|  |  |
| --- | --- |
| Internship Project Title | Classification Model - Build a Model that Classifies the Side Effects of a Drug |
| Name of the Company | TCS-ion |
| Name of the Industry Mentor | DEBASHIS ROY |
| Name of the Institute | Tatyasaheb Kore Institute of Engineering and Technology. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| SEP-5-2023 | 0CT-6-2023 | 125 | COLAB  VISUAL STUDIO CODE | PYTHON SKLEARN PLOTLY MATPLOTLIB  NTLK |

|  |  |
| --- | --- |
| **TABLE OF CONTENT** | |
| **Content** | **Page number** |
| 1. Acknowledgement | 3 |
| 1. Objective | 4 |
| 1. Introduction | 5 |
| 1. Description | 6 |
| 1. Internship Activities | 7 |
| 1. Approach\methodology | 11 |
| 1. Assumptions | 11 |
| 1. Exceptions\exclusion | 11 |
| 1. Charts, Table, Diagrams | 12 |
| 1. Algorithms | 20 |
| 1. Challenges & Opportunities | 21 |
| 1. Risk Vs Reward | 22 |
| 1. Reflections on the Internship | 23 |
| 1. Recommendations | 23 |
| 1. Outcome / Conclusion | 24 |
| 1. Enhancement Scope | 24 |
| 1. LINK TO CODE AND EXECUTABLE FILE | 24 |
|  |  |

1. **ACKNOWLEDGEMENT**

First I would like to thank TCSION for giving me the opportunity to do an internship within the organization.

I also would like all the people that worked along with me to support with their patience and openness they created an enjoyable working environment.

It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of these individuals.

I am highly indebted TCSion authorities for the facilities provided to accomplish this internship.

I would like to thank my mentor Mr. Debashius Roy for his constructive criticism throughout my internship.

I would like to thank ICT internship for their support and advices to get and complete internship in above said organization.

I am extremely great full to my colleagues and friends who helped me in successful completion of this internship.

1. **OBJECTIVE**

* To create a Classification Model.
* Build a Model that Classifies the Side Effects of a Drug.

1. **INTRODUCTION**

Machine learning (ML) is a state-of-the-art approach that has extensive applications in categorization, prediction, and forecasting. Machine learning techniques are being used in a variety of fields such as medicine, engineering, education, manufacturing and production, weather forecasting, traffic management, robotics, and more. It is one of the most advanced concepts of artificial intelligence (AI), and provides a strategic approach to developing automated, intricate and unbiased algorithmic techniques for multimodal and dimensional biomedical or mathematical data analysis. Machine learning has already shown potential in pharmaceuticals and medicine for finding ways to effectively collect and use various types of data for better analysis, prevention, and treatment of individuals.

Healthcare is an important industry that offers value-based care to millions of people. Healthcare specialists and stakeholders around the world are looking for innovative ways to deliver on quality, value and outcome. Machine learning (ML) based applications embedded with real-time patient data available from different healthcare systems in multiple countries can increase the efficacy of new treatment options which were previously unavailable. It has found wide applications in precision medicine and personalized treatments. Using ML techniques, side effects of drugs both beneficial and adverse can be classified into categories. This can help make more intelligent decisions for precision medicine, personalized treatments, and drug repurposing. Drug classifiers based on side effects can also be an informational resource designed to assist licensed healthcare practitioners in caring for their patients and/or to serve consumers viewing this service as a supplement to, and not a substitute for, the expertise, skill, knowledge and judgment of healthcare practitioners.

A side effect is an unwanted secondary effect that occurs in addition to the desired therapeutic effect of a drug or medication. Side effects can vary for each individual depending on their disease state, age, weight, gender, ethnicity, and general health. They can occur when commencing, decreasing/increasing dosages, or ending a drug or medication regimen and may lead to non-compliance with prescribed treatment. Severe side effects may require adjusting the dosage or prescribing a second medication. Lifestyle or dietary changes may also help minimize side effects. Classifying the side effects for each drug is a challenging task. Machine learning techniques can make such tasks easier without compromising accuracy. Pharmacogenetic research has uncovered significant differences among racial and ethnic groups in the metabolism, clinical effectiveness, and side-effect profiles of many clinically important drugs. These differences must be taken into account in the design of cost management policies such as formulary implementation, therapeutic substitution, and step-care protocols. These programs should be broad and flexible enough to enable rational choices and individualized treatment for all patients, regardless of race or ethnic origin.

1. **DESCRIPTION OF INTERNSHIP**

Machine learning models have been developed to classify side effects of drugs based on age and gender using a dataset with user-generated text acquired by scraping the WebMD site. The dataset includes both demographic and clinical data and provides user reviews on specific drugs along with related conditions, side effects, age, sex, and ratings reflecting overall patient satisfaction. WebMD is an organization that provides information, support, and reference material about health subjects through a team of doctors and health experts across a broad range of specialty areas. The dataset contains 362806 instances and 12 features including categorical, numerical, and text data. Dataset provides user reviews on specific drugs along with related conditions, side effects, age, sex, and effectiveness reflecting overall patient satisfaction. The structure of the data is that a patient with a unique ID purchases a drug that meets his condition and writes a review and rating for the drug he/she purchased on the date. Afterwards, if the others read that review and find it helpful, they will click UsefulCount.

1. **INTERNSHIP ACTIVITIES**

**Activity 1: Familiarize the topic and search for a dataset.**

* Familiarize the TCSION platform and understands what I have to do and what kind of dataset I want. For that I search on different sites like Kaggle. And finalize a WebMD dataset of 368206 entries.

**Activity 2: Finalize a dataset from Kaggle and do basic analysis.**

* Do the basic analysis like info, describe, columns, shape, null values, datatypes, unique count etc. Then I familiarize what each column in the dataset about.

**Activity 3: Data preprocessing**

* Missing value handling, Data Wrangling, Cleaning irrelevant data, data with inconsistent datatypes etc.

**Activity4: Exploratory data analysis non graphical**

* Count of gender, Drugs, DrugId, most mentioned drug and DrugId, most mentioned condition etc.

**Activity 5: Exploratory data analysis graphical**

* Plotting of graph based on each parameter. Count plot of top 20 drugs used, ease of use satisfaction rate etc.

**Activity 6: Text processing**

* Text cleaning
* Make text lowercase, remove text in square brackets, remove punctuation and remove words containing numbers.

**Activity 7: Text processing**

* Import the stopwords module from the nltk library and creates a set of English stop words.
* Applies a lambda function to each element in the Review column.
* The lambda function splits each review into individual words using the split () method.
* The filtered words back together into a single string, separated by spaces.
* The filtered text is stored in a new column.

**Activity 8: Lemmatization**

* Downloads the WordNet corpus, which is a lexical database for the English language used by NLTK for lemmatization.
* Imports the WordNetLemmatizer class from the nltk. stem module.
* Creates an instance of the WordNetLemmatizer class.
* Splits each review into individual words using the split () method.
* The lemmatized words back together into a single string, separated by spaces.
* Finally, the lemmatized text is stored in a new column.

**Activity 9: Text blob sentimental analysis**

* Lambda function to each element in the review\_clean column.
* The lambda function applies the Text Blob sentiment analysis to each review.
* The Text Blob sentiment analysis returns a **named tuple** of the form Sentiment (polarity, subjectivity).

**Activity 10:** **Sentimental analysis**

* SentimentIntensityAnalyzer (). Polarity\_scores () method to calculate sentiment scores

**Activity 11: Word cloud visual representation**

* Word cloud visual representation of positive negative and neutral reviews.
* The word cloud visualization for neutral sentiment.
* It uses the Word Cloud class from the word cloud library to create a word cloud image based on the frequencies of words.

