An Optimized Machine Learning Approach for Improved Sentiment Detection and Enhanced Recommendation Systems Using Drug Reviews

BY

RASEL SARKER

ID: 221-15-5663

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Sazzadur Ahamed
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

13 JANUARY 2025

APPROVAL

This Thesis paper titled **"An Optimized Machine Learning Approach for Improved Sentiment Detection and Enhanced Recommendation Systems Using Drug Reviews"**, submitted by Md. Rasel Sarker, ID No: 221-15-5663 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13 January 2025.

BOARD OF EXAMINERS

Dr. Md. Zahid Hasan — **Chairman**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mohammad Monirul Islam — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Afjal Hossan Sarower — **Internal Examiner**
Sr. Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Ahmed Wasif Reza — **External Examiner**
Professor
Department of Computer Science and Engineering
East West Universit

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Sazzadur Ahamed, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Md. Sazzadur Ahamed

Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

13.01.25

MD. Rasel Sarker
ID:221-15-5663
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing making us possible to complete the final year project/internship successfully.

We are really grateful and wish our profound indebtedness to **Md. Sazzadur Ahamed**, **Assistant Professor**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Deep Learning" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Sheak Rashed Haider Noori**, **Professor & Head**, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE Department of Daffodil International University.

We would like to thank our entire course mates in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Sentiment analysis as a branch of Natural Language Processing (NLP) is becoming more useful in healthcare by helping understand patient feedback about medicines. This study aims to improve how we evaluate drug effectiveness by combining advanced NLP techniques and machine learning methods. We also propose creating a Drug Recommendation System to support healthcare professionals in choosing the right medicines. Our study takes this further by introducing five sentiment levels: Frustrated, Bad, Neutral, Good, and Excited, based on patient ratings. We utilized a dataset obtained from the UCI Machine Learning Repository for this research and collected additional data to balance the dataset. The text data is cleaned and prepared using NLP techniques such as breaking text tokenization involves dividing or cutting the text into small pieces and eliminating punctuation and unnecessary words, and converting words to the root or base forms (stemming and lemmatization). For understanding text better, we used methods like Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Word2Vec, and manual feature creation. To handle uneven data and improve results collected additional data and also we used SMOTE-SO-MAK, a technique that creates extra samples for less common sentiment categories. Among the different machine learning models tested, Logistic Regression (LR) gave the best results. We checked the system's accuracy and performance using measures like precision, recall, F1-score, and overall accuracy. This study improves drug recommendation systems by integrating the latest NLP, machine learning algorithms and data balancing, and testing methods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLES **PAGE**

CHAPTER 1
Introduction

1.1 Introduction

Sentiment analysis involves categorizing opinions expressed intext to assess an individual's attitude towards a particular area. Significant research has been focused on areas like electronic products, movies, restaurants, etc. but in public health and medical fields, different sentiment analysis is less studied. To understand aspects of how drug reviews may be used to reveal previously undocumented side effects and how it may be used to turn health care professionals' attention to what may be a dangerous side effect associated with a particular medication. Earlier literature review has revealed that sentiments are usually categorized into positive, negative and neutral [1]. Our study introduces a more detailed classification system with five sentiment levels: Frustrated, Bad, Neutral, Good, and Excited. These categories are derived from user ratings and help identify potential adverse reactions or side effects. The dataset used in this study, sourced from the UCI Machine Learning Repository, organizes ratings into Excited (7-10), Good (6), Neutral (5), Bad (4), and Frustrated (1-3). The dataset includes 161,297 reviews distributed as follows: Excited with 106,866 entries, Frustrated with 35,063 entries, Neutral with 8,013 entries, Good with 6,343 entries, and Bad with 5,012 entries. For handling class imbalance in the given dataset, we collected additional data and used SMOTE technique with a change in density of minority classes to be 70% of the majority class. This approach generates synthetic samples, improving the model's ability to detect patterns in all classes and enhancing predictive accuracy. After applying SMOTE, our models achieved a higher accuracy of 86.86%. Preprocessing of text and feature extraction are important in enhancing the

performance of any machine learning models. In this study, preprocessing steps included converting text to lowercase, tokenization, removing punctuation, eliminating stopwords, stemming, and lemmatization. These steps ensured clean and consistent data for analysis. We used TF-IDF and Count-Vectorizer with various n-gram ranges, such as unigrams, bigrams, trigrams, and higher N-grams to capture contextual patterns in the text. We employed eleven machine learning models to predict patient sentiments based on drug reviews. These models include Logistic Regression, Support Vector Classifier, Decision Trees, Extra Trees Classifier, Random Forest, Naive Bayes, K-Nearest Neighbors, AdaBoost, Bagging Classifier, Gradient Boosting Decision Tree, and XGBoost. Each model was evaluated using metrics like accuracy, precision, recall, and F1-score. The best-performing model was validated employing 10-fold cross-validation for improved reliability of the results. In addition to sentiment analysis, we developed a recommendation system to suggest medicines based on user input. By converting textual data into numerical features using Count-Vectorizer, the system identifies the most relevant medicines for user concerns, offering a practical tool to support better healthcare decisions. This study aims to refine the role of sentiment analysis applied to the drug reviews through advanced machine learning techniques and provide a reliable recommendation system for effective healthcare support.

Two bullet point:

- To refine sentiment analysis in drug reviews by employing a multi-class text classification system and advanced machine learning models to better capture nuanced user sentiments and enhance the recommendation process.
- Enhanced the accuracy of previous research by selecting optimal feature extraction methods and effectively applying machine learning algorithms and developing a recommender system to assist most relevant medicines based on a user's specific input problem.

1.2 Motivation

This research is based on the emerging awareness of analyzing sentiment data growing in the healthcare sector, specifically regarding patients' views on medications. Existing research predominantly categorizes sentiments into broad classes like positive, negative, or neutral, which may not capture the nuanced experiences and concerns that patients have about drug effectiveness. This study seeks to advance the field by introducing a multi-class sentiment classification system with five levels—Frustrated, Bad, Neutral, Good, and Excited—allowing for a more detailed and accurate analysis of patient sentiments. By employing advanced NLP techniques and machine learning methods, this research aims to improve drug effectiveness evaluation and support healthcare professionals in making better medication choices. The Drug Recommendation System developed in this study offers personalized recommendations based on detailed sentiment analysis, overcoming the constraints of the conventional system that mostly focuses on the names of medicines.

1.3 Rationale of the Study

Data-Driven Drug Risk Assessment**:** By applying sentiment analysis to patient reviews, the study aims to identify potential drug risks from real-world data rather than relying solely on clinical data, which may miss rare or delayed adverse effects. Enhanced Predictive Models**:** The study utilizes advanced machine learning models to predict ADRs based on patient feedback, aiming for greater accuracy. This involves using text-based reviews to capture nuanced patient experiences.

Comprehensive Feature Engineering**:** By employing various methods of feature extraction – for instance, TF-IDF, n-grams, and advanced preprocessing steps—the study enhances model performance, allowing for a more precise understanding of

sentiment toward drug effectiveness and side effects. User-Friendly Health Decision Support**:** The recommender system developed in this study is designed to provide clear insights for users, helping them make informed decisions about medication based on aggregated patient experiences and the likelihood of ADRs. Potential for Broader Applications: Beyond ADR detection, this approach could be adapted for other healthcare applications, such as monitoring patient satisfaction, identifying mental health concerns from textual data, or gauging responses to treatment plans based on patient sentiment.

1.4 Research Questions

Why is a recommendation system needed in healthcare, particularly for drug effectiveness evaluation?

Why should sentiment analysis be applied to drug review data for better healthcare outcomes?

Why is consulting a recommendation system crucial for understanding potential side effects before purchasing medicine?

Can AI models fully understand the complexities of human emotions expressed in text reviews of medications?

Is the predicted result useful for that??

1.5 Expected Output

The expected contribution to this study is the designing of a next generation Drug Recommendation System that employs both sentiment analysis coupled with

machine learning in improving the detection and, perhaps as well, anticipation of ADRs. By moving beyond broad sentiment categories, the system aims to offer a more granular understanding of patient experiences with medications, incorporating five sentiment levels: Frustrated, Bad, Neutral, Good, and Excited. The study also addresses class imbalance through techniques like SMOTE-SO-MAK, improving model performance and accuracy. It is also expected to design advanced feature engineering methods that will be based on the patient's feedback to provide medication recommendations. The intent is to provide a tool that is easy to use, functional and that ultimately enhances patients' results while negating the potential adverse consequences. **Contribution to Public Health**: Support for public health initiatives by identifying potential risks associated with medications, thereby enhancing drug safety and overall healthcare quality. So, being consistent and following the structure, there will be a positive outcome in the long run, day by day. **Development of a Drug Recommendation System**: Creation of a system that assists healthcare professionals and patients in making informed medication choices based on sentiment analysis. **Enhanced Sentiment Analysis**: Improved accuracy in classifying drug reviews, leading to better identification of adverse drug reactions and insights into patient experiences. **Research Contributions to Public Health Literature**: Documentation of findings and methodologies that contribute to the academic understanding of sentiment analysis applications in the medical field, supporting future research and public health initiatives.

1.6 Report Layout

Chapter 1: The research topic's historical and contextual information is presented in the introduction, along with the investigation's challenge or query and the study's goals and relevance. This section includes the introduction of the paper in 1.1, the

inspiration for the subject of the study in 1.2, the justification for the study's conduct in 1.3, the anticipated results of this article in 1.4, and the summary or format of the document in 1.5.

Chapter 2: An initial appraisal that provides a brief synopsis of the research conducted on this topic is included in the background study. The applicable intelligence technology research is described here. Moreover, the challenges we faced while doing this study demonstrated the size of the issue.

Chapter 3: The main concepts of data set handling and model generation have been comprehensively covered in this section. In 3.1, the research approach is introduced; in 3.2, how the dataset is assembled; in 3.3, how the dataset is sterilized; in 3.4, the preliminary processing strategy of the dataset; and in 3.5 and 3.6, respectively, the recommended approach and the implementational prerequisites.

Chapter 4: This section assessed and looked into the output of our predictive framework. For ease of understanding, it incorporates all the results from the graphical description. This section comprises the evaluation of the paper and the experimental findings. The introductory part in 4.1, the result investigation in 4.2, the confusion matrix and classification report resemblance of the outcomes in 4.3 and 4.4, the precision of the validation and training in 4.5, and the discussion of the results segment in 4.6.

Chapter 5: The repercussions of marine life freshness on the community, the surroundings, as well as sustainability are briefly addressed in 5.1, 5.2, 5.3, and 5.4 accordingly.

Chapter 6: In accordance with 6.1, 6.2, and 6.3, an overview of the accomplished study, a conclusion, and potential future research are shown in this section.

CHAPTER 2
Background Study

2.1Terminologies

There has been substantial progress in applying sentiment analysis (SA) with machine learning algorithms. Many research studies have explored drug-related SA using various preprocessing and feature extraction techniques. However, achieving high accuracy in drug SA remains a challenge due to inappropriate preprocessing methods, inadequate feature extraction, and unsuitable training parameters. Additionally, the lack of ground-truth datasets often limits these studies, with many relying on small datasets to train machine learning or deep learning algorithms. Many types of research generally practice binary (positive and negative) or multiclass (positive, neutral, and negative) categorizations, which are not very detailed. For this research, we employed the Drugs.com dataset obtained from the UCI Machine Learning Database, which is a widely popular provider of machine learning datasets based at the University of California, Irvine. This study advances drug recommendation systems by integrating state-of-the-art natural language processing (NLP) techniques, machine learning cross-validation, and data balancing methods to implement a robust multiclass text classification system. Our approach categorizes sentiments into five levels: Excited, Good, Neutral, Bad, and Frustrated, providing a more nuanced understanding of user feedback and supporting informed decision-making in medication management. Analysis of sentiments in drug reviews may currently provide significant information on the experiences of individuals as regards the certain drugs, supplements, or health products. Critiques may also be identified through sentiment analysis and catalogue automatically according to the sort of feel being portrayed and this information may assist healthcare suppliers, manufacturers and researchers to know the efficacy and

downsides of medication. As people increasingly share their experiences online, extracting accurate sentiments from the vast amount of data has become a challenging task. This research applies to the difficulty of performing sentiment analysis in the medical domain by dealing with defying issues like lack of labeled data that are crucial for emotion categorization. Further, we provided healthcare providers and manufacturers with recommendations to help analyse their strengths and weaknesses in an attempt to refine the delivery of health care services and patients satisfaction.

2.2 Related Work

Wael M.S. et al. [20] emphasizes the analysis of Arabic-language health-related content, focusing on herbal treatments for diabetes through YouTube comments. The ADHTD dataset was created specifically for this purpose. The preprocessing techniques included stemming and stop-word removal, crucial for handling Arabic text effectively. The authors tackled data imbalance with SMOTE, which generates that creates synthetic samples for minority classes and balances the dataset significantly. The study demonstrates the importance of preprocessing and data balancing for achieving high-performance sentiment classification, especially in non-English contexts.

Vijayaraghavan et al. [19] explored the role of Natural Language Processing (NLP) in improving drug review classifications and user rating predictions. They examined the contextual significance of words in reviews and tested various machine learning (ML) and deep learning (DL) models, including SVM, Neural Networks, and RNNs. By comparing Count Vectorizer (CV) and TF-IDF for feature extraction, they provided valuable insights into which techniques work best

for sentiment analysis tasks. This research sheds light on the potential of NLP algorithms in refining sentiment predictions and understanding user feedback in drug reviews.

Suhartono et al. [21] highlights the use of convolutional neural networks (CNNs) with advanced word embedding techniques like GloVe and Word2Vec. GloVe embeddings achieved a higher accuracy (84.56%) compared to Word2Vec (80.52%). They further experimented with transformer-based models like BERT and RoBERTa, demonstrating the enhanced capabilities of modern deep learning techniques in handling complex sentiment classification tasks. This work showcases how state-of-the-art architectures improve performance and sets a benchmark for future research.

Garg et al. [18] developed a drug recommender system that uses sentiment analysis to assist medical professionals in making informed decisions. They employed machine learning classifiers with feature extraction techniques like TF-IDF. The LinearSVC model stood out, achieving 93% accuracy. This study emphasizes the integration of sentiment analysis in recommender systems for healthcare, highlighting its potential to enhance treatment accuracy and reduce errors, especially in resource-constrained settings like rural areas.

