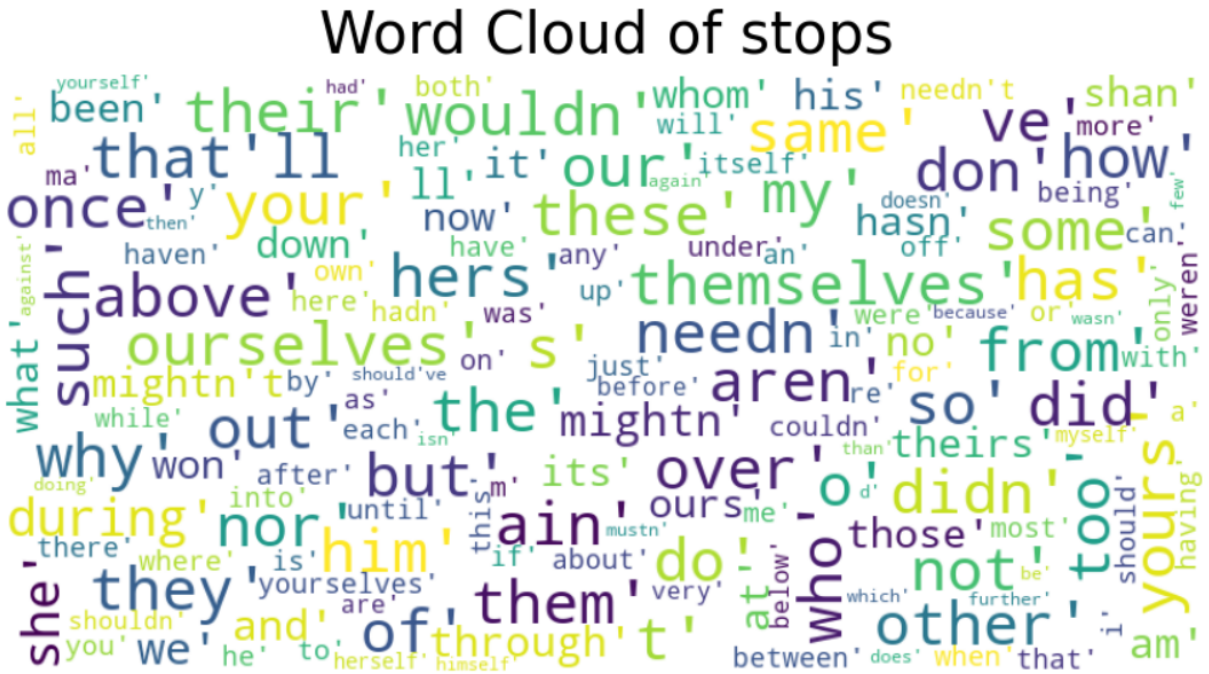


Fig 3.12 This is a wordcloud representation of the stop words that need to be removed from our data for it to consume less memory and make it more accurate while predicting.

Chapter 04: PERFORMANCE ANALYSIS

4.1 Screenshots of the Performance Analysis

LightGBM
When sentiments are not considered. Target: usefulcount

```
solution = df_test['sentiment']
confusion_matrix(y_pred=sub_preds, y_true=solution)
```

```
array([[    0, 15942],
       [    0, 37072]])
```

```
from sklearn.metrics import accuracy_score
accuracy_score(solution, sub_preds)
```

```
0.6992869807975252
```