**Activity 12: N gram plotting**

* + - N gram plotting of top 10 positive and negative reviews.
* The CountVectorizer class from the sklearn. feature\_extraction.text module to extract the top n-gram words from a given corpus.
* The function takes three parameters: corpus, ngram\_range, and n. It returns a list of the most frequently occurring n-gram words in the corpus, sorted in descending order of frequency.

**Activity 13:** **Train dataset.**

* For that I use the TfidfVectorizer class from the sklearn. feature\_extraction.text module to create a term frequency-inverse document frequency (TF-IDF) matrix for the given corpus of text data.
* The TfidfVectorizer class is used to convert a collection of raw documents into a matrix of TF-IDF features.

**Activity 14: Modelling**

* LGBM CLASSIFIER AND XGBM CLASSIFIER

**Activity 15:** **Modelling continues**

* XGBOOST classifier got error. Try to resolve the error.

**Activity 16: XGBOOST Classifier**

* Correcting error. And got accuracy of 70.

**Activity 17: Naïve Bayes Bernoulli Algorithm**

* Accuracy is 0.6633333333333333Precision is 0.6633333333333333Recall is 0.6633333333333333F1 score is 0.6633333333333333.
* So the accuracy is less than LGBM Classifier.

**Activity 18: Random forest**

* The Accuracy of the train\_data is: 0.9982244897959184
* The Accuracy of the model is: 0.8639523809523809
* Accuracy is less than LGBM but greater than NB

**Activity 19: Logistic regression**

* The Accuracy of the train\_data is: 0.9067755102040816
* The Accuracy of the model4 is: 0.8937619047619048
* Accuracy is good and almost same as LGBM

**Activity 20: DECISION TREE**

* The Accuracy of the train\_data is: 0.6969795918367347
* The Accuracy of the model is: 0.6804285714285714
* Accuracy is very low. Not acceptable.

**Activity 21: GradientBoostingClassifier**

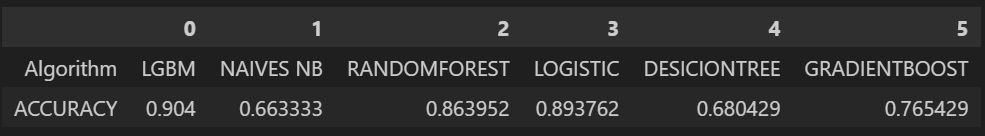
* The Accuracy of the train\_data is: 0.7796530612244897
* The Accuracy of the model is: 0.7654285714285715
* Accuracy is 0.7654285714285715
* Precision is 0.7654285714285715
* Recall is 0.7654285714285715
* F1 score is 0.7654285714285715
* Very less accuracy than other algorithms.

**Activity 22: KNN CLASSIFIER**

* Accuracy is 0.37323809523809526. The most least accuracy got by this.

**Activity 23:** COMPARING ALL ALGORITHMS

* Compared all the algorithms and realize that LGBM and Logistic regression having the most accuracy.



* I select logistic regression for the modelling.
* Accuracy Report and Confusion Matrix I concluded that Logistic Regression model is best fitted model.

**Activity 24: checking the prediction**

**Activity 25: just for a practice I try to make a web app front end page.**

**Activity 26: I prepared a backend page.**

**Activity 27: I connect front end and backend and successfully hosted in local host.**

**Activity 28: I try to host in python anywhere and other hosting app abut NLTK takes too much storage space. So I drop it. And decided to upload until modelling.**

**Activity 29.: Upload all documents in GITHUB.**

**Activity 30: Preparation of report.**

1. **APPROACH / METHODOLOGY**

The dataset includes both demographic and clinical data. It contains 362806 instances and 12 features including categorical, numerical and text data. Dataset provides user reviews on specific drugs along with related conditions, side effects, age, sex, and effectiveness reflecting overall patient satisfaction. The structure of the data is that a patient with a unique ID purchases a drug that meets his condition and writes a review and rating for the drug he/she purchased on the date. Afterwards, if the others read that review and find it helpful, they will click UsefulCount.

1. **ASSUMPTIONS**

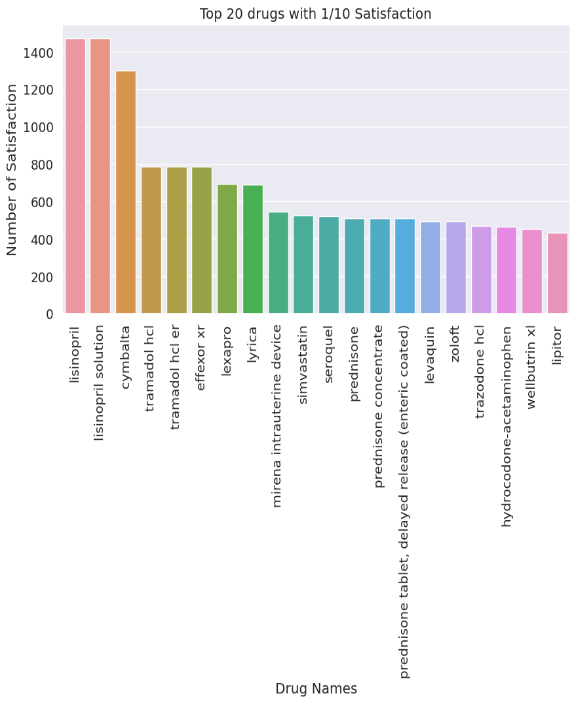
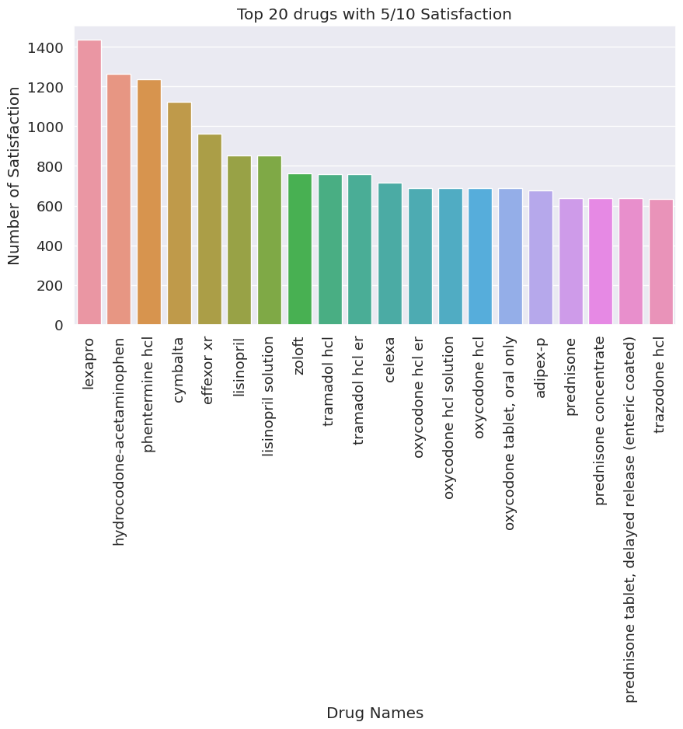
* The data collected from WebMD is representative of the population as a whole.
* The data collected is accurate and free from errors.
* The machine learning techniques used to classify side effects are appropriate for the dataset.
* The classification model developed is generalizable to other datasets.
* The age and gender information provided in the dataset is accurate and complete.