Shreehar Joshi et al. [22] focuses on the multiclass classification of drug reviews into positive, neutral, and negative sentiments. Using models like SVM, Naive Bayes, and Random Forest, the study revealed that SVM achieved the highest accuracy of 80%. This underscores the efficiency of SVM in handling text data and its superiority in achieving reliable predictions in multiclass sentiment classification tasks.

Tharunya et al. [17] proposed a system combining sentiment analysis and machine learning to recommend drugs. The study employed models such as Logistic

Regression, SVM, Decision Tree, SGD, and Naive Bayes for sentiment classification. The authors also suggested integrating demographic and chemical details into the system to enhance accuracy. This study points toward a future direction for personalized medicine based on sentiment analysis.

Marthin et al. [16] using TF-IDF and Latent Semantic Analysis (LSA) for feature extraction, this study achieved high accuracies with Random Forest (84%) and SVM (83%). The focus was on optimizing traditional machine learning models for efficiency and accuracy, making this study a valuable reference for researchers working on resource-efficient SA models.

Sivakumar et al. [15] used LSTM which falls under category of Recurrent Neural Networks to classify the drug reviews into sentiment; positive, neutral negative. As a rule-based sentiment analysis tool, Vader Lexicon was applied for pre-processing the sets. With the explanatory results equal to 82.6%, the study proves the efficiency of deep learning approaches to the sentiment classification based on the text material.

Mohammed Nazim et al. [23] compared several ML models, including SVM, Naive Bayes, Logistic Regression, and Random Forest, for sentiment classification of drug reviews. With an accuracy of 85%, SVM emerged as the best-performing model. This reinforces the idea that SVM remains a strong choice for sentiment analysis tasks due to its ability to handle high-dimensional data effectively.

Garg et al. [10] explored binary sentiment classification using several ML algorithms. Logistic Regression (LR) delivered the highest accuracy of 91%, outperforming other methods. By testing various feature extraction techniques, the research highlighted the role of robust algorithms and preprocessing in achieving high accuracy in binary sentiment classification.

Kyaing et al. [11] focusing on a multiclass dataset collected from WebMD, this study applied a linguistic approach to drug sentiment analysis. It achieved an accuracy of 69%, surpassing the performance of SVM models used for comparison. This work emphasizes the importance of domain-specific linguistic features in enhancing sentiment classification.

Korkontzelos et al. [12] investigated sentiment analysis features to detect adverse drug reactions (ADRs) in online posts. Using binary classification, their models achieved an accuracy of 80%. This research highlights the importance of identifying ADRs from social media, providing a practical application for sentiment analysis in pharmacovigilance.

Sridevi et al. [14] proposing an ontology-based model, this research integrates domain knowledge like drug names and medical conditions into sentiment analysis. This approach improves classification accuracy by resolving ambiguities in textual data, showcasing the potential of combining domain expertise with machine learning techniques.

Balahur et al. [9] used Twitter datasets to explore unigram and bigram features with SVM models. This research demonstrated the value of combining simple textual features for supervised machine learning approaches to sentiment analysis, providing foundational insights for future work.

Salas-Zárate et al. [7] applied aspect-based sentiment analysis on diabetic-related tweets, benchmarking three N-gram extraction methods. The "N-gram around" technique, which considers words before and after the aspect, proved most effective. This study highlights the importance of aspect-level analysis in understanding context within sentiments.

Jianqiang et al. [6] combining prior polarity scores with n-gram features, this study created an ensemble classifier for sentiment analysis. Logistic Regression, with an accuracy of 86%, outperformed the baseline, showcasing the strength of ensemble models in improving sentiment analysis accuracy.

Whitehead et al. [5] explored ensemble methods such as bagging and boosting, demonstrating their superiority over single classifiers. By applying these techniques on diverse datasets, the study provided practical insights into ensemble learning for sentiment analysis.

Noferesti et al. [4] focuses on sentiment classification of patient reviews. The paper's emphasis on preprocessing and data processing methods provides transferable insights for sentiment analysis in healthcare applications.

T. Chen et al. [3] proposed a fuzzy-rough feature selection model, using Bag-of-Words (BOW) and TF-IDF techniques for sentiment classification. Achieving 67% accuracy with Random Forest, this research highlights the importance of feature selection in handling noisy datasets.

2.3 Comparative Analysis and Summary

The evolution of sentiment analysis (SA) techniques in healthcare, particularly for drug reviews, highlighting diverse methodologies and datasets. Preprocessing and data handling are critical, with Wael M.S. et al. emphasizing stemming, stop-word removal, and SMOTE to balance imbalanced datasets. Similar importance on preprocessing was noted by Noferesti et al. and Korkontzelos et al., especially for tasks like adverse drug reaction detection. Traditional machine learning models, such as Logistic

Regression (LR) and Support Vector Machine (SVM), demonstrated consistent performance, with Garg et al. [10] achieving 91% accuracy in binary classification using LR, and Shreehar Joshi et al. [22] attaining 80% accuracy for multiclass classification with SVM. Ensemble methods, explored by Whitehead et al., proved superior to standalone models, while advanced techniques like Bagging and Boosting further enhanced performance in tasks requiring robust generalization. Deep learning methods, including CNNs, LSTMs, and transformers like BERT and RoBERTa, showcased superior accuracy in capturing complex sentiment patterns, as evidenced by Suhartono et al., with CNNs achieving 84.56% accuracy using GloVe embeddings. However, these approaches often come with higher computational costs. Feature extraction techniques also play a vital role, with TF-IDF frequently outperforming Count Vectorizer by capturing contextual nuances, as demonstrated in studies by Vijayaraghavan et al. Advanced embeddings like GloVe and Word2Vec, as utilized by Suhartono et al., further improved performance in deep learning applications. Specialized approaches, such as aspect-based sentiment analysis by Salas-Zárate et al. and drug recommendation systems by Garg et al. [18], underscore the value of tailoring models to specific domains. The integration of demographic and domain-specific features, as suggested by Tharunya et al., points to promising advancements in personalized medicine. In summary, while traditional machine learning models like LR and SVM remain effective for structured datasets, the future of SA lies in leveraging deep learning techniques, ensemble methods, and domain-specific optimizations to address the growing complexity of healthcare data. This progression underscores the importance of combining advanced algorithms, preprocessing strategies, and domain expertise for robust and reliable sentiment classification. By evaluating model performance, preprocessing methods, and user feedback insights, this research

seeks to identify the most effective approaches for analyzing patient sentiments. The expected outputs will significantly enhance patient safety, empower informed decision-making, and improve healthcare outcomes. By implementing a robust drug recommendation system, healthcare professionals will be better equipped to support patients in navigating medication choices. The findings of this research will contribute to advancing public health initiatives and enhancing the overall quality of healthcare delivery. This research advances drug recommendation systems by integrating state-of-the-art NLP techniques, machine learning, deep learning, cross-validation, and collected additional data and data balancing methods, Multi-class Text Classification, leading to improved performance and supporting more informed decision-making in medication management.

2.4 Scope of the Problem

The generalisation of the findings is also limited because the research mainly draws inference from data available at the UCI Machine Learning Repository, additional variables that may influence the sentiment of a patient have not been included. This limitation can perhaps hamper actualization of the model in real life health facilities since patient experiences may not always be the same. Patient satisfaction may also be different around the world and patient culture may affect attitude and this work does not reflect that. As far as I know, no one prior to me examined multiclass SAs of drug reviews that were crawled from drugs.com and which garnered a great accuracy score utilizing ML algorithms. If patients provide inaccurate information, the drug recommendation system may not be struggle to identify appropriate

medications. This can lead to misguided recommendations, as the system relies on the accuracy of user-generated data to evaluate sentiments and assess drug effectiveness. Consequently, erroneous feedback can distort the underlying analysis, potentially resulting in adverse outcomes for patients and undermining the system's credibility. Ensuring data accuracy and encouraging patients to provide truthful and detailed information are crucial for the reliability of the recommendation system. While our research addresses numerous medication side effects, many drugs may not be included in the analysis. This absence can lead to confusion within the recommendation system, as it may not struggle to provide accurate guidance for medications not represented in the dataset. When patients inquire about these unlisted drugs, the system may be unable to generate reliable recommendations or accurately assess potential side effects, ultimately compromising patient safety and care. To enhance the system's effectiveness, it would be beneficial to continually update the dataset and incorporate a wider range of medications and their associated side effects. Many medications share similar side effects, which can lead to confusion among patients when using the recommendation system. This overlap makes it challenging for patients to distinguish between the side effects of different drugs, potentially resulting in misinterpretation of their experiences and concerns. When patients encounter common side effects, they may struggle to identify which specific medication is causing their. Symptoms, leading to uncertainty about treatment decisions. To mitigate this confusion, the recommendation system should include clear, comprehensive information on side effects, along with contextual guidance to help patients understand the relationships between different medications and their associated risks.

2.5 Challenges

Advance NLP Techniques: Implementing advanced Natural Language Processing (NLP) techniques can be complex due to the need for accurate text preprocessing, feature extraction, and model selection, which may require extensive expertise and fine-tuning. Machine Learning Model**:** Developing robust machine learning models necessitates careful selection of algorithms, hyperparameter tuning, and validation to ensure accurate and reliable predictions. **Cross-Validation:** In this regard, cross-validation is a critical part of model testing, but integrating this process takes time and becomes increasingly complex with large datasets and may obscure the analysis results. **Data Balancing Methods**: Effectively collected additional data in various dataset, internet source and also applying data balancing techniques (like SMOTE) is critical to address class imbalances, but finding the right approach and ensuring the quality of synthetic data can be challenging. **Multi-class Text Classification:** Accurately classifying text data into multiple sentiment categories requires sophisticated algorithms and may be hindered by ambiguous or overlapping data points. **Improved Performance:** Achieving significant performance improvements across various metrics (accuracy, precision, recall) necessitates a comprehensive approach to model development, evaluation, and continuous refinement. **Recommender Systems Build:** Building an effective recommender system that can adapt to user preferences and personalized recommendations involves challenges in understanding user needs and integrating diverse data sources than provide medicines.

CHAPTER 3
Research Methodology

3.1 Introduction

This research uses modern Natural Language Processing (NLP) and high-level embrace of Machine Learning (ML) to accomplish the necessary task of sentiment analysis of the drug reviews efficiently. The process starts with data cleansing to eliminate all sorts of data inconsistencies that may hinder the analysis process. Some of these includes: converting all text to lowercase, collapsing sentences to words, eliminating special characters and eradicating stop words. Stop words and word stemming and lemmatization are used to preprocess the data and getting it in a usable form for training the model. For this study, we employed a database acquired from UCI official site of Machine Learning Database and merged it with other dataset for balancing the data. Also, there is class imbalance in the dataset; to tackle this, there is a combining technique known as SMOTETomek This technique uses SMOTE to create a new instance, especially for the minority class while Tomek Links to remove noisy data or overlapping data. This combined two-part approach makes the data more balanced and the results more accurate in each class so that there is less chance of any classes being favored or dismissed. Eleven machine learning: a disparate collection models is implemented for sentiment classification. These include Multinomial Logistic Regression (LR), Linear Support Vector Classifier (Linear SVC), Decision Trees (DTC), Extra Trees (ETC), Random Forests (RF), Multinomial Naive Bayes (NB), K-Nearest Neighbors (KN), AdaBoost, Bagging Classifier (BGC), Gradient Boosting Decision Tree (GBDT), and XGBoost. These models were trained using different feature extraction

methods which include unigram and bigram, Trigram, up to N-grams features which consider both the word and the content surrounding it. This, allows the models to define and analyze certain categories of the sentiments, behind drug reviews. The type of data that was used in this study comprises the drug reviews amounting to 161,297 retrieved from the UCI Machine Learning Repository; records were balanced from the DrugLib dataset containing 1,022 records, the Medical Reviews dataset got from Kaggle having 68,192 records and 14,118 records gotten from other websites. To enhance the efficiency of the assessment of the cross-validation, it is utilised in ten folds to contribute to the lowering of the bias of the model. This method ensures that the models are tested on different subsets of data, providing a robust measure of their ability to generalize. Stratified sampling is used during training and testing to maintain proportional representation of each sentiment class, ensuring balanced performance across all subsets. Model performance is assessed using multiple evaluation metrics, including accuracy, precision, recall, F1-score, and the ROC (Receiver Operating Characteristic) curve. Among all models, the Logistic Regression (LR) achieves the best results, with an impressive accuracy score of 86.86%. This highlights its exceptional ability to differentiate between sentiment classes, making it the most effective model for this task.