Fig 4.1 Accuracy of LGBM model with target value usefulcount

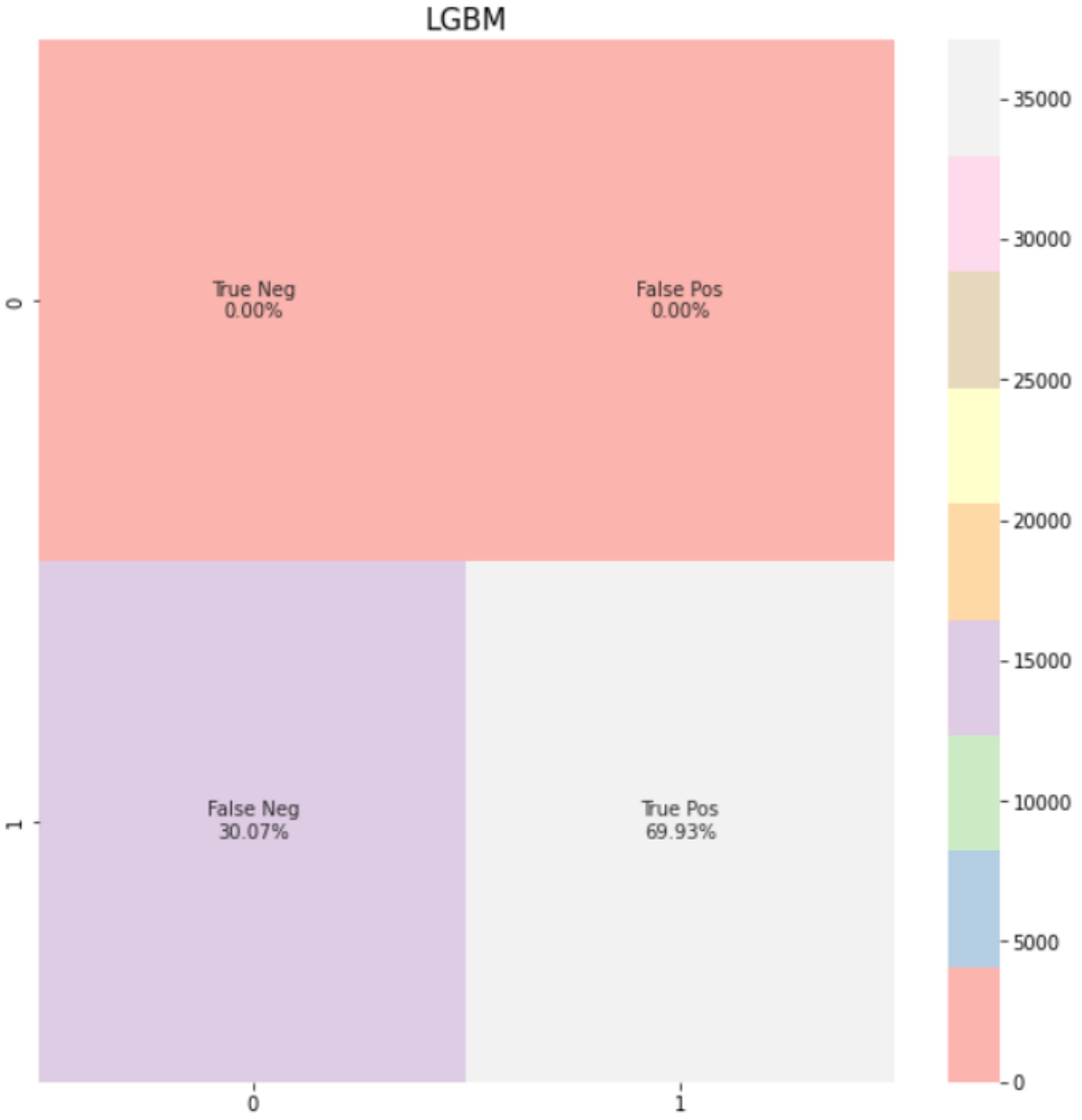


Fig 4.2 This is the confusion matrix that we have finally achieved at the end of our LightGBM model without sentiments.

LightGBM

When Sentiment Analysis is performed and target values are sentiments

```
confusion_matrix(y_pred=sub_preds, y_true=solution)
```

```
array([[10524,  5418],
       [ 2784, 34288]])
```

```
accuracy_score(solution, sub_preds)
```

```
0.8452861508280831
```