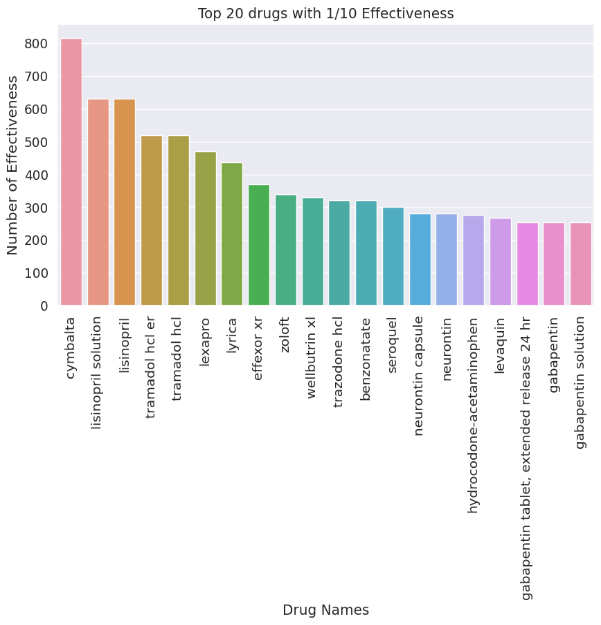
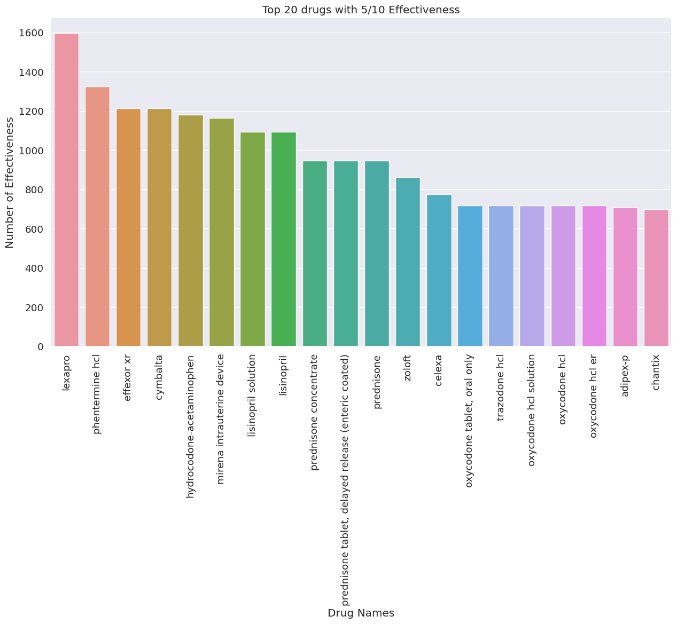
1. **EXCEPTIONS / EXCLUSIONS**

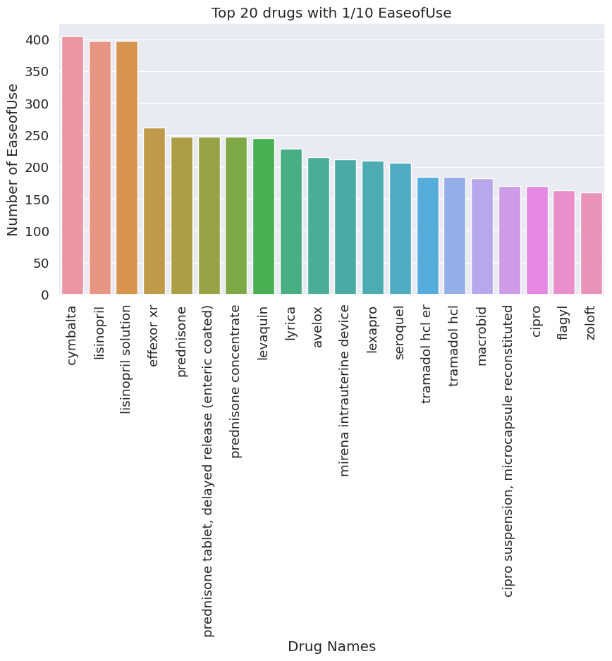
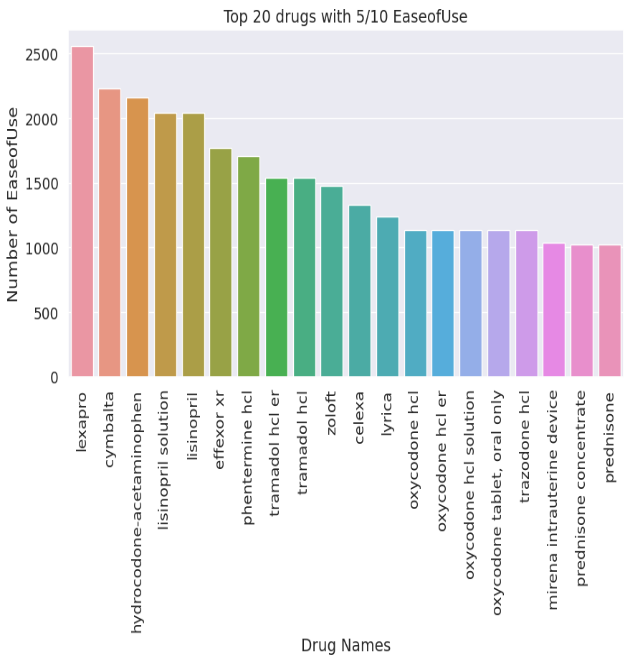
* The dataset may not include all possible side effects for each drug.
* The dataset may not include information on all drugs or medications.
* The dataset may not be up-to-date or accurate.
* The classification model developed may not be able to accurately predict side effects for all individuals.
* The classification model developed may not be able to account for all factors that can impact side effects.