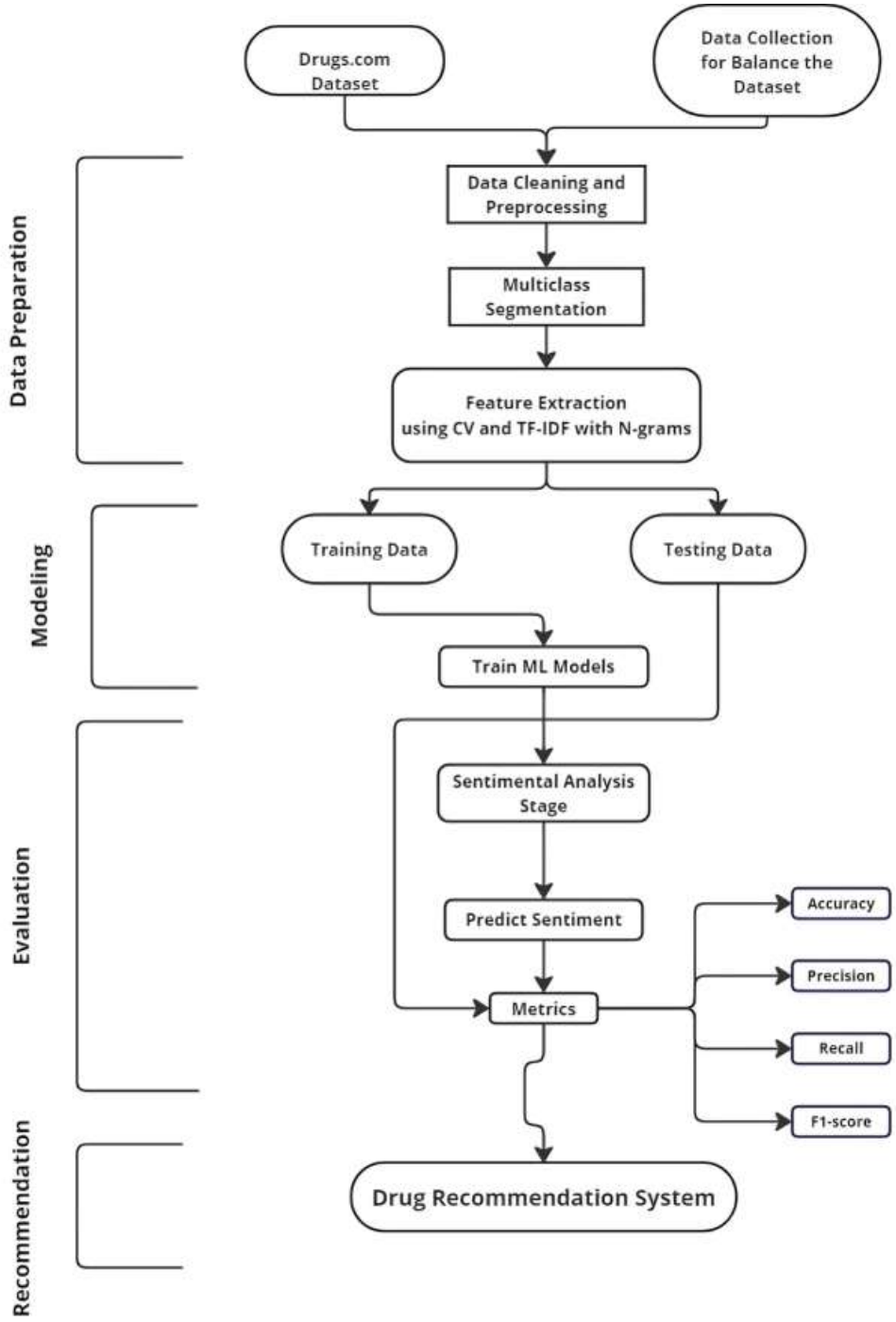
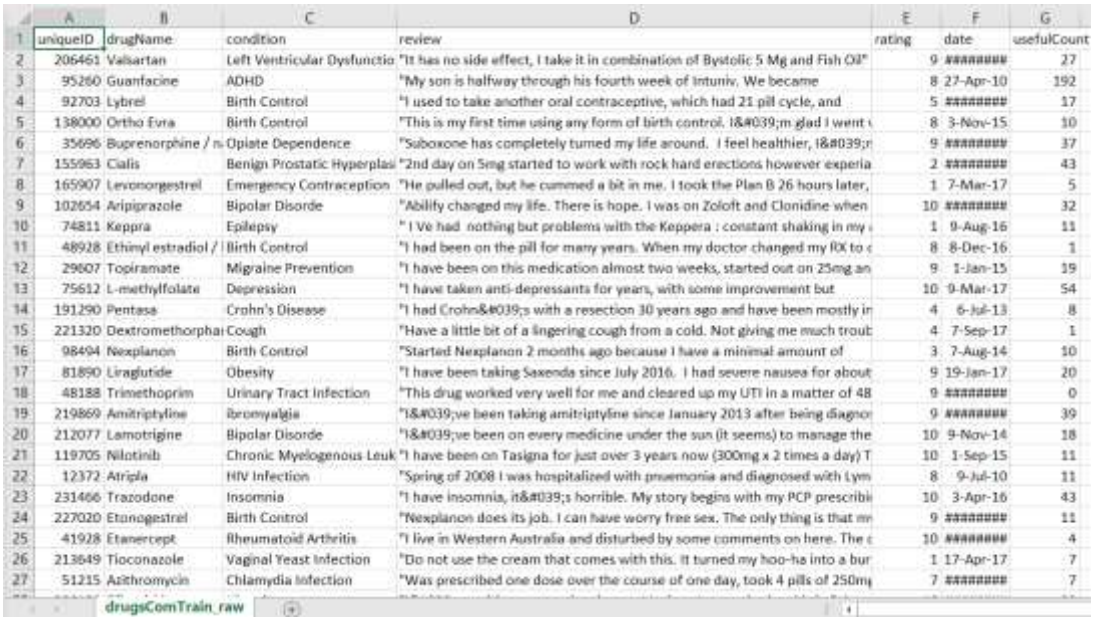


Figure 3.1: Pipeline of the Project

3.2 Dataset Description

The data set for this investigation has been sourced from the UCI Machine Learning Repository a popular online database of datasets for Machine Learning research. The repository, created in 1987 by the University of California, Irvine, is widely regarded as a leading resource for datasets in machine learning, artificial intelligence, and data mining. It is frequently utilized for empirical research, particularly in fields like sentiment analysis, natural language processing, and healthcare modeling.



	A	B	C	D	E	F	G
1	uniqueID	drugName	condition	review	rating	date	usefulCount
2	206461	Valsartan	Left Ventricular Dysfunctio	"It has no side effect, I take it in combination of Bystolic 5 Mg and Fish Oil"	9	########	27
3	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of Intuniv. We became	8	27-Apr-10	192
4	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, which had 21 pill cycle, and	5	########	17
5	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth control. I'm glad I went	8	3-Nov-15	10
6	35696	Buprenorphine / n	Opiate Dependence	"Suboxone has completely turned my life around. I feel healthier, I'r	9	########	37
7	155963	Cialis	Benign Prostatic Hyperplasi	"2nd day on 5mg started to work with rock hard erections however experia	2	########	43
8	165907	Levonorgestrel	Emergency Contraception	"He pulled out, but he cummed a bit in me. I took the Plan B 26 hours later,	1	7-Mar-17	5
9	102654	Aripiprazole	Bipolar Disorde	"Abilify changed my life. There is hope. I was on Zoloft and Clonidine when	10	########	32
10	74811	Keppra	Epilepsy	" I Ve had nothing but problems with the Keppera : constant shaking in my	1	9-Aug-16	11
11	48928	Ethinyl estradiol /	Birth Control	"I had been on the pill for many years. When my doctor changed my RX to	8	8-Dec-16	1
12	29807	Topiramate	Migraine Prevention	"I have been on this medication almost two weeks, started out on 25mg an	9	1-Jan-15	19
13	75612	L-methylfolate	Depression	"I have taken anti-depressants for years, with some improvement but	10	9-Mar-17	54
14	191290	Pentasa	Crohn's Disease	"I had Crohn's with a resection 30 years ago and have been mostly in	4	6-Jul-13	8
15	221320	Dextromethorphai	Cough	"Have a little bit of a lingering cough from a cold. Not giving me much troub	4	7-Sep-17	1
16	98494	Nexplanon	Birth Control	"Started Nexplanon 2 months ago because I have a minimal amount of	3	7-Aug-14	10
17	81890	Liraglutide	Obesity	"I have been taking Saxenda since July 2016. I had severe nausea for about	9	19-Jan-17	20
18	48188	Trimethoprim	Urinary Tract Infection	"This drug worked very well for me and cleared up my UTI in a matter of 48	9	########	0
19	219869	Amitriptyline	ibromyalgia	"I've been taking amitriptyline since January 2013 after being diagno	9	########	39
20	212077	Lamotrigine	Bipolar Disorde	"I've been on every medicine under the sun (it seems) to manage the	10	9-Nov-14	18
21	119705	Nilotinib	Chronic Myelogenous Leuk	"I have been on Tasigna for just over 3 years now (300mg x 2 times a day) T	10	1-Sep-15	11
22	12372	Atripla	HIV Infection	"Spring of 2008 I was hospitalized with pnuemonia and diagnosed with Lym	8	9-Jul-10	11
23	231466	Trazodone	Insomnia	"I have insomnia, it's horrible. My story begins with my PCP prescribi	10	3-Apr-16	43
24	227020	Etonogestrel	Birth Control	"Nexplanon does its job. I can have worry free sex. The only thing is that m	9	########	11
25	41928	Etanercept	Rheumatoid Arthritis	"I live in Western Australia and disturbed by some comments on here. The c	10	########	4
26	213649	Tioconazole	Vaginal Yeast Infection	"Do not use the cream that comes with this. It turned my hoo-ha into a bur	1	17-Apr-17	7
27	51215	Azithromycin	Chlamydia Infection	"Was prescribed one dose over the course of one day, took 4 pills of 250mg	7	########	7

drugsComTrain_raw

Figure 3.2: Sample Dataset

The specific dataset used from Drugs.com which contains a total of 161,297 entries and balanced it collected additional data incorporating 1,022 records from the

DrugLib dataset, 68,192 records from the Medical Reviews dataset on Kaggle, and 14,118 records collected from various websites.

Table 3.2: Collected Data Sources and Quantities

Origins of collected data	Total entries
UCI Machine Learning Repository	161,297
Medical Reviews Dataset (Kaggle)	68,192
DrugLib Dataset	1,022
Various Websites	14,118

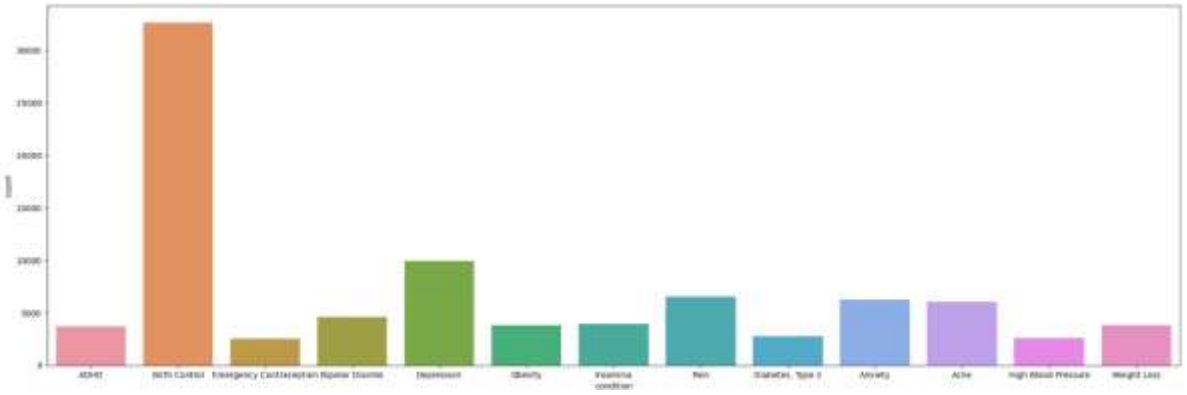


Figure 3.2.1: Particular condition for various DrugsName in Dataset

3.3 Data Cleaning, Preprocessing & Visualization

The dataset is cleaned by addressing missing values, duplicated entries, and NaN values. Missing values are identified and either replaced with appropriate substitutes or removed if necessary. Duplicate records are detected and eliminated to ensure data consistency. The text data is standardized by making all the characters in the text lowercase since all the data will be required to have the same format. Tokenization is then applied to break down the text into smaller units, such as individual words or phrases, facilitating analysis. Following this, punctuation is removed to focus solely on meaningful words. Common stopwords are filtered out to retain only the most relevant terms for analysis. Thus, stemming is used to cut down the words and eliminate suffix string, whereas lemmatization always aims at converting the words to the entries of the word base while retaining the sense. These preprocessing steps in aggregate improve the quality of the given dataset and make it ready for further analysis or to be given further to machine learning algorithms.

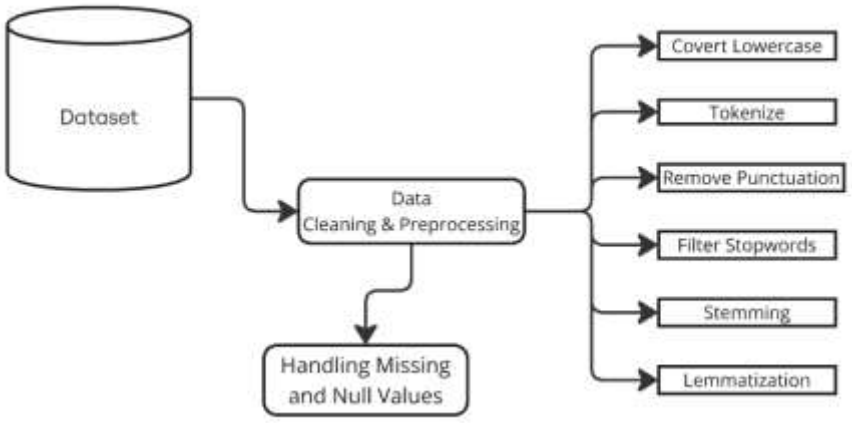


Figure 3.3: Data Cleaning and Preprocessing Workflow

3.4 Data Labeling & Balancing

The dataset from Drugs.com contains a total of 161,297 entries, categorized based on ratings from 1 to 10. Among the ratings, the highest number of entries, 50,989, is associated with a rating of 10, followed by 27,531 entries for a rating of 9. A rating of 8 corresponds to 18,890 entries, while 9,456 entries are rated 7, and 8,013 are rated 5. Lower ratings include 6,343 entries for a rating of 6, 6,513 for a rating of 3, 6,931 for a rating of 2, 5,012 for a rating of 4, and the lowest, 21,619 entries, for a rating of 1. This distribution highlights that the majority of entries are skewed toward higher ratings, with the largest share of reviews receiving a perfect score.