Fig 4.3 Accuracy of LGBM with sentiment analysis

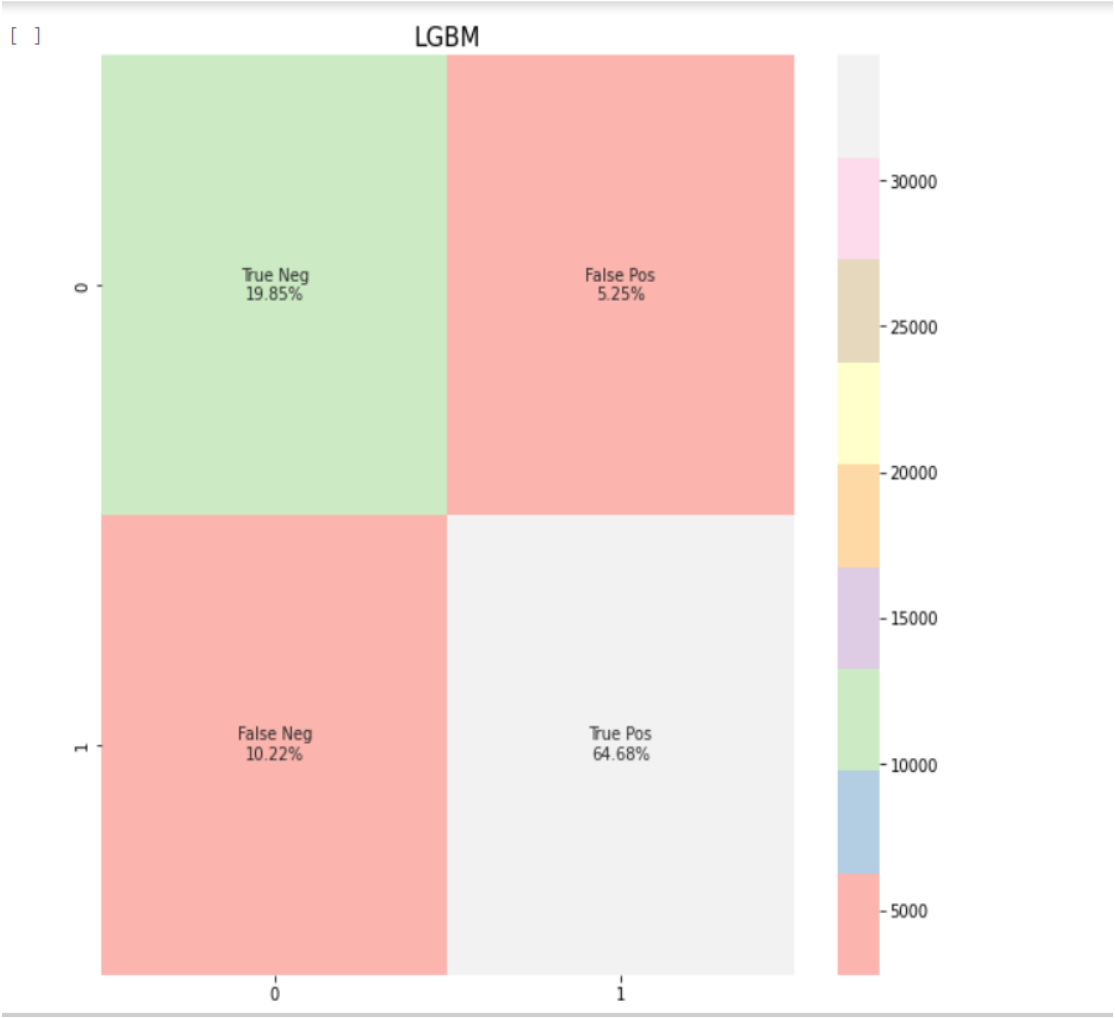
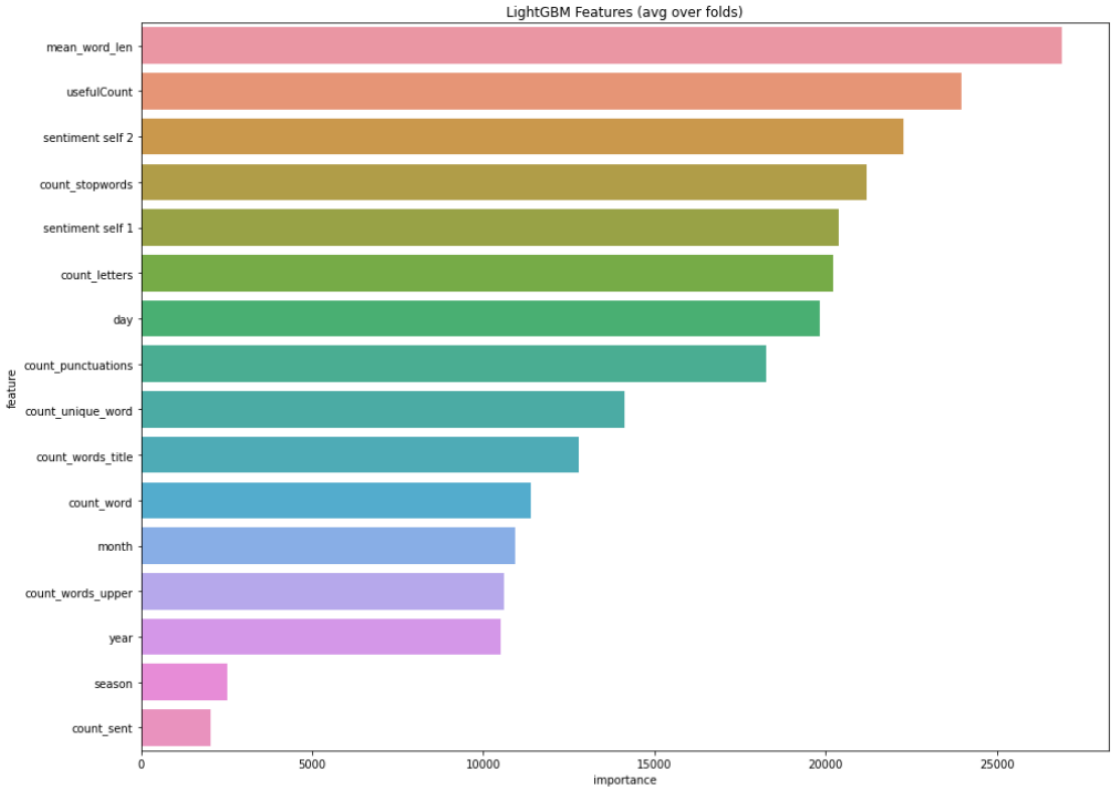