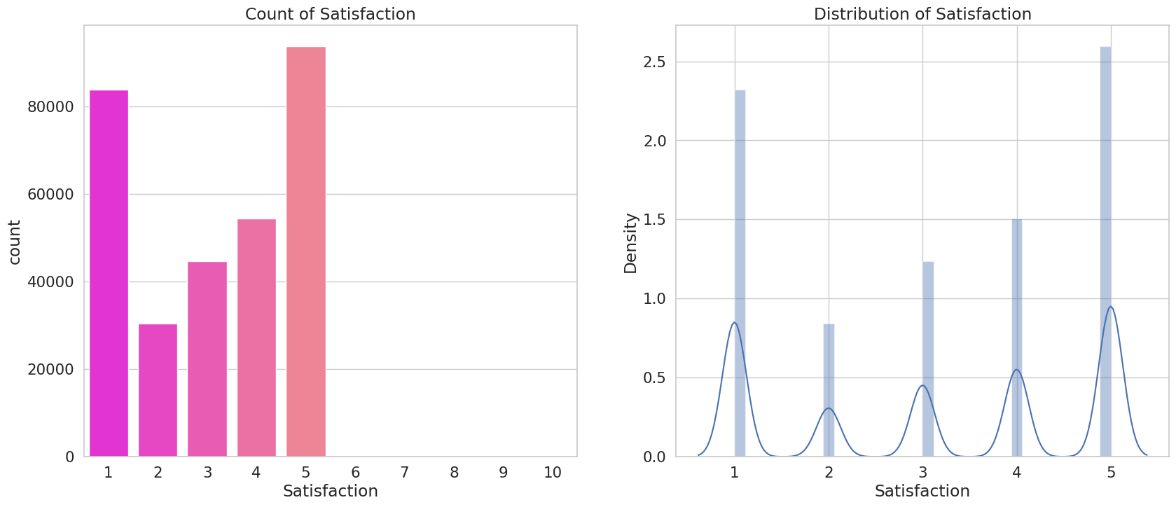
1. **CHARTS, TABLES, DIAGRAMS**

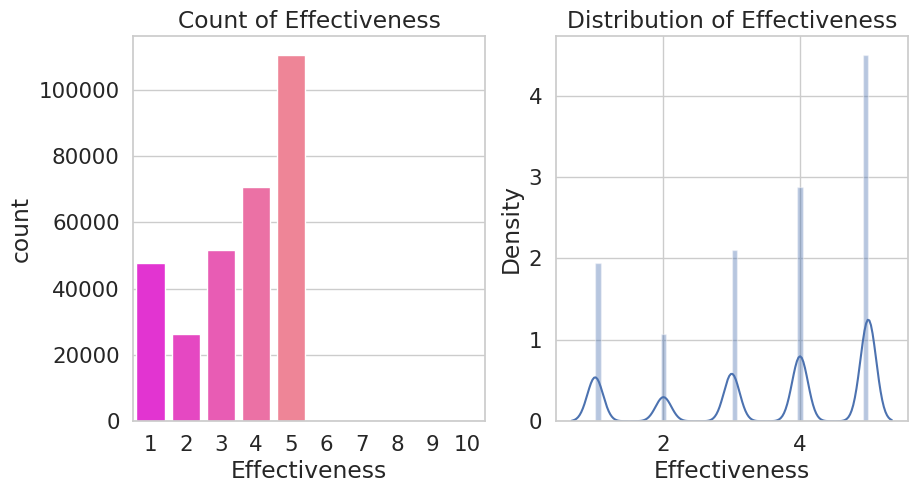
**Exploratory data analysis**

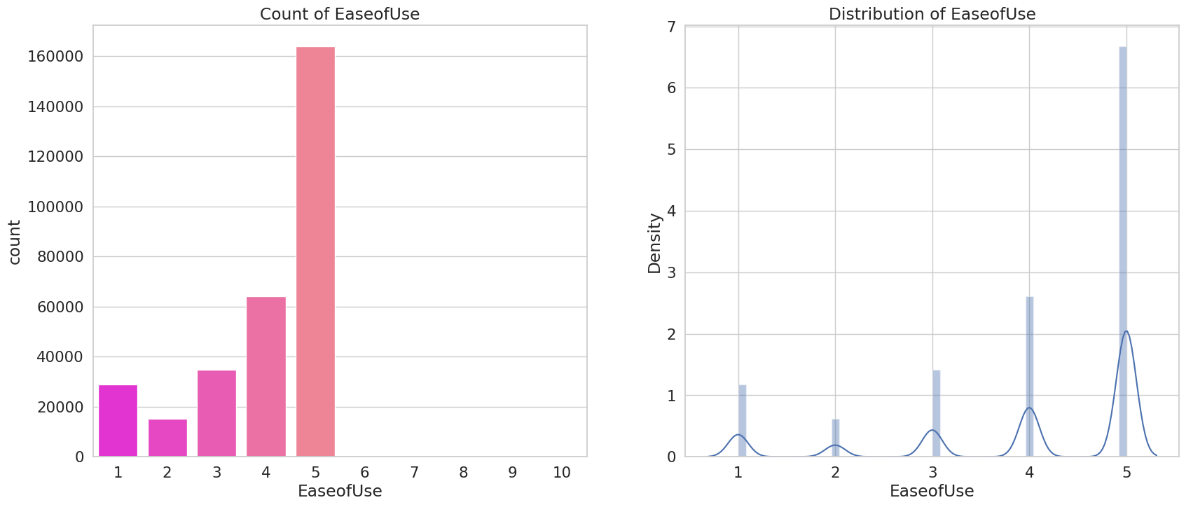
****

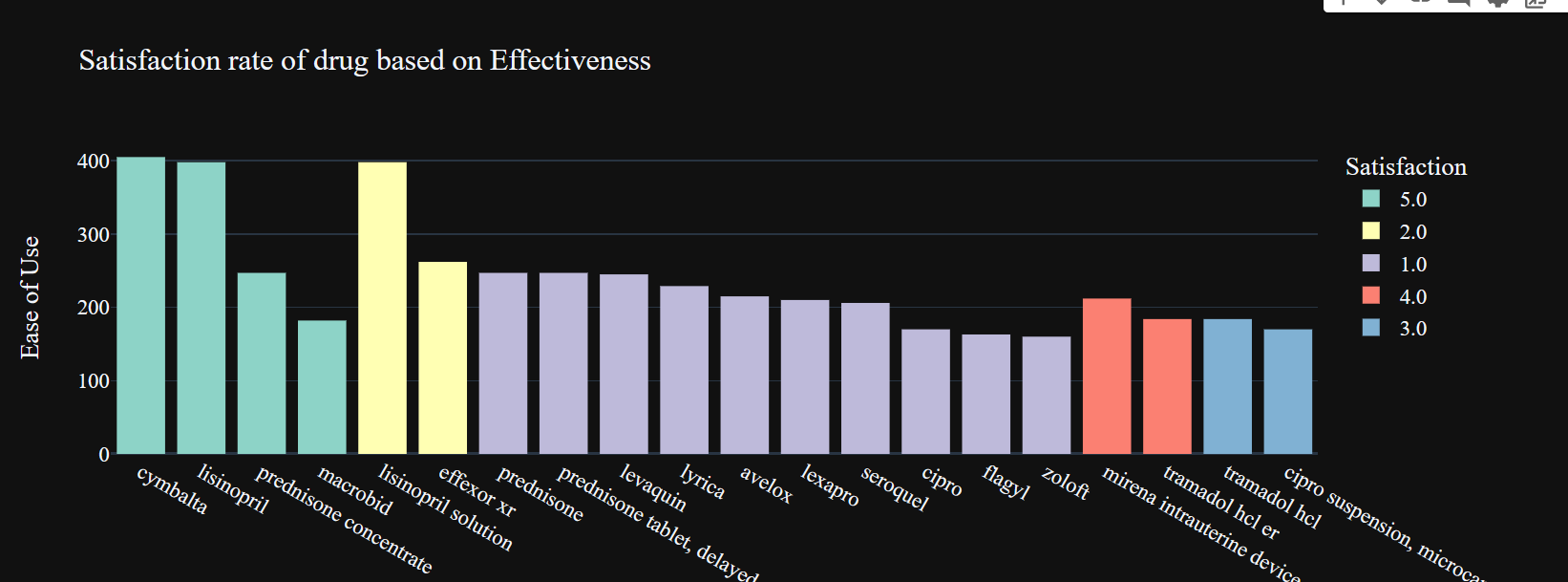
****

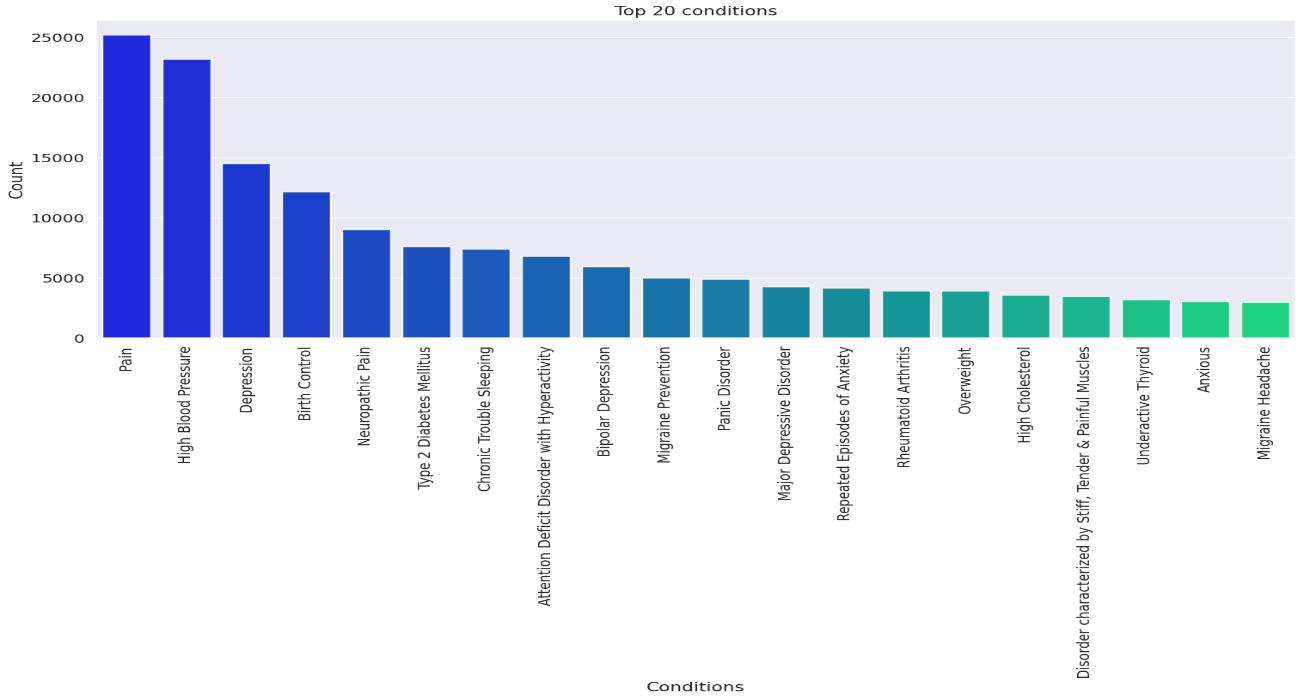
****

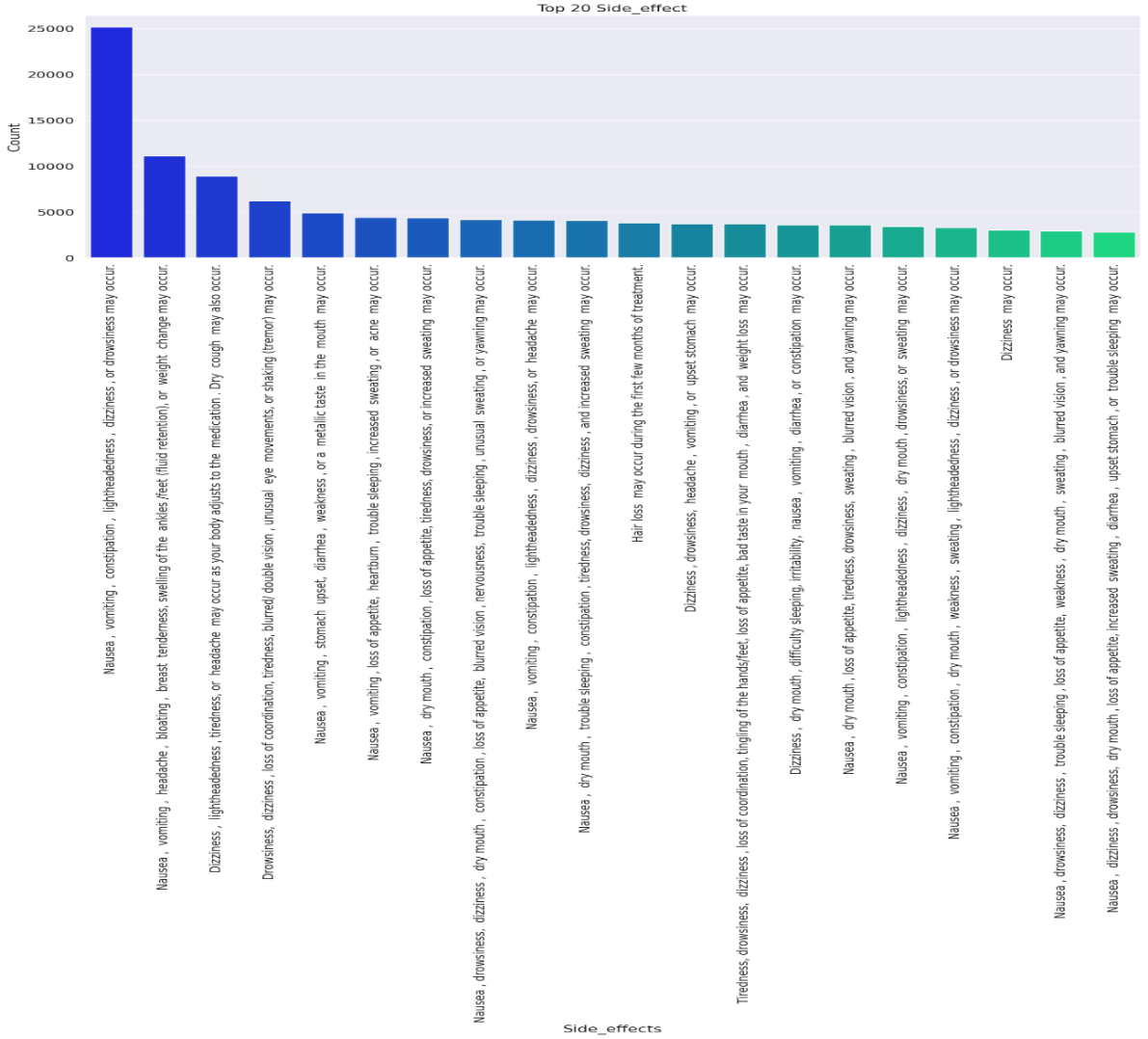
****

****

****

****

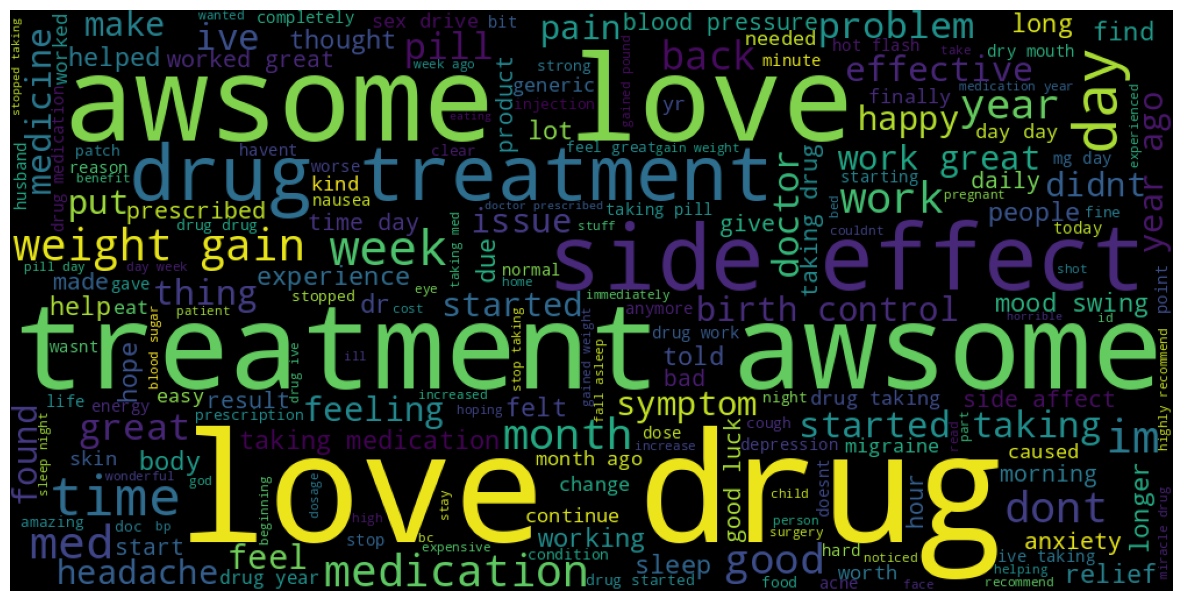