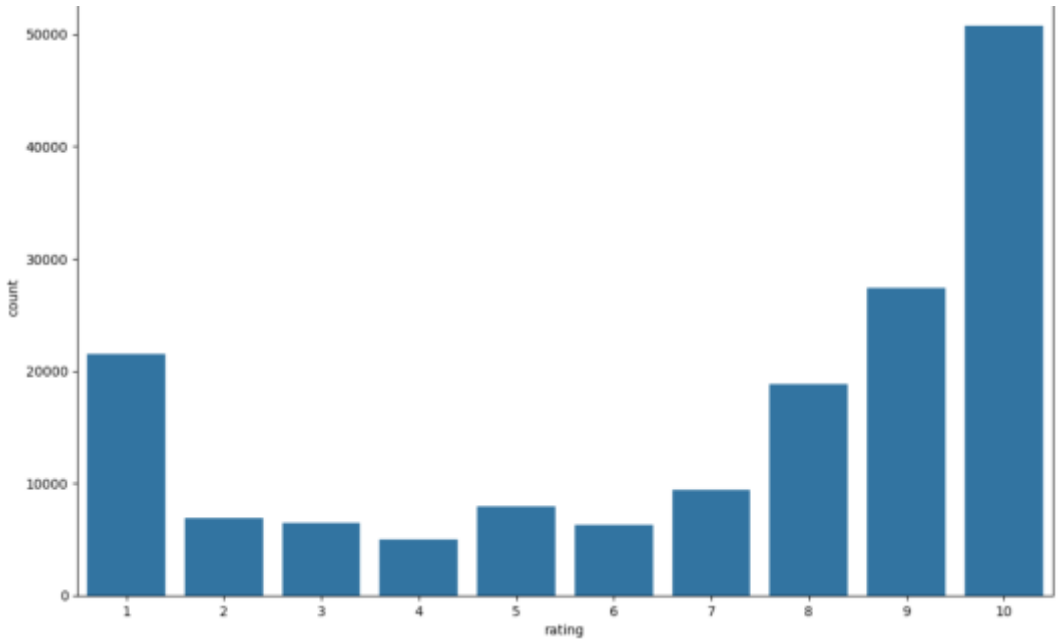


Figure 3.4.1: Drugs.com Dataset Rating Distribution

The dataset from Drugs.com, which initially contained 161,297 entries, was further enhanced by adding data from several external sources. Specifically, 1,022 records were obtained from the DrugLib dataset, 68,192 records from the Medical Reviews dataset on Kaggle, and 14,118 records collected from various websites. The expanded dataset now includes a more balanced distribution of ratings. The entries are distributed as follows: 50,989 entries with a rating of 10 and 27,531 with a rating of 9 and 26,751 with a rating of 7 and 23,733 with a rating of 5 and 22,141 with a rating of 2. Additionally, there are 21,619 entries with a rating of 1 and 19,950 with a rating of 3 and 18,890 with a rating of 8 and 18,402 with a rating of 6 and 14,623 with a rating of 4. While the dataset is now more some balanced.

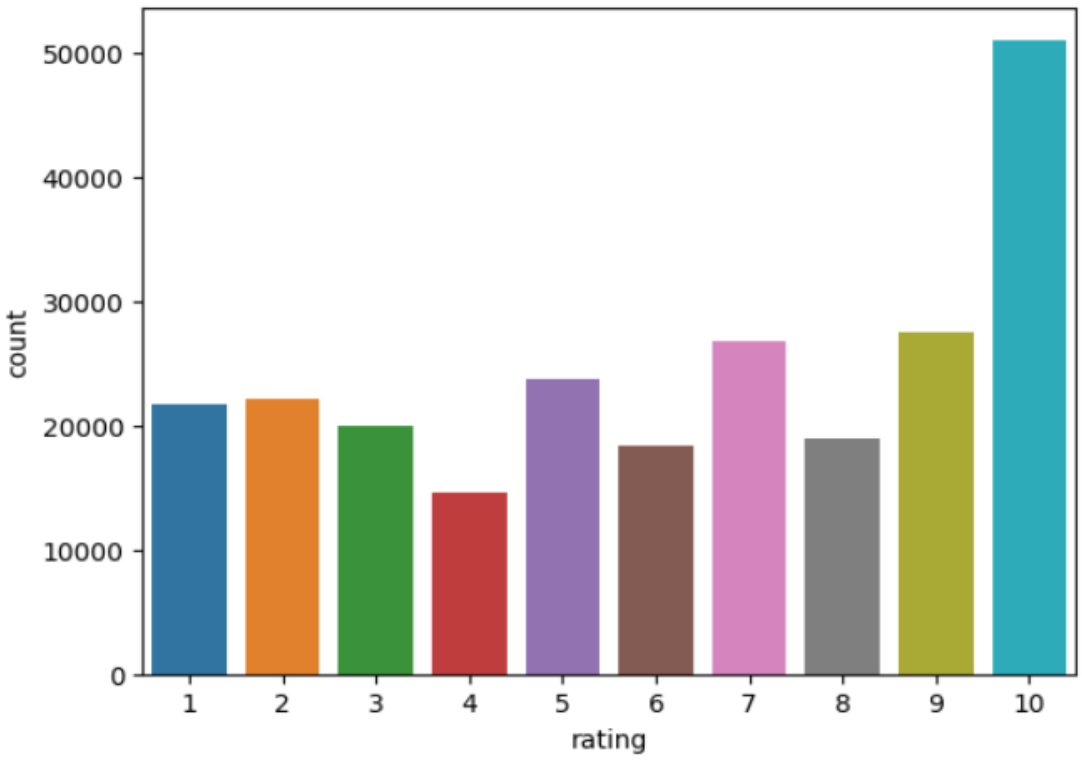


Figure 3.2.2: Rating Distribution in the Final Collected Dataset

Smith et. al in this research, drug review sentiments were usually grouped into three broad categories: positive, neutral, and negative[1] based on rating. In our study, we take this further by introducing a detailed five-level sentiment system: Frustrated, Bad, Neutral, Good, and Excited. This detailed system helps us better identify possible issues like adverse drug reactions or side effects. The sentiment levels are: Excited (ratings 7-10), showing high satisfaction or enthusiasm; Good (rating 6), showing general satisfaction; Neutral (rating 5), showing a balanced or indifferent response; Bad (rating 4), showing mild dissatisfaction; and Frustrated (rating 1-3), showing strong frustration. After dividing the dataset into multiple classes, the distribution of entries across the different classes is as follows: Class 5 contains the largest number of entries, with 124,161 records. This is followed by Class 1, which has 63,710 entries. Class 3 has 23,733 entries, while Class 4 contains 18,402 entries. Finally, Class 2 has the fewest entries, with 14,623 records. This distribution highlights that Class 5 significantly outweighs the other classes in terms of the number of entries, with the remaining classes showing progressively fewer records.

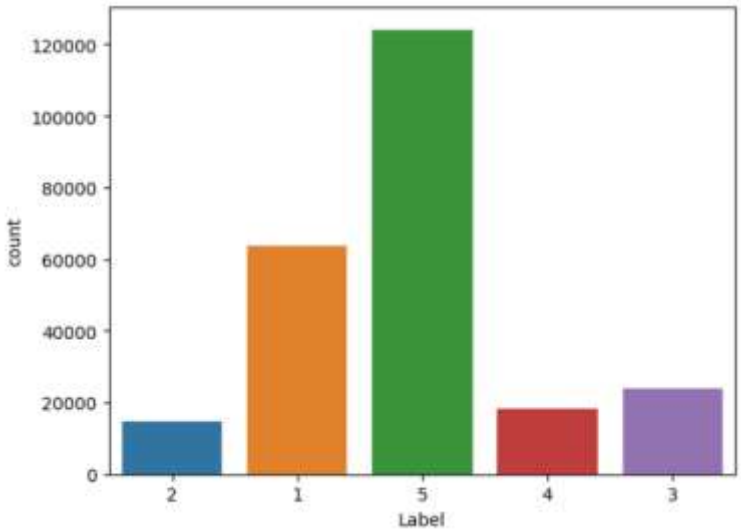


Figure 3.2.3: Multiclass Data Distribution

The dataset under study exhibits an unequal distribution across its five sentiment classes. To address this class imbalance, a process of oversampling and undersampling was exercised with the help of SMOTETomek which is the Synthetic Minority Over-sampling Technique along with Tomek Links. First of all, SMOTE was used to over-sample the minority classes to 70% of the biggest class. This process generates synthetic samples for the underrepresented classes, effectively enhancing their prevalence within the dataset. The integration of Tomek Links further refines the dataset by identifying and removing borderline or noisy samples, ensuring a cleaner and more balanced distribution. These combined techniques enhance the reliability of the training data, reducing bias in the models and improving the accuracy of sentiment classification.

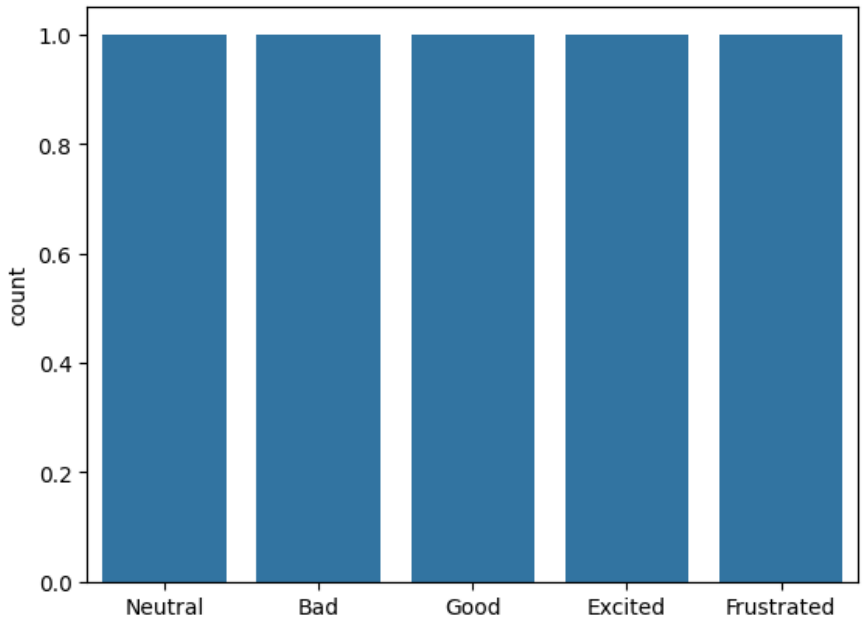


Figure 3.2.4: Multiclass Data Distribution After Applying SMOTETomek

This approach enhances the model's ability to recognize patterns in both classes,

reducing bias toward the majority class and improving predictive accuracy. However the model has achieved a higher accuracy level that were above 70% when applying Synthetic Minority Oversampling Technique (SMTOE). The effectiveness of all kinds of classifiers on SA in ML streams largely depends on text preprocessing and feature extraction.

In the dataset, user reviews are provided with ratings ranging from 1 to 10 for various drug conditions. After preprocessing the data, the ratings are mapped into a multi-class classification system to facilitate sentiment analysis. These categories include Excited (ratings 7-10), which indicate high satisfaction. Good (6), representing general satisfaction. Neutral (5), signifying a balanced or indifferent response. Bad (4), reflecting mild dissatisfaction. and Frustrated (ratings 1-3), representing strong dissatisfaction or frustration.

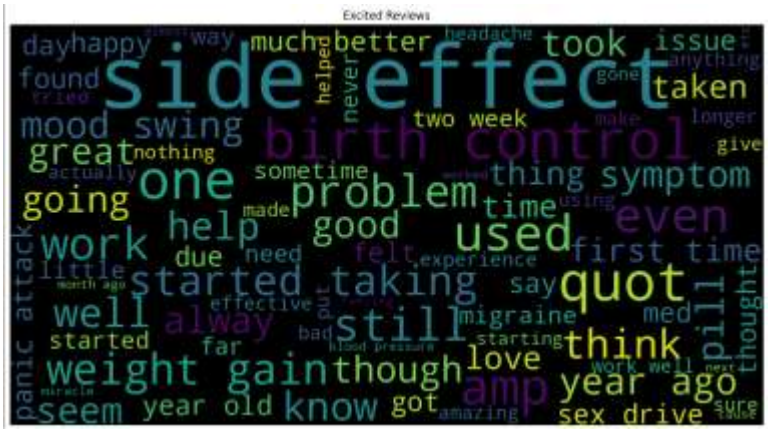


Figure 3.3.1: A Visual Exploration for Excited Sentiments

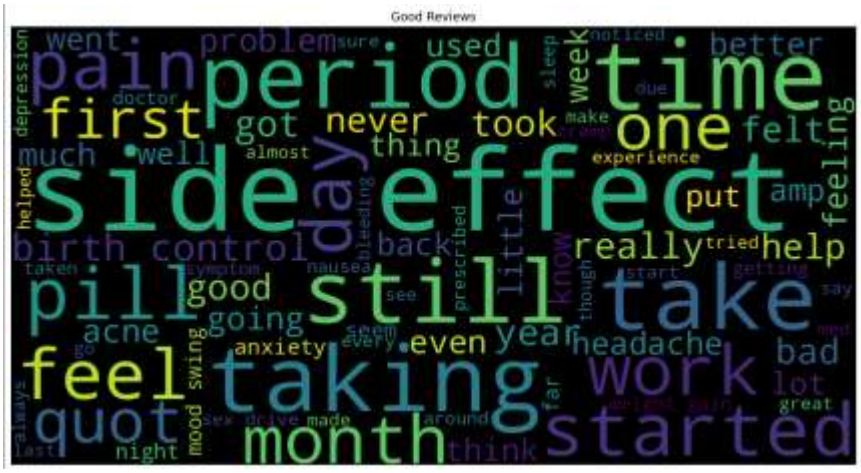


Figure 3.3.2: A Visual Exploration for Good Sentiments

Figure 3.3.3: A Visual Exploration for Neutral Sentiments

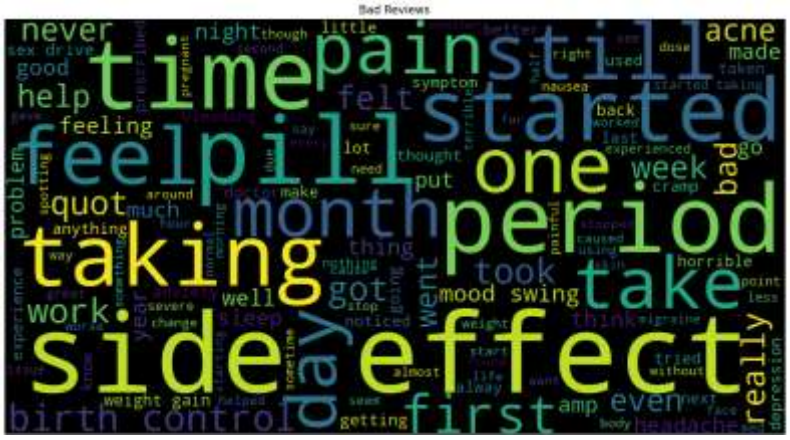


Figure 3.3.4: A Visual Exploration for Bad Sentiments

Figure 3.3.5: A Visual Exploration for Frustrated Sentiments

This structured categorization enables a more granular understanding of user sentiment, which is critical for identifying patterns in drug effectiveness and adverse drug reactions across varying satisfaction levels. The resulting multi-class sentiment labels are then utilized as target variables for machine learning models to evaluate and predict drug review sentiments effectively.