Fig 4.4 This is the confusion matrix that we have finally achieved at the end of our LightGBM model with sentiments as targets



Graph 4.1 LightGBM Features (avg over folds)

Predicted Results

The following results were calculated by adding the df_test of LightGBM on Sentiments Model and the Sentiments via dictionary and then multiplying the result with normalized usefulcount -

```
def userful_count(data):
    grouped = data.groupby(['condition']).size().reset_index(name='user_size')
    data = pd.merge(data,grouped,on='condition',how='left')
    return data
#_____________________________________________________________
df_test =  userful_count(df_test)
df_test['usefulCount'] = df_test['usefulCount']/df_test['user_size']
```

```
df_test['machine_pred'] = sub_preds

df_test['total_pred'] = (df_test['machine_pred'] + df_test['sentiment_by_dic'])*df_test['usefulCount']
```

```
df_test = df_test.groupby(['condition','drugName']).agg({'total_pred' : ['mean']})
df_test
```

Fig 4.5 normalizing useful count and overall dftest to find the recommended medicines

		total_pred
		mean
condition	drugName	
ADHD	Adderall	0.051084
	Adderall XR	0.032110
	Adzenys XR-ODT	0.008153
	Amantadine	0.003407
	Amphetamine	0.011073
...	...	...
t Care	Salvax Duo	0.000000
von Willebrand's Disease	Desmopressin	4.125000
zen Shoulde	Nabumetone	8.333333
	Relafen	19.000000
	Voltaren	3.666667

5227 rows × 1 columns

Fig 4.6 Results of recommendation system

After model building and evaluation, this is the final version of our recommendation of medicinal drugs according to various conditions based on the order of value, therefore, a medicine recommendation system which recommends all the medicines/drugs with their mean predicted value. These predicted values tell us which of the medicines are more accurate and have been proven useful as well have no harmful effects, and therefore are safer to use. A medicine is better when the mean predicted value is more. so, doctors prescribe the best medicine by considering the highest mean value predicted.