**\**

****

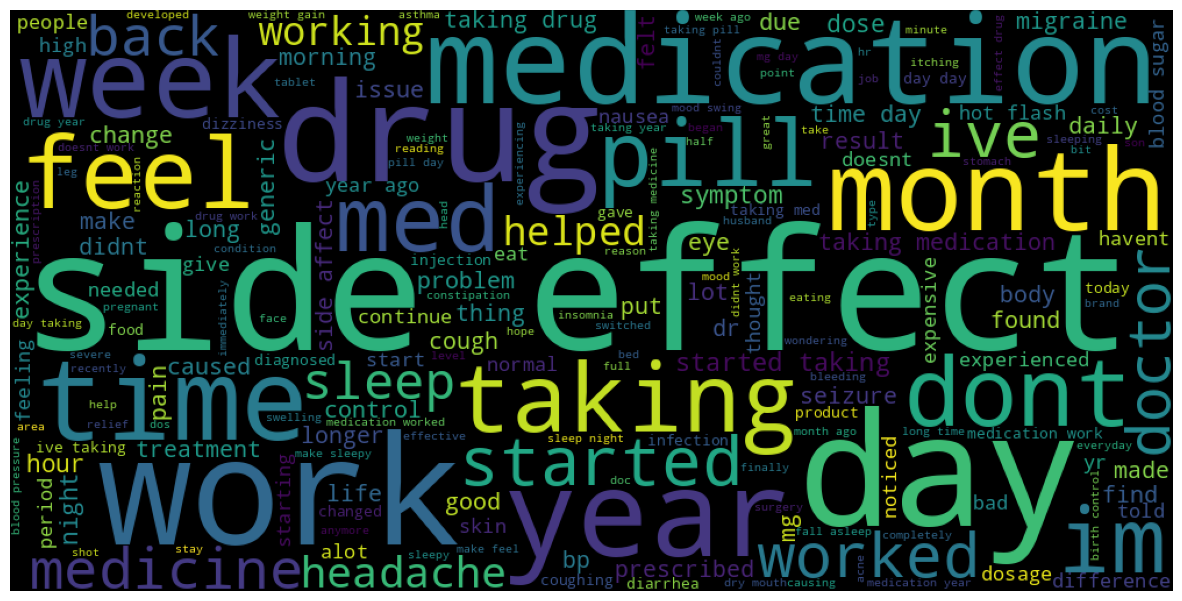
For the Lisinopril drug as example, I concluded as this drug have side effects but its effectiveness and satisfaction rate is good. Females are more affected by the side effects and females are the main users. age group of 55-64 having adverse effects. 45-64 is almost the same side effects. reviews count decreased