3.5 Feature Extraction

Trained on two different feature extractors which were converted into feature vectors after going through the process of TF-IDF Vectorization and CV.

TF-IDF Vectorizer:

TF is the ratio of the frequency of a term in the document to the frequency of all terms in the document and IDS is the number of documents in the corpus containing the term. The TF-IDF value can be calculated using the following

Formula:

$$\text{TF} - \text{IDF}(t, d) = \frac{TF(t, d)}{\text{timeslog}\left(\frac{N}{DF(t)}\right)} \dots\dots\dots\dots\dots (\boldsymbol{i})$$

To enhance the performance of the model, these changes were made: TF-IDF using different n-grams: unigram (single words), bigram (pairs of words), trigram (three-word sequences), and higher-order n-grams. Each of these n-grams captures more complex patterns in the text. By testing these different n-grams with TF-IDF, we can find the best way to represent the text for machine learning models, which helps improve the accuracy of the sentiment classification. This method ensures that the most meaningful words and patterns in the text are captured for better prediction.

Count Vectorizer:

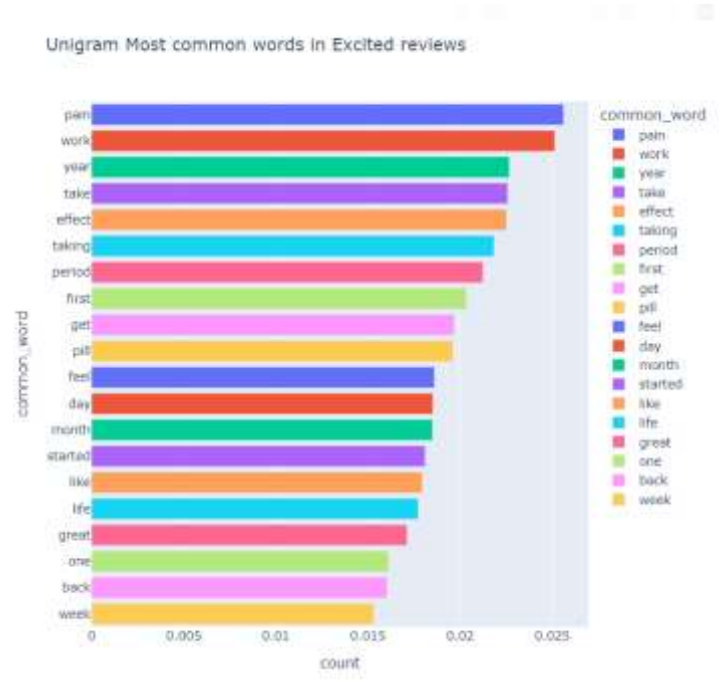
Count Vectorizer is a technique that aims to translate a set of text documents into a matrix of token frequencies. It measures the frequency of each term in a document.

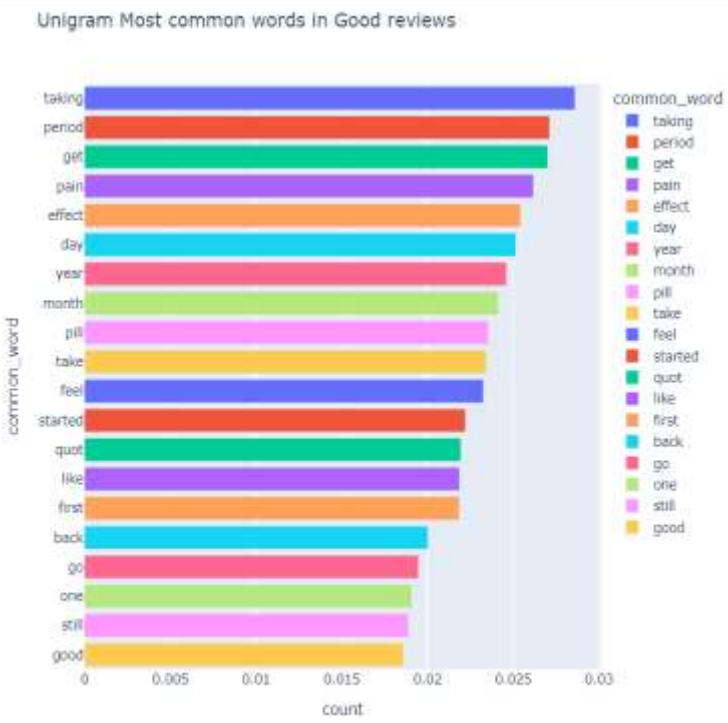
Formula:

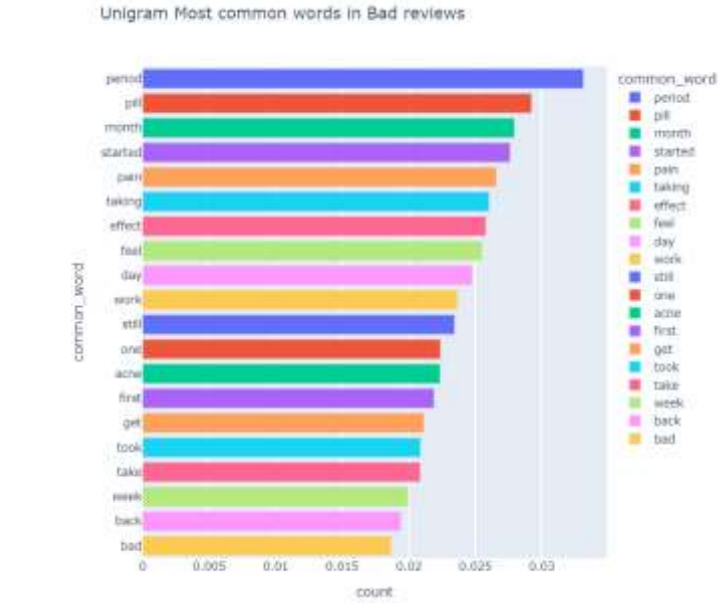
$$\text{Count}(t, d) = \text{Number of times term t appears in document d} \dots\dots\dots\dots (\boldsymbol{ii})$$

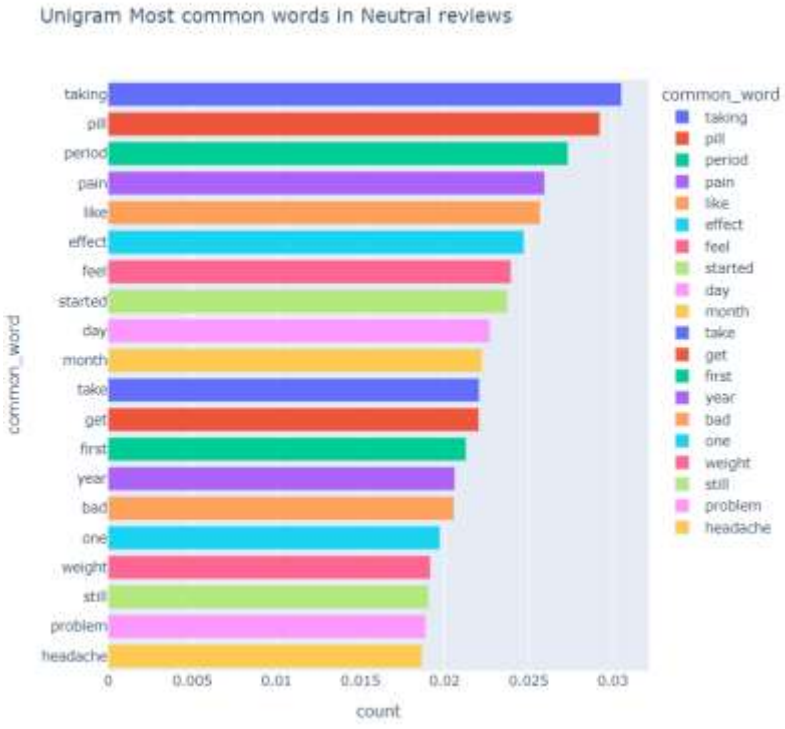
To enhance the model's performance, we also utilized **Count Vectorizer (CV)** with a range of **n-grams**, including unigrams, bigrams, trigrams, and higher-order.

To achieve more effective results, these sentiment categories are further enriched by extracting textual features using unigrams, bigrams, trigrams, and higher-order n-grams. This approach captures a broader contextual understanding of user sentiment towards drug conditions. Unigrams capture the most frequently occurring words across all sentiment categories, providing a clear understanding of common language patterns associated with each category. Bigrams identify pairs of adjacent words that frequently co-occur across all sentiment categories, helping to highlight common phrase combinations associated with different sentiment levels. Trigrams capture sequences of three consecutive words that frequently appear together across all sentiment categories, providing deeper insights into specific sentiment contexts and relationships. Higher n-grams, including quad grams and longer sequences, capture complex linguistic features and relationships across all sentiment categories.


Unigram Most common words in Excited reviews
common_word
pain
work
year
take
effect
taking
period
first
get
pill
feel
day
month
started
like
life
great
one
back
week
common_word
count
0
0.005
0.01
0.015
0.02
0.025


Unigram Most common words in Good reviews
common_word
taking
period
get
pain
effect
day
year
month
pill
take
feel
started
quot
like
first
back
go
one
still
good
common_word
count
0
0.005
0.01
0.015
0.02
0.025
0.03


Unigram Most common words in Bad reviews
common_word
period
pill
month
started
pain
taking
effect
feel
day
work
still
one
acne
first
get
took
take
week
back
bad
common_word
count
0
0.005
0.01
0.015
0.02
0.025
0.03


Unigram Most common words in Neutral reviews
common_word
taking
pill
period
pain
like
effect
feel
started
day
month
take
get
first
year
bad
one
weight
still
problem
headache
common_word
count
0
0.005
0.01
0.015
0.02
0.025
0.03