Chapter 05: RESULTS AND CONCLUSION

5.1 Discussion on the Results Achieved

The project was selected as a motivation for recommending the right medicinal drug as per the condition of the patients by checking the reviews from the dataset and then proceeded the project starting with the exploration data analysis phase, followed by the data preprocessing where we need to make the data easy for further analysis and modelling. In the data exploration stage, we used the statistical techniques as well as the visualisation techniques to understand the data and its features.In the same section, we had to find the best n-grams that could represent the relationship and emotions with the features like rating or data. The next part of the project was the data preprocessing stage. Here we had to remove the missing, defective and unwanted data from the set as well as any such condition which had less than two drugs for recommendation since it won't be as reliable. In the process of modelling, to handle the limitations which were in NLP, we decided to use Lightgbm to overcome it. It is one of the fastest machine learning algorithms which is based on the decision tree classifier. Alongside, we conducted an emotional analysis or sentiment analysis using NLTK's Wordnet and SentiWordnet as well as using a word dictionary. Apart from this, we also normalised the biassed usefulcount by condition for better efficiency and reliability. All these steps allowed us to measure the total mean predicted result values for all the drugs under every condition which would help in recommending the right drug by the order of its value.

5.2 Application of the Project

Since the pandemic began, many people with problems other than covid have avoided going to hospitals for a thorough examination in order to avoid coming into contact with the virus. This has resulted in online treatments and the use of drugs found on the internet. Furthermore, many doctors and medical students have begun practising medicine and recommending drugs to patients with limited knowledge and experience, resulting in errors and mistakes in their judgement and a number of deaths.
To avoid such mistakes, we provide a medicine recommendation system which could help the doctors or people who want to treat themselves and can be used by them while prescribing or taking medicines respectively.

5.3 Limitation of the Project

In conclusion, these are the limitations we had during the project.

- Sentiment word dictionaries for sentiment analysis is not a great way since it has low reliability if the rate of categorised good and bad words are limited. Therefore, we could have provided a criteria where, if the number of sentiment words was 5 or less, we could exclude the observations to avoid biassed results.

- We normalised usefulCount to ensure the reliability of the predicted values, and multiplied it to the predicted values,but as the review gets older, the usefulCount may be higher for them since the number of visitors increases from the date when the reviews are posted. Therefore, we should have also considered time when normalising usefulCount.
- If the emotion is positive, the reliability should be increased to the positive side, and if it is negative, the reliability should be increased toward the negative side. However, we simply multiplied the usefulCount for reliability and did not consider this part. So we should have multiplied considering the sign of usefulCount according to different kinds of emotion.

5.4 Future Work

Our future work would be to work on the limitations of these projects.
The results could've been better if we'd have used deep learning to train the model. Apart from that, we need to consider time when normalising the usefulcount since the count increases if the review was older.

References

Garg, Satvik. (2021).” Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning”.

L. Wang, Q. Zhang, Q. Qian, J. Wang, W. Cheng and J. Feng, "An Internet Medical Service Recommendation Method based on Collaborative Filtering," 2020 International Conference on Service Science (ICSS), 2020, pp. 31-35, doi: 10.1109/ICSS50103.2020.00013.

T. Venkat Narayana Rao, Anjum Unnisa, Kotha Sreni,”Medicine Recommendation System Based On Patient Reviews” ,International Journal of Scientific & Technology Research Volume 9, Issue 02,February 2020

Varun A.Goyal , Dilip J. Parmar , Namaskar I. Joshi , Prof. Komal Champanerkar, ”Medicine Recommendation System”,International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 03 | Mar 2020

Benjamin Stark , Constanze Knahl , Mert Aydin , Karim Elish “A Literature Review on Medicine Recommender Systems”,(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 8, 2019

Y. Bao and X. Jiang, "An intelligent medicine recommender system framework," 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), 2016, pp. 1383-1388, doi: 10.1109/ICIEA.2016.7603801.

Na, Jin-Cheon & Kyaing, Wai. (2015). “Sentiment Analysis of User-Generated Content on Drug Review Websites”. Journal of Information Science Theory and Practice. 3. 6-23. 10.1633/JISTaP.2015.3.1.1.