yearly. That means users body are adjusted to side effects or however users have the mentality to cure the sickness and not too much worry about side effects.

**TEXT PROCESSING**

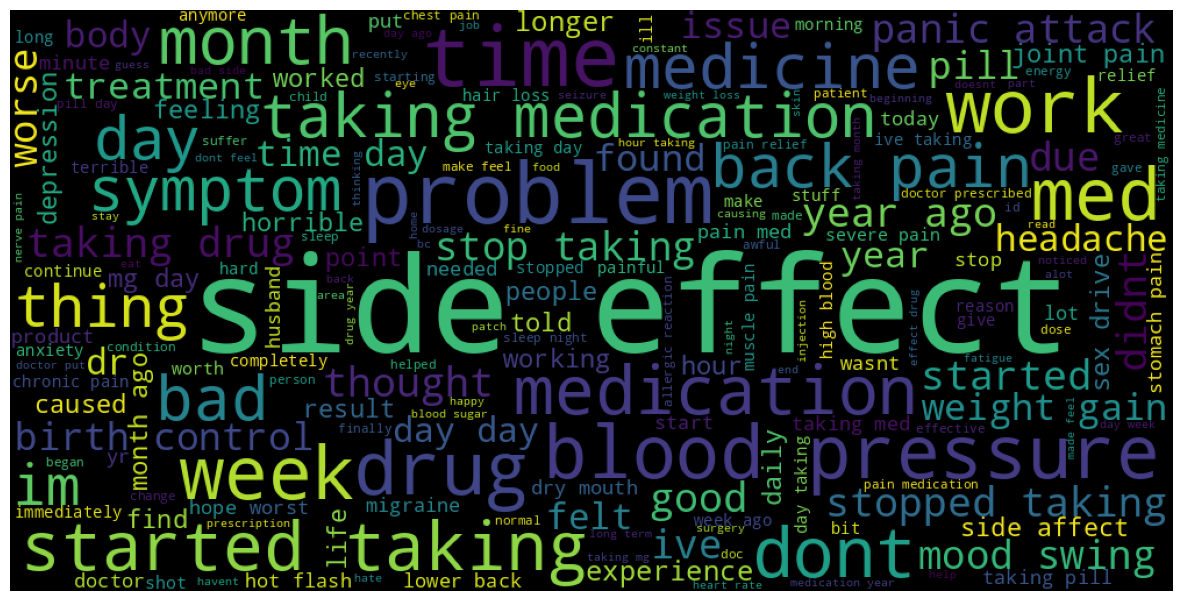
1. Word cloud of the Clean reviews with +ve sentiments



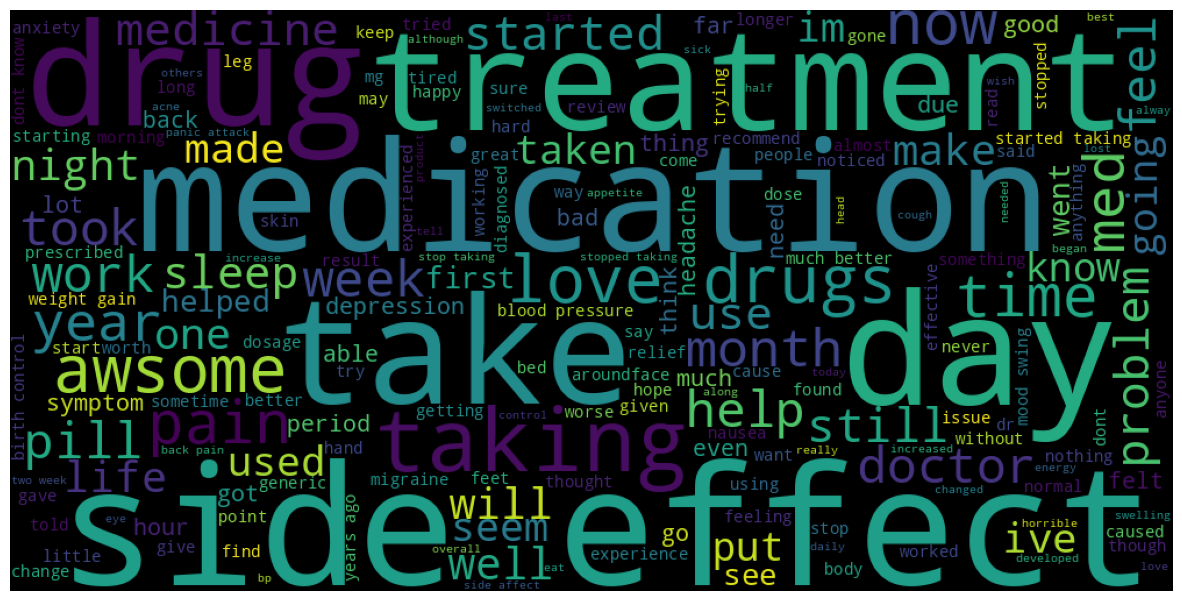
1. Word cloud of the Clean reviews with -ve sentiments