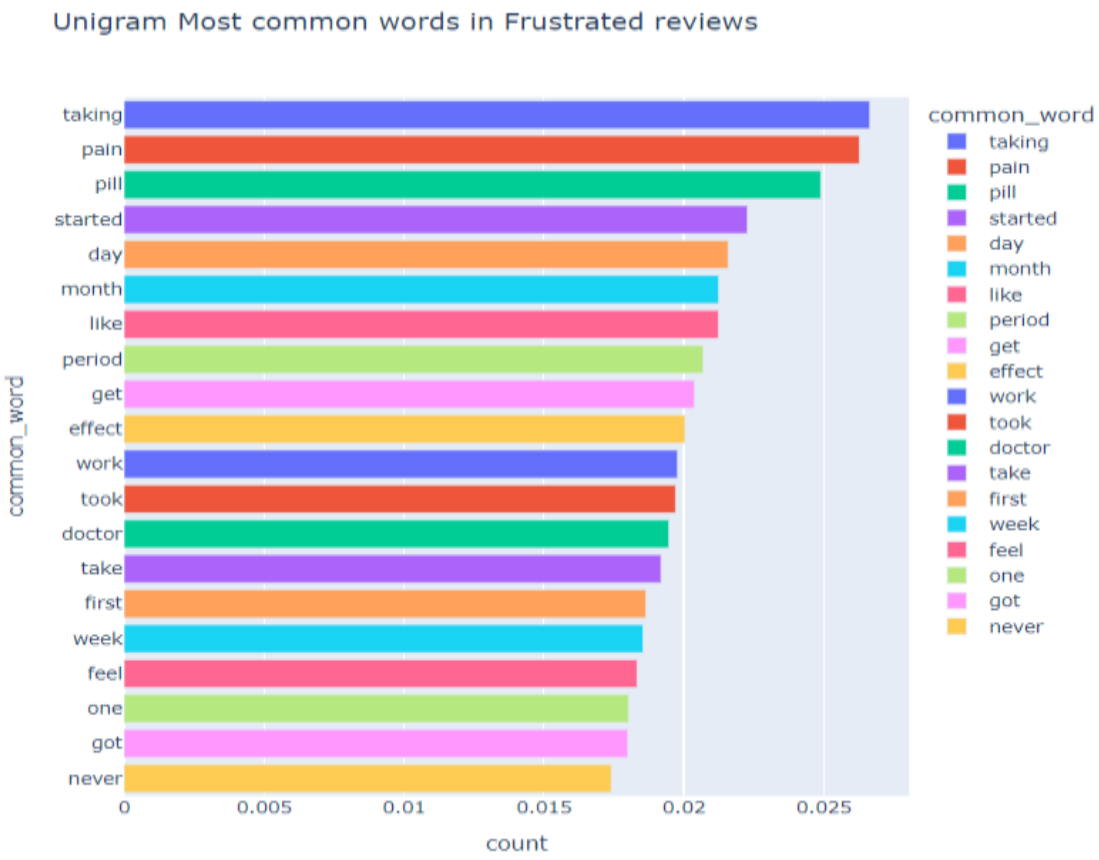
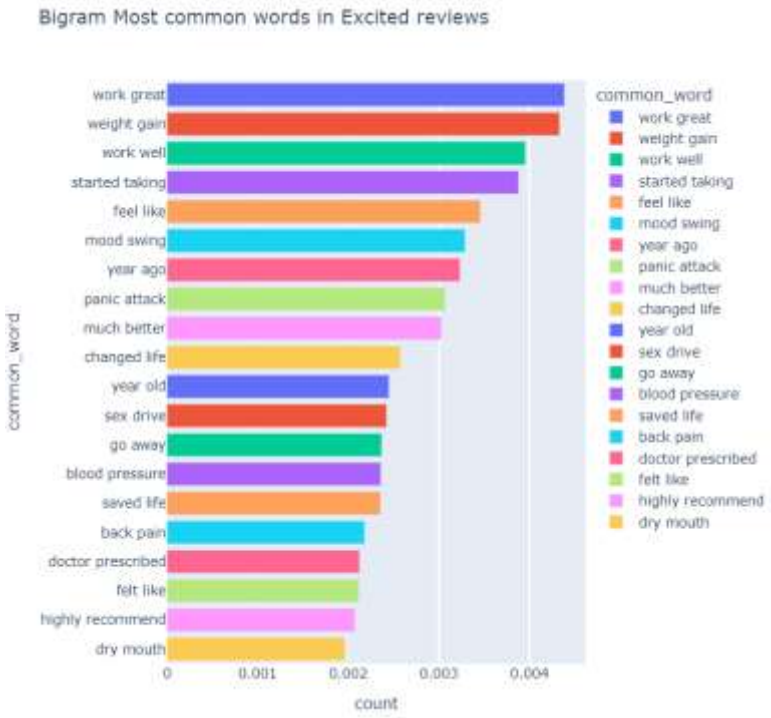
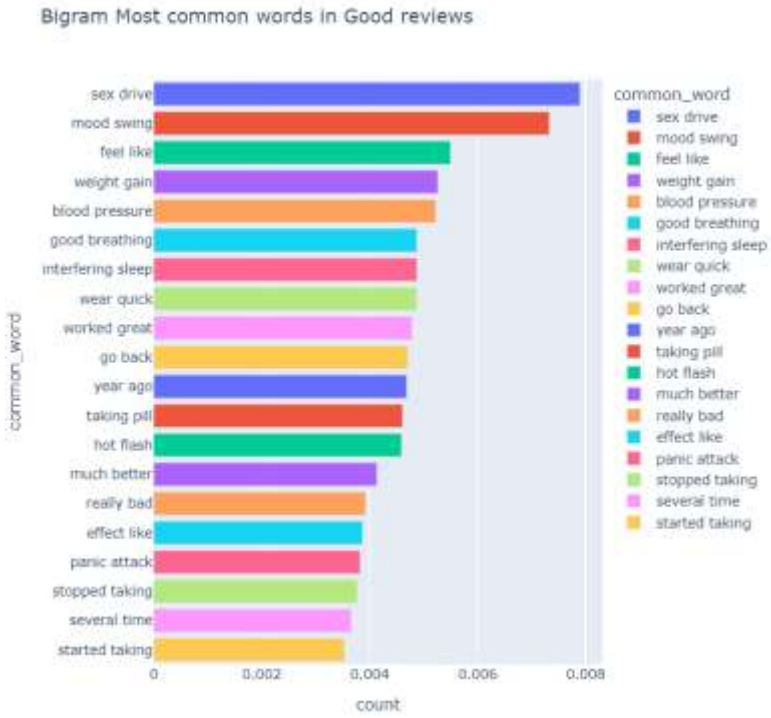
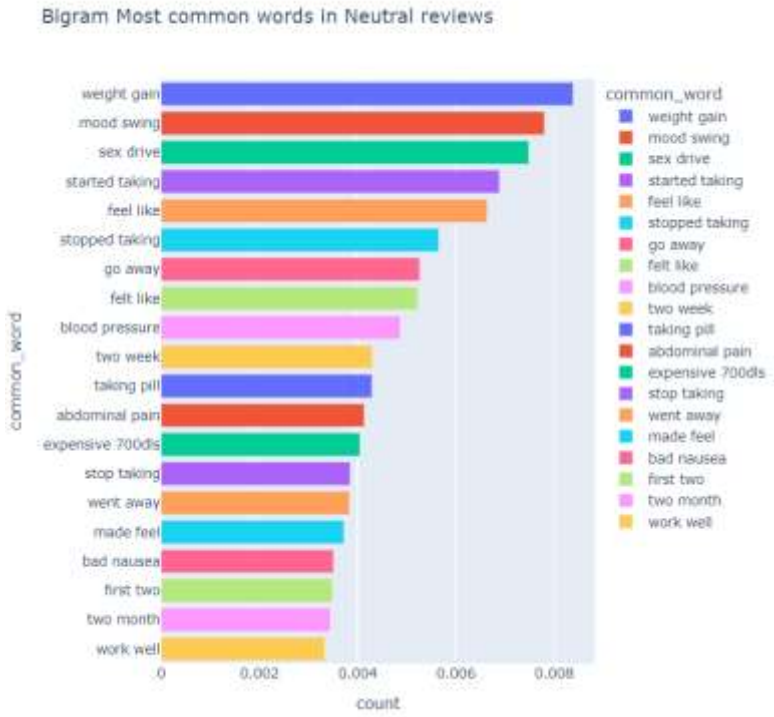
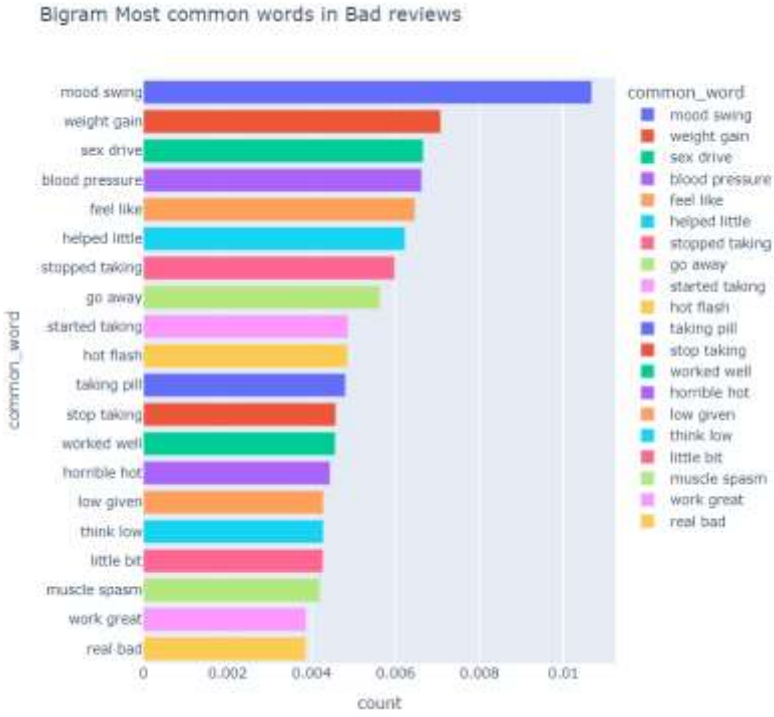


Figure 3.3.6: Unigram most common words for Excited, Good, Neutral, Bad, Frustrated Sentiments


Bigram Most common words in Excited reviews
work great
weight gain
work well
started taking
feel like
mood swing
year ago
panic attack
much better
changed life
year old
sex drive
go away
blood pressure
saved life
back pain
doctor prescribed
felt like
highly recommend
dry mouth
common_word
0
0.001
0.002
0.003
0.004
count


Bigram Most common words in Good reviews
sex drive
mood swing
feel like
weight gain
blood pressure
good breathing
interfering sleep
wear quick
worked great
go back
year ago
taking pill
hot flash
much better
really bad
effect like
panic attack
stopped taking
several time
started taking
common_word
0
0.002
0.004
0.006
0.008
count


Bigram Most common words in Neutral reviews
common_word
weight gain
mood swing
sex drive
started taking
feel like
stopped taking
go away
felt like
blood pressure
two week
taking pill
abdominal pain
expensive 700dls
stop taking
went away
made feel
bad nausea
first two
two month
work well
common_word
0
0.002
0.004
0.006
0.008
count


Bigram Most common words in Bad reviews
common_word
mood swing
weight gain
sex drive
blood pressure
feel like
helped little
stopped taking
go away
started taking
hot flash
taking pill
stop taking
worked well
horrible hot
low given
think low
little bit
muscle spasm
work great
real bad
common_word
0
0.002
0.004
0.006
0.008
0.01
count

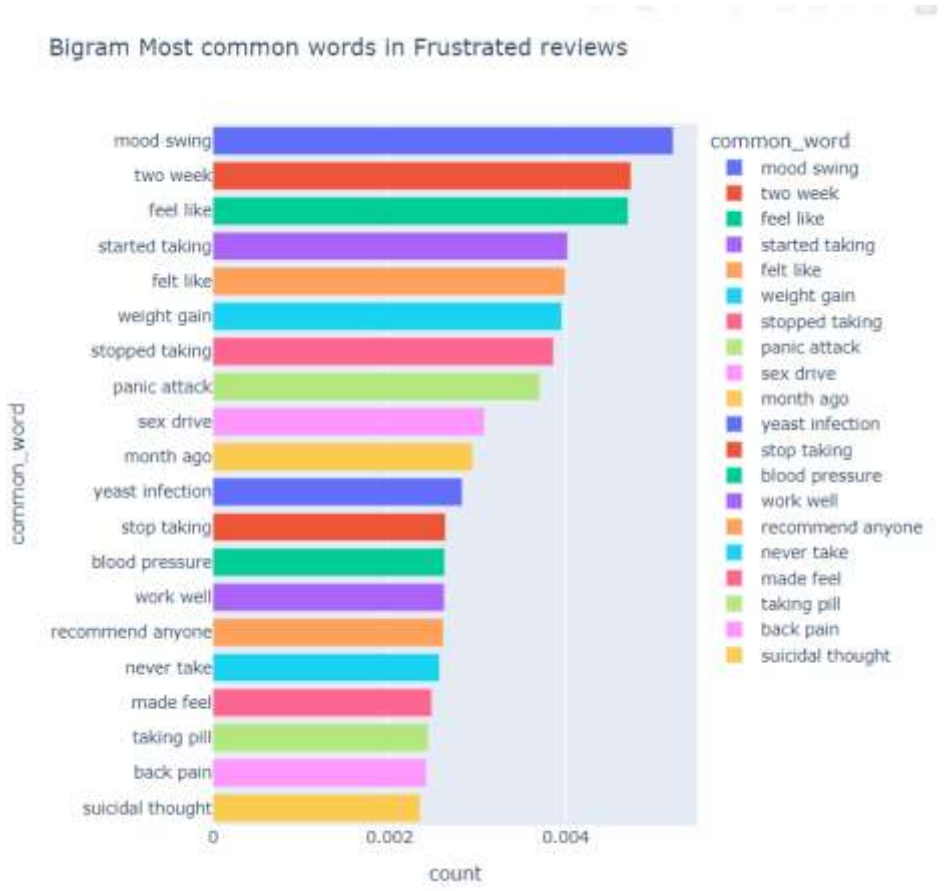
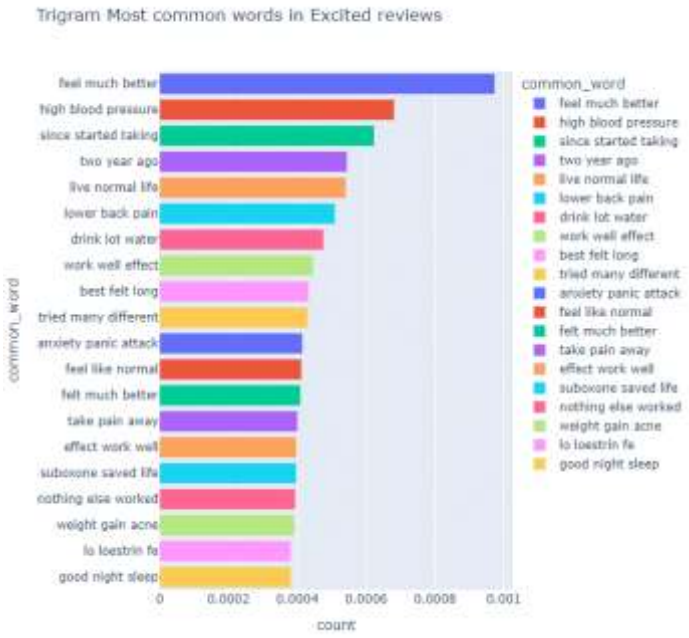
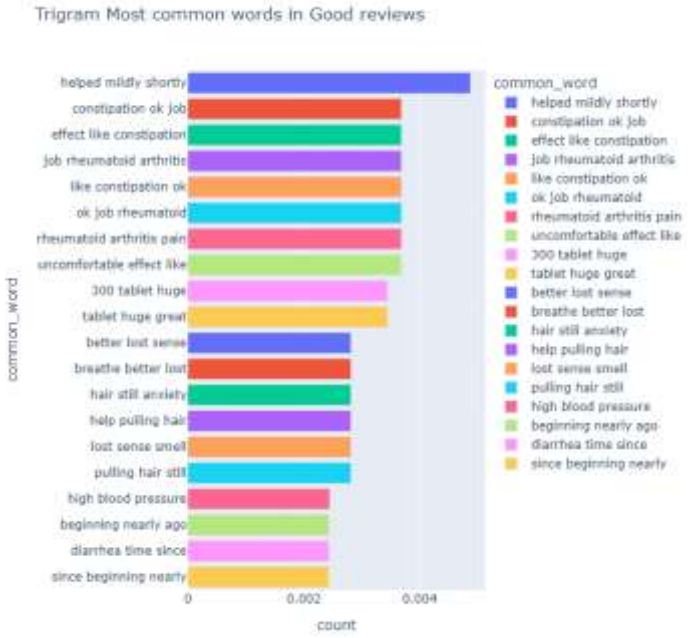
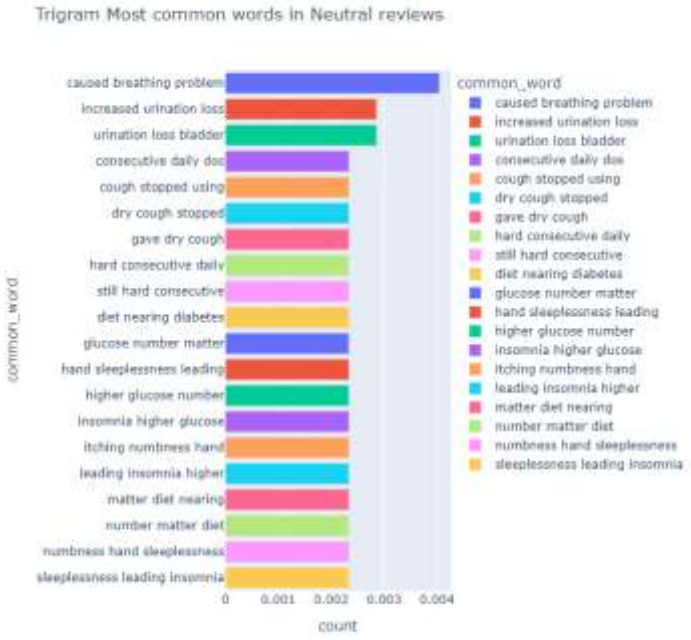
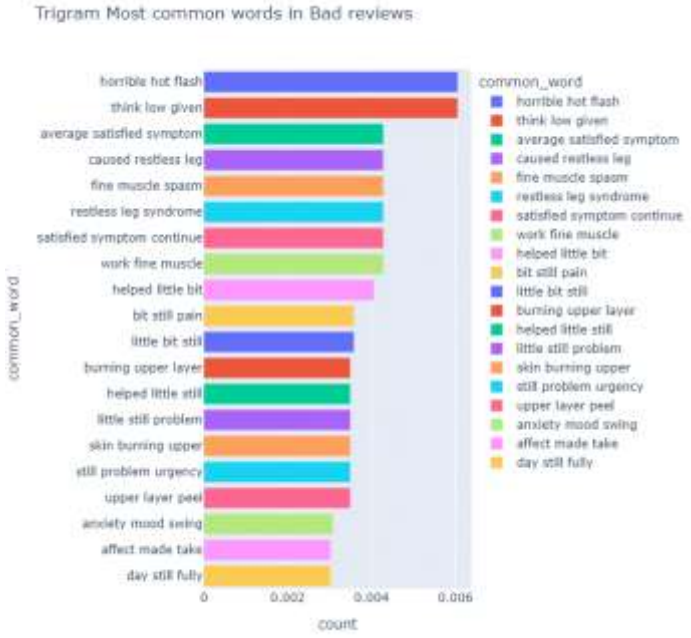


Figure 3.3.7: Bigram most common words for Excited, Good, Neutral, Bad, Frustrated Sentiments


Trigram Most common words in Excited reviews
feel much better
high blood pressure
since started taking
two year ago
live normal life
lower back pain
drink lot water
work well effect
best felt long
tried many different
anxiety panic attack
feel like normal
felt much better
take pain away
effect work well
suboxone saved life
nothing else worked
weight gain acne
lo loestrin fe
good night sleep
common_word
0
0.0002
0.0004
0.0006
0.0008
0.001
count


Trigram Most common words in Good reviews
helped mildly shortly
constipation ok job
effect like constipation
job rheumatoid arthritis
like constipation ok
ok job rheumatoid
rheumatoid arthritis pain
uncomfortable effect like
300 tablet huge
tablet huge great
better lost sense
breathe better lost
hair still anxiety
help pulling hair
lost sense smell
pulling hair still
high blood pressure
beginning nearly ago
diarrhea time since
since beginning nearly
common_word
0
0.002
0.004
count


Trigram Most common words in Neutral reviews
caused breathing problem
increased urination loss
urination loss bladder
consecutive daily dos
cough stopped using
dry cough stopped
gave dry cough
hard consecutive daily
still hard consecutive
diet nearing diabetes
glucose number matter
hand sleeplessness leading
higher glucose number
insomnia higher glucose
itching numbness hand
leading insomnia higher
matter diet nearing
number matter diet
numbness hand sleeplessness
sleeplessness leading insomnia
common_word
0
0.001
0.002
0.003
0.004
count
common_word
caused breathing problem
increased urination loss
urination loss bladder
consecutive daily dos
cough stopped using
dry cough stopped
gave dry cough
hard consecutive daily
still hard consecutive
diet nearing diabetes
glucose number matter
hand sleeplessness leading
higher glucose number
insomnia higher glucose
itching numbness hand
leading insomnia higher
matter diet nearing
number matter diet
numbness hand sleeplessness
sleeplessness leading insomnia


Trigram Most common words in Bad reviews
horrible hot flash
think low given
average satisfied symptom
caused restless leg
fine muscle spasm
restless leg syndrome
satisfied symptom continue
work fine muscle
helped little bit
bit still pain
little bit still
burning upper layer
helped little still
little still problem
skin burning upper
still problem urgency
upper layer peel
anxiety mood swing
affect made take
day still fully
common_word
0
0.002
0.004
0.006
count
common_word
horrible hot flash
think low given
average satisfied symptom
caused restless leg
fine muscle spasm
restless leg syndrome
satisfied symptom continue
work fine muscle
helped little bit
bit still pain
little bit still
burning upper layer
helped little still
little still problem
skin burning upper
still problem urgency
upper layer peel
anxiety mood swing
affect made take
day still fully

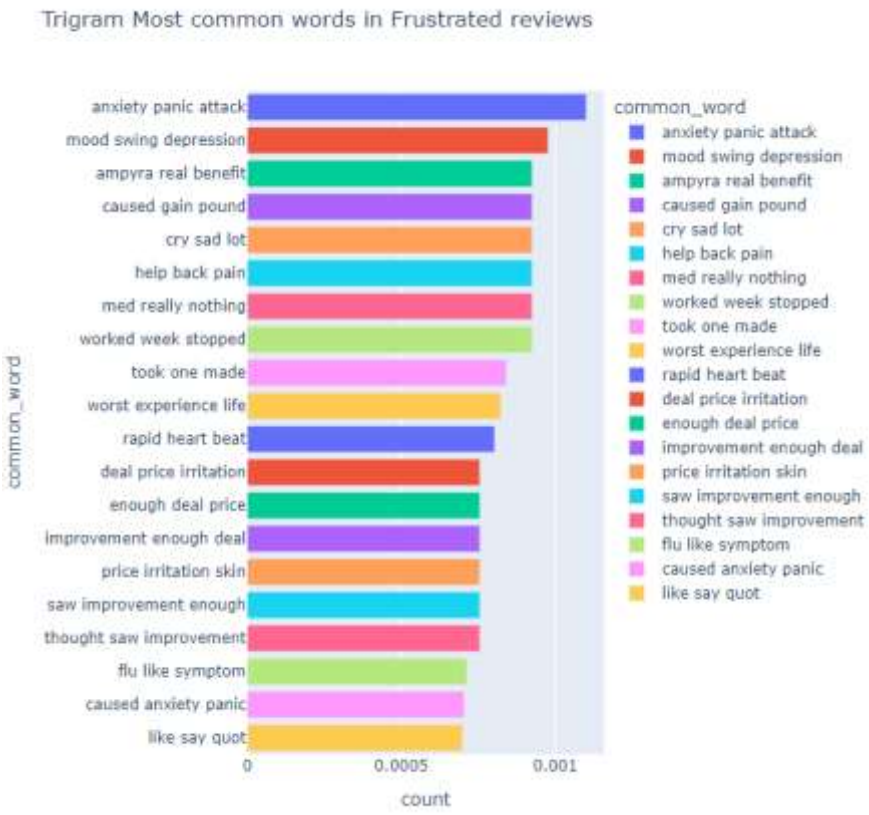
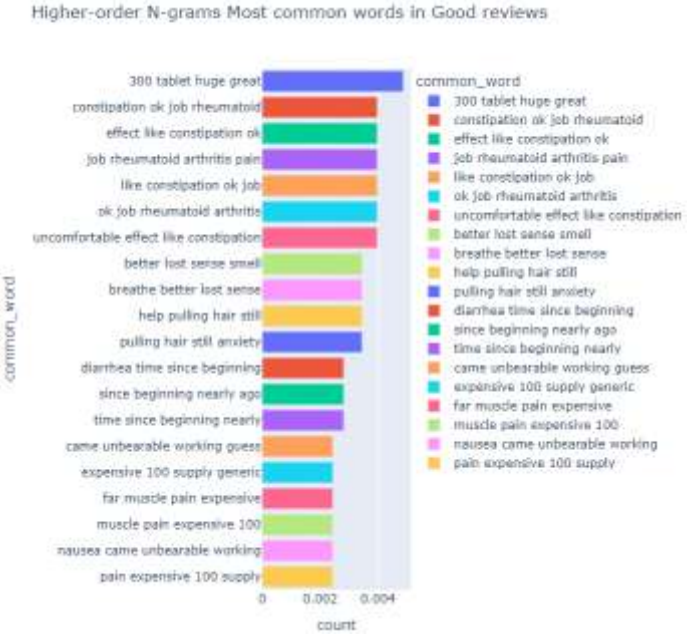
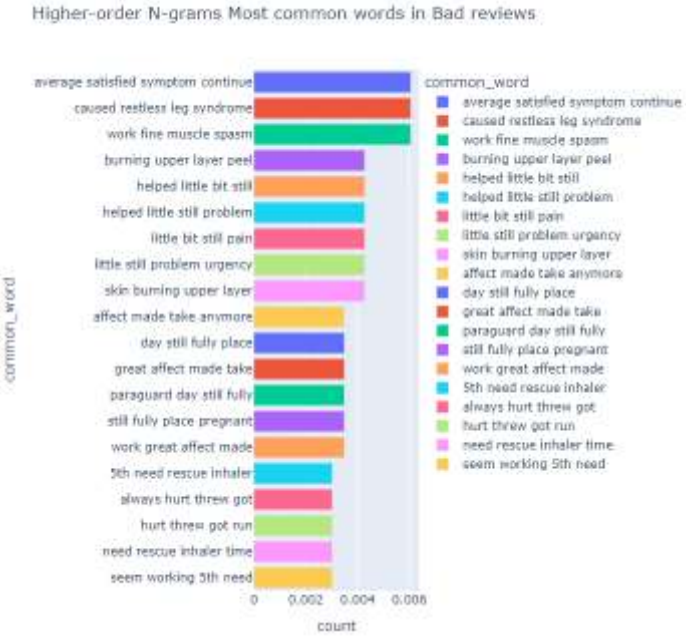


Figure 3.3.8: Trigram most common words for Excited, Good, Neutral, Bad, Frustrated Sentiments


Higher-order N-grams Most common words in Excited reviews
anything else ever tried
found helped anything else
helped anything else ever
amp sedation great pain
constipation amp sedation great
sedation great pain relief
work great feel like
better anything else tried
feel like normal person
post traumatic stress disorder
always work well effect
amazing month gained 20lbs
binge drinker several year
definitely take away hunger
excellent cleared symtoms 24hours
expensive work negative effect
femara year reaction pleased
gave life back 150mg
great effect especially adult
great made seizure controlled
common_word
0
100µ
200µ
300µ
400µ
count


Higher-order N-grams Most common words in Good reviews
300 tablet huge great
constipation ok job rheumatoid
effect like constipation ok
job rheumatoid arthritis pain
like constipation ok job
ok job rheumatoid arthritis
uncomfortable effect like constipation
better lost sense smell
breathe better lost sense
help pulling hair still
pulling hair still anxiety
diarrhea time since beginning
since beginning nearly ago
time since beginning nearly
came unbearable working guess
expensive 100 supply generic
far muscle pain expensive
muscle pain expensive 100
nausea came unbearable working
pain expensive 100 supply
common_word
0
0.002
0.004
count


Higher-order N-grams Most common words in Neutral reviews
increased urination loss bladder
dry cough stopped using
gave dry cough stopped
hard consecutive daily dos
still hard consecutive daily
glucose number matter diet
hand sleeplessness leading insomnia
higher glucose number matter
insomnia higher glucose number
itching numbness hand sleeplessness
leading insomnia higher glucose
matter diet nearing diabetes
number matter diet nearing
numbness hand sleeplessness leading
sleeplessness leading insomnia higher
tingling itching numbness hand
antiinflamatory better taking vicodin
bad antiinflamatory better taking
better taking vicodin percocet
cleared cystitis got rash
common_word
0
0.001
0.002
0.003
0.004
count
common_word
increased urination loss bladder
dry cough stopped using
gave dry cough stopped
hard consecutive daily dos
still hard consecutive daily
glucose number matter diet
hand sleeplessness leading insomnia
higher glucose number matter
insomnia higher glucose number
itching numbness hand sleeplessness
leading insomnia higher glucose
matter diet nearing diabetes
number matter diet nearing
numbness hand sleeplessness leading
sleeplessness leading insomnia higher
tingling itching numbness hand
antiinflamatory better taking vicodin
bad antiinflamatory better taking
better taking vicodin percocet
cleared cystitis got rash


Higher-order N-grams Most common words in Bad reviews
average satisfied symptom continue
caused restless leg syndrome
work fine muscle spasm
burning upper layer peel
helped little bit still
helped little still problem
little bit still pain
little still problem urgency
skin burning upper layer
affect made take anymore
day still fully place
great affect made take
paraguard day still fully
still fully place pregnant
work great affect made
5th need rescue inhaler
always hurt threw got
hurt threw got run
need rescue inhaler time
seem working 5th need
common_word
0
0.002
0.004
0.006
count
common_word
average satisfied symptom continue
caused restless leg syndrome
work fine muscle spasm
burning upper layer peel
helped little bit still
helped little still problem
little bit still pain
little still problem urgency
skin burning upper layer
affect made take anymore
day still fully place
great affect made take
paraguard day still fully
still fully place pregnant
work great affect made
5th need rescue inhaler
always hurt threw got
hurt threw got run
need rescue inhaler time
seem working 5th need

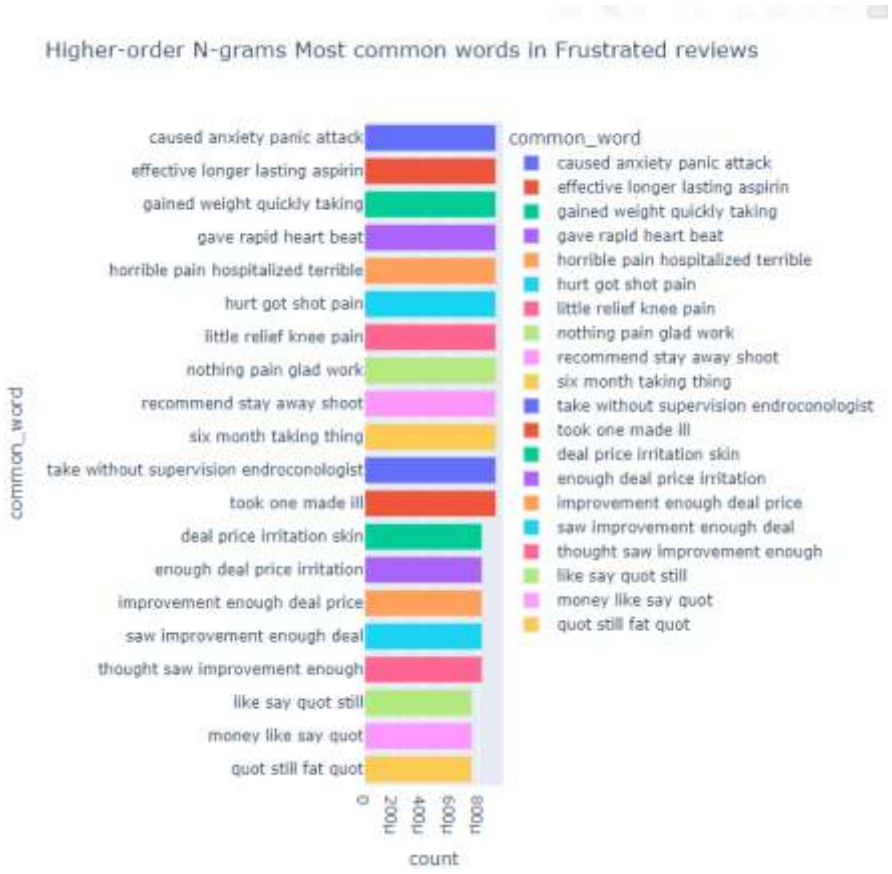


Figure 3.3.9: Higher n-grams most common words for Excited, Good, Neutral, Bad, Frustrated Sentiments

3.6 Proposed Methodology

During the preprocessing stage, the data undergoes cleaning by addressing missing values, duplicates, and NaN entries, followed by text normalization involving conversion to lowercase, punctuation removal, tokenization, and stopword removal. Stemming and lemmatization are applied to ensure words are reduced to their root forms and dictionary meanings, respectively. The user ratings are then mapped into five sentiment categories for multi-class classification: Excited (ratings 7-10), Good (rating 6), Neutral (rating 5), Bad (rating 4), and Frustrated (ratings 1-3). To handle the inherent class imbalance, the SMOTETomek technique is applied, where SMOTE generates synthetic samples for minority classes and Tomek Links removes noisy samples, resulting in a balanced dataset with minority classes reaching 70% of the majority class. Text data is transformed into numerical representations using Term Frequency-Inverse Document Frequency (TF-IDF) and Count Vectorizer (CV) with n-grams (unigrams, bigrams, trigrams, and higher-order n-grams). These n-grams capture both individual words and contextual relationships, essential for sentiment analysis. Eleven machine learning models, including Logistic Regression, Linear Support Vector Classifier, Decision Trees, Random Forests, Naive Bayes, K-Nearest Neighbors, AdaBoost, Bagging Classifier, Gradient Boosting Decision Trees, Extra Trees, and XGBoost, are trained and evaluated. Cross-validation, specifically K-fold (K=10), and stratified sampling are utilized to ensure robust model performance by reducing bias and maintaining class representation. Performance is measured using accuracy, precision, recall, F1-score, of the ROC curve. The results indicate that Logistic Regression (LR) is the best-performing model, achieving an impressive accutacy score of 86.86%, showcasing its superior ability to classify sentiment accurately using Count Vectorizer with higher order n-grams. Accurate sentiment

classification is pivotal in this study, as it ensures that the recommender system built on this analysis can provide precise, personalized drug recommendations based on user feedback. A clear and reliable sentiment analysis is essential to improve user trust, optimize drug selection, and ultimately enhance patient outcomes. This methodology effectively addresses class imbalance, enhances feature representation, and leverages state-of-the-art machine learning models for accurate sentiment classification in drug reviews.

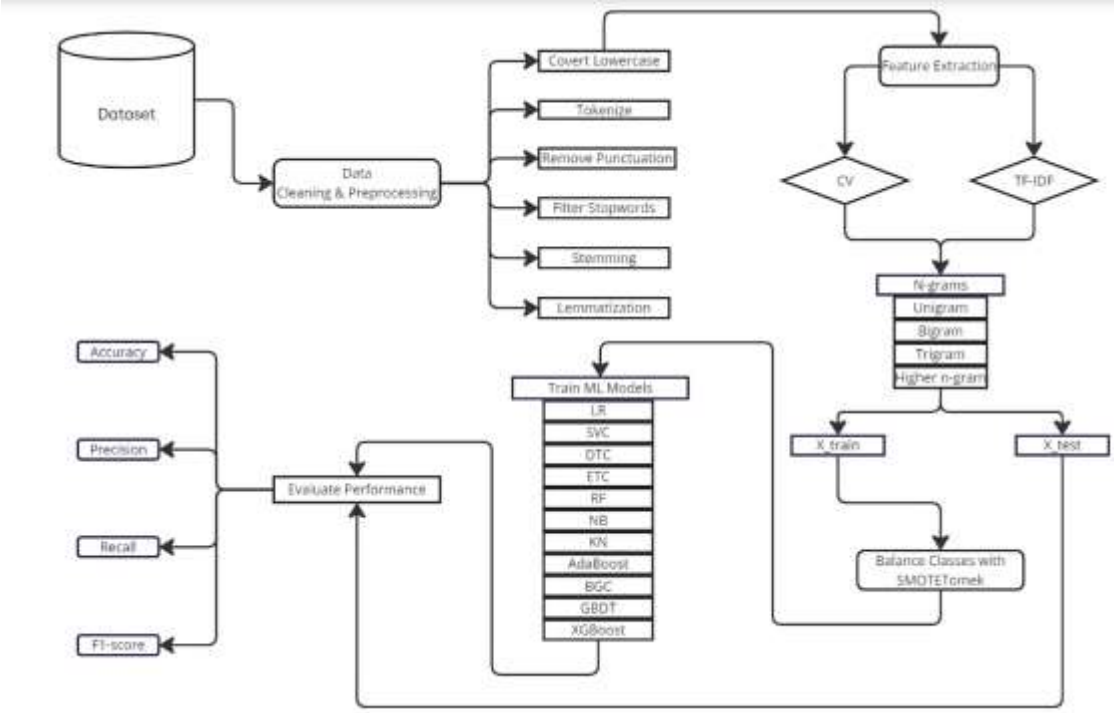


Figure 3.3.12: Design Flow of the Proposed Model

3.7 Model Training

Eleven machine learning models were trained to classify sentiment in the drug review dataset. These models include Logistic Regression (LR), Linear Support Vector Classifier (Linear SVC), Decision Trees (DTC), Random Forests (RF), Naive Bayes (NB), K-Nearest Neighbors (KNN), AdaBoost, Bagging Classifier (BGC), Gradient Boosting Decision Trees (GBDT), Extra Trees Classifier (ETC), and XGBoost. Each model was selected for its suitability in handling multi-class classification tasks and its ability to process complex data patterns. The dataset underwent transformation They were able to use two techniques of feature extraction: TF-IDF and Count Vectorizer (CV), incorporating unigrams, bigrams, trigrams, and higher n-grams to capture both individual words and their contextual relationships. To further optimize Logistic Regression (LR) performance, cross-validation with K-fold (K=10) and stratified sampling were applied. These methods help reduce model bias and ensure that the proportions of each class are consistently represented in both the training and testing sets. However, cross-validation with K-fold (K=7) provided the best accuracy for Count Vectorizer with higher order n-gram features, so this method was selected for the final model evaluation. In LR, three different sampling methods were tested to split the data:

- **Linear Sampling**: Data is split in a sequential order without shuffling.
- **Shuffled Sampling**: Data is shuffled randomly before being split.
- **Stratified Sampling**: Class proportions in the target variable remain practically the same in the training and test sets.

Given the nature of our data and task, stratified sampling was found to be the best fit. The training set and the test set preventing bias due to class imbalance and improving the proposed model has offered good prediction accuracy of both classes

of the signs. Used to measure the performance of the feature extraction using LR, Accuracy, Precision, Recall, F1 score and the ROC curve. To give you an idea, these yield an overall idea of how the LR model performs when classifying the sentiment of the drug review set. Moreover, other performance measures like precision and recall for positive and negatives sentiments, as well as for overall classification, in terms of accuracy, and F1—Score confirm that LR is a suitable algorithm for the specified task of sentiment analysis and that the model provides reliable consistent results.

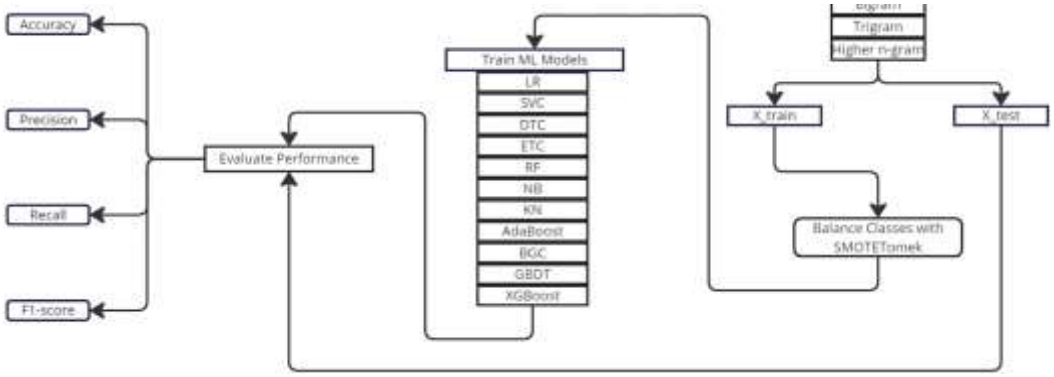


Figure 3.7: Design Flow of Train ML Models

3.6.1 Logistic Regression (LR)

Logistic Regression is a linear model classification algorithm. Indeed, it estimates the likelihood that a data point is of a given class through the logistic (sigmoid) function which maps calculated values to the [0.0, 1.0] range. The algorithms simple to implement, and performs particularly well for linearly separable data; however, it used mostly for binary classification problem although it can be generalized to multi-class problems by directly extending of one vs rest or softmax.

$$P(y = 1|X) = \frac{1}{1 + e^{-z}} \dots\dots\dots\dots\dots (\boldsymbol{iii})$$

Where:

$$z = w^T X + b \ldots\ldots\ldots\ldots (iv)$$

Where:

• yi is the true label (0 or 1) for the i-th sample.

• P(yi = 1|Xi) is the predicted probability that the i-th sample is of class

3.6.2 Linear Support Vector Classifier (Linear SVC)

Linear SVC is among the family of the support vector machine, also known as SVC that aims at determining a plane that optimally differentiates various categories of data compartments. It offers maximum margin which is defined as the largest gap between hyperplane and the closest points of every class possible either in relation to distance or can also be said that it is the maximum separation of the pattern vectors of two classes. It is best used for linearly separable data and is also good when used in high-dimensional data.

$$\boldsymbol{Minimize}\ \frac{1}{2} \| \boldsymbol{w} \|^2 \ldots\ldots\ldots\ldots (v)$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1\ \forall i \ldots\ldots\ldots\ldots (vi)$$

Where:

• w is the weight vector, and b is the bias.

• xi is the feature vector for the i-th sample.

• yi is the label for the i-th sample

3.6.3 Decision Trees (DTC)

Decision Trees are tree structures where data is bifurcated into branches by comparing the values of a feature. Each internal node is equivalent of a decision rule and each leaf node is equivalent of a class label or outcome. It grows by splitting the data until each node sends another node or till a stopping threshold is reached at the leaf node. DTs are highly comprehensible but may be hazardous by overfitting.

3.6.4 Random Forests (RF)

Random Forests are Decision Trees type classifiers that use different Decision Trees to enhance the classification or regression performance. Data is bootstrapped, meaning each tree operates on a random selection of the data set and any feature can be used for a split. The final prediction is done by a method called regression for measurement variables or classification for categorical variables. This helps to decrease over training and increases the model's ability to generalize.

$$\boldsymbol{Gini(D) = 1 - \sum_{i=1}^{k} p^2\, i} \ldots\ldots\ldots\ldots\ldots (\boldsymbol{vii})$$

Where:

- pi is the proportion of samples of class i in the dataset D.
- k is the number of possible classes.

3.6.5 Naive Bayes (NB)

Naïve Bayes is an approach to building a classifier based on Bayes 'theory to estimate the probability of the class coming with feature values. It is a greedy algorithm Because it assumes that every one feature is conditionally independent

of all other features given the other features present which though makes computation easy but is not necessarily true. It still surprises many people in its good work in some of the real-world applications as well as text classification and spam filtering even though it is based on a naive assumption.

$$\boldsymbol{P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X|C) \cdot P(C)} \ldots\ldots\ldots\ldots (vii)}$$

Where:

- P(C|X) is probability of Class C given the feature vector X.
- P(X|C) is the likelihood, and the likelihood refers to the probability of observing features X given class C.
- P(C) is the prior probability of class C.
- P(X) is the evidence, the probability of the features X.

3.6.6 K-Nearest Neighbors (KN)

K-Nearest Neighbors is an intrinsically comprehensible algorithm that estimates the class of a target data point depending upon the majority of the nearest neighbors in the feature space. The distance, for example, Euclidean is used to identify the nearest neighbors. It is distribution-free, so it does not estimate the form of any distribution of the data, although it may face high computational costs in the event of large data samples.

$$\boldsymbol{y\hat{} = majorityvote(yi|distance(x, xi) \leq k) \ldots\ldots\ldots\ldots (ix)}$$

Where:

- x is the test sample.
- xi are the k-nearest neighbors to x, and yi are their corresponding labels

3.6.7 AdaBoost

AdaBoost stands for Adaptive Boosting and is an ensemble learning method which uses a number of weak classifiers, usually Decision Trees, to create a strong classifier. It operates in batches – the weights of misclassified samples are modified to pay extra attention to them on the next loop. The last of them is the final prediction of the over complete number of weak classifiers, which doesn't let the model overfit.

3.6.8 Bagging Classifier (BGC)

In ensemble learning Bootstrap Aggregating abbreviated as Bagging is an example of a model that is worked out to make two predictions based on the different sections of data. This not only decreases the model variation effect and enhances the model stability. The Bagging Classifier often applies decision trees as their base learners and Random forest is an example of Bagging.

3.6.9 Gradient Boosting Decision Trees (GBDT)

Gradient Boosting is a sequential ensemble method the the objective of which is to build Decision Trees that reduce the errors of previous models. Optimization is done through gradient descent on a loss function, trying to minimize the residuals (error) gradually. Every tree learns from each of the previous trees and is probably the best tree for prediction. As we know, GBDT brings forward a group of highly efficient models but is known to be more computationally expensive.

3.6.10 Extra Trees (ETC)

Extra Trees or Extremely Randomized Trees are analogous to Random Forests with the difference of how splits are made. Compared with the method that search the optimal splitting, Extra