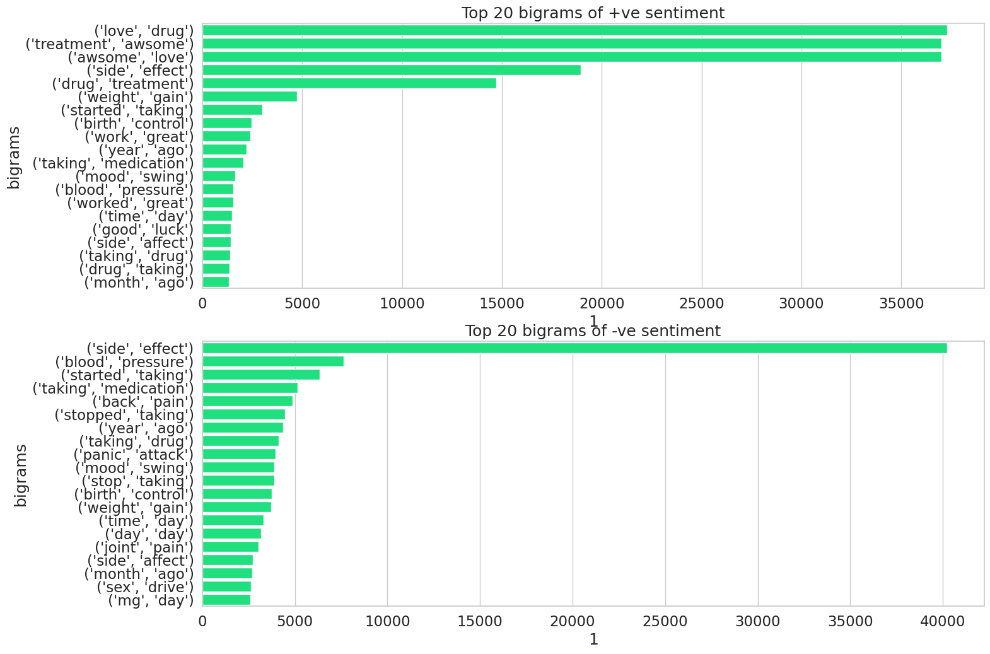
1. Word cloud of the Clean reviews with neutral sentiments

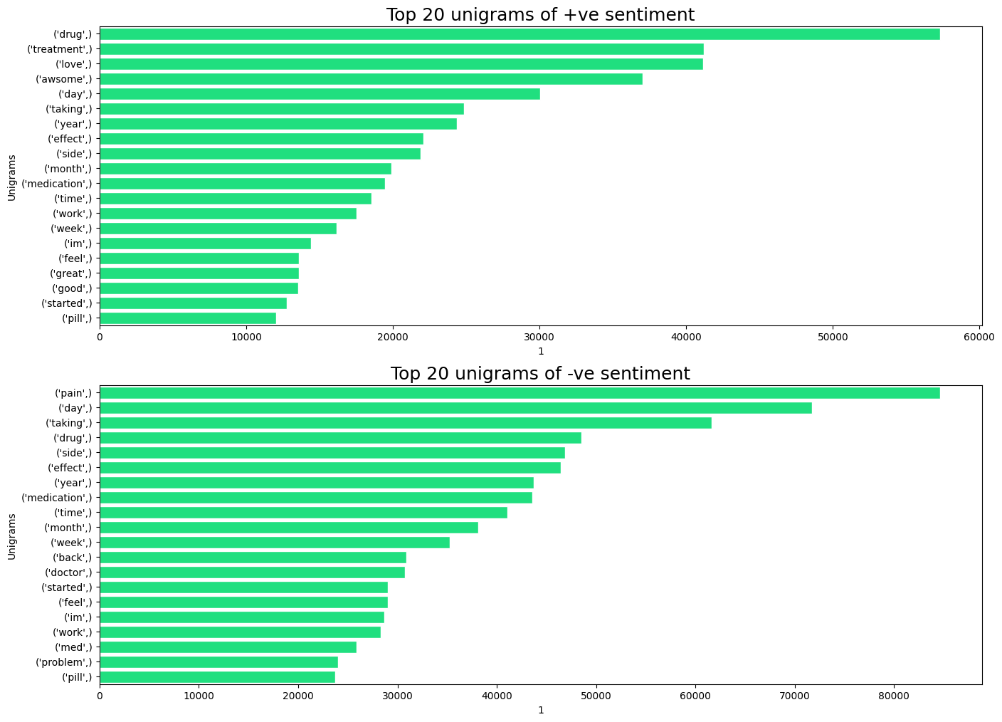


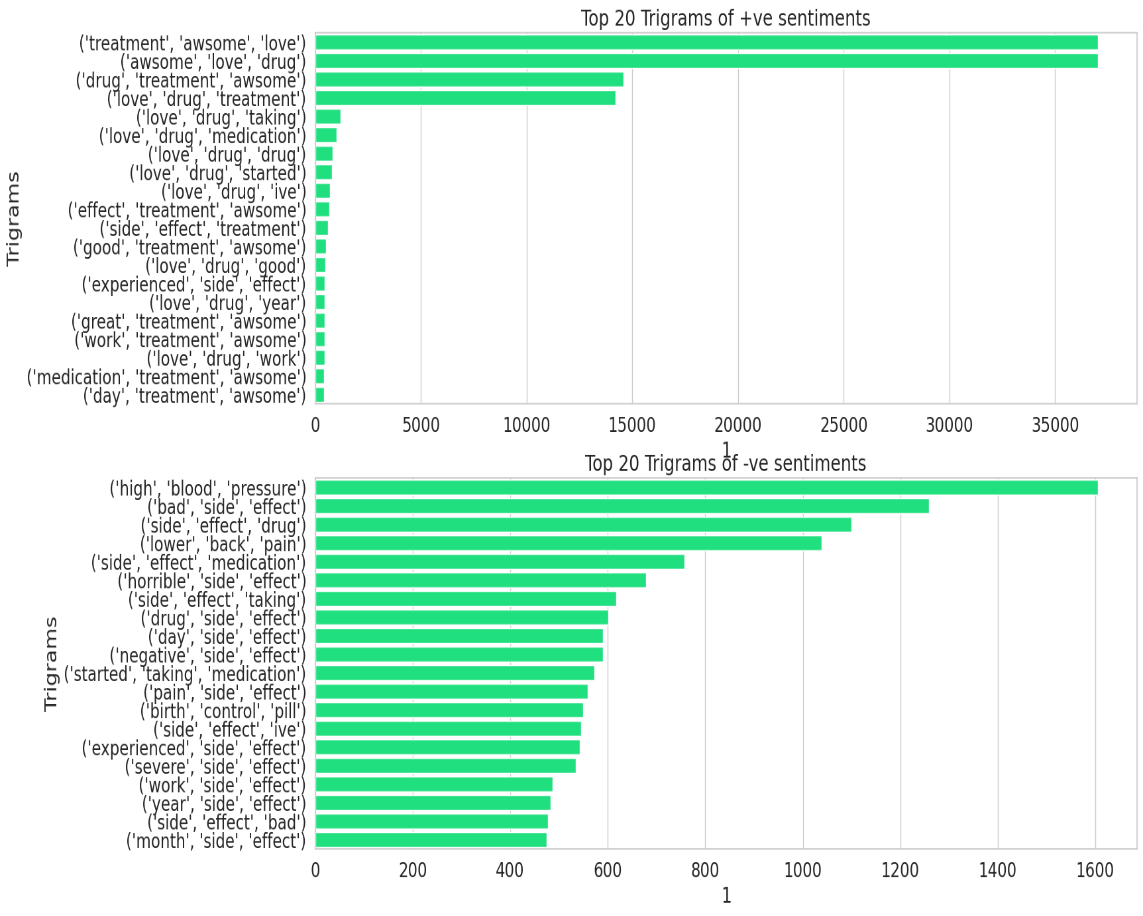
1. Word cloud of the unClean reviews with +ve sentiments



1. N grams





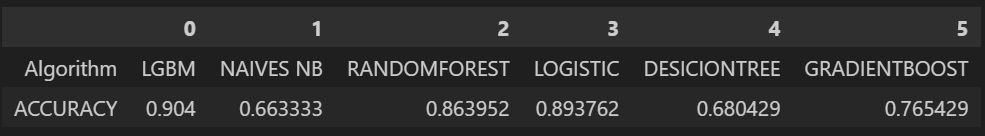


The analyzer parameter determines whether the feature should be made of word or character n-grams. In this case, it is set to 'word', which means that the TF-IDF matrix will be based on individual words. From this we can understand the most used words in reviews as per the sentiments of the users. From this we can model the columns.

1. **ALGORITHMS**

* **LightGBM (LGBM)**: A gradient boosting framework that uses tree-based learning algorithms. It is designed to be efficient and scalable in handling large datasets.
* **Naive Bayes (NaiveNB)**: A probabilistic algorithm that applies Bayes’ theorem with the assumption of independence between every pair of features. It is commonly used for text classification.
* **Random Forest**: An ensemble learning method that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
* **Logistic Regression**: A statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).
* **Decision Tree**: A decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.
* **Gradient Boosting**: An ensemble technique that combines weak learners to create a strong learner by iteratively training models to minimize the loss function. It is used for both regression and classification problems.
* **K-Nearest Neighbors (KNN)**: A non-parametric algorithm that stores all available cases and classifies new cases based on a similarity measure

**COMPARISON OF ALL ALGORITHMS**



* I select logistic regression for the modelling.
* Accuracy Report and Confusion Matrix I concluded that Logistic Regression model is best fitted model.

1. **CHALLENGES & OPPORTUNITIES**

* The dataset may not be representative of the entire population.
* The data collected may not be accurate or complete.
* The classification model developed may not be generalizable to other datasets.
* The age and gender information provided in the dataset may not be accurate or complete.
* Developing a classification model that can accurately predict side effects for all individuals is a challenging task.
* The dataset may not include all possible side effects for each drug.
* The dataset may not include information on all drugs or medications.
* The dataset may not be up-to-date or accurate.
* Developing a classification model that can accurately predict side effects for all individuals is a challenging task.
* There is a need to account for all factors that can impact side effects.
* However, developing such a model can help healthcare professionals make more informed decisions about prescribing medications and help patients make more informed decisions about their health.
* Another opportunity is that the model can be used to identify previously unknown side effects of drugs, which can lead to new discoveries and better treatment options. Additionally, the model can be used to identify patterns in side effects across different drugs and patient populations, which can help researchers better understand the underlying mechanisms of drug side effects. Finally, the model can be used to develop personalized treatment plans based on an individual's age, gender, and other factors that may impact their risk of experiencing side effects.

1. **RISK VS REWARD**

|  |  |
| --- | --- |
| 1. **Risk** | **Reward** |
| As a reward, Healthcare professionals make more informed decisions about prescribing medications and help patients make more informed decisions about their health. It can also be used to identify previously unknown side effects of drugs, which can lead to new discoveries and better treatment options. Additionally, the model can be used to identify patterns in side effects across different drugs and patient populations, which can help researchers better understand the underlying mechanisms of drug side effects. | As a risk, developing a classification model that can accurately predict side effects for all individuals is a challenging task. There is a need to account for all factors that can impact side effects, and the dataset may not include all possible side effects for each drug. The dataset may also not include information on all drugs or medications, and it may not be up-to-date or accurate. It is important to carefully consider the risks and rewards of any treatment or intervention before making a decision. Healthcare professionals should work closely with their patients to ensure that they understand the potential benefits and harms of different treatments and can make informed decisions about their care. |

1. **REFLECTIONS ON THE INTERNSHIP**

Developing a classification model that can accurately predict side effects of drugs based on age and gender. The project also highlighted the importance of data preprocessing and cleaning to ensure that the data is accurate and complete. Additionally, the project demonstrated the potential of machine learning techniques to identify previously unknown side effects of drugs and patterns in side effects across different drugs and patient populations. Future research could focus on developing more accurate classification models that can account for all factors that can impact side effects and can be generalized to other datasets.

1. **RECOMMENDATIONS**

* Data Collection: Collect data from multiple sources to increase the diversity of the dataset and improve the generalizability of the model.
* Feature Engineering: Explore different feature engineering techniques to extract meaningful information from the dataset. Consider incorporating additional features such as drug dosage, treatment duration, and patient medical history to improve the model’s performance.
* Model Selection: Experiment with different machine learning algorithms and evaluate their performance using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. Consider ensemble methods or deep learning models to further enhance the model’s predictive capabilities.
* Hyper parameter Tuning: Optimize the hyper parameters of the selected model using techniques such as grid search or random search. This can help improve the model’s generalization ability and prevent overfitting.
* Validation and Testing: Use appropriate validation techniques such as cross-validation to estimate the model’s performance on unseen data. Perform rigorous testing on an independent test set to assess the model’s real-world performance.
* Interpretability: Use interpretable machine learning models such as decision trees or logistic regression to provide insights into the factors influencing side effects.
* Continual Improvement: Regularly update the model with new data to ensure its relevance and accuracy over time. Monitor its performance in real-world scenarios and refine it based on feedback from healthcare professionals and end-users.

1. **OUTCOME / CONCLUSION**

In conclusion, from the EDA analysis, I concluded as this drug have side effects but its effectiveness and satisfaction rate is good. Females are more affected by the side effects and females are the main users. age group of 55-64 having adverse effects. 45-64 is almost the same side effects. Reviews counts decreased yearly. That means users body are adjusted to side effects or however users have the mentality to cure the sickness and not too much worry about side effects.

Developing a classification model for side effects of drugs using the WebMD dataset is a challenging task that requires careful consideration of the data, machine learning techniques, and evaluation metrics. While there are limitations and challenges associated with this approach, such as the need to account for all factors that can impact side effects and the accuracy of the dataset, there are also significant opportunities for improving patient care and advancing medical research. By developing more accurate classification models, researchers can help healthcare professionals make more informed decisions about prescribing medications and help patients make more informed decisions about their health. Additionally, these models can be used to identify previously unknown side effects of drugs and patterns in side effects across different drugs and patient populations, which can lead to new discoveries and better treatment options. Overall, the development of classification models for side effects of drugs using the WebMD dataset has the potential to significantly improve patient outcomes and advance medical research.

1. **ENHANCEMENT SCOPE**

Regularly update the model with new data to ensure its relevance and accuracy over time. Monitor its performance in real-world scenarios and refine it based on feedback from healthcare professionals and end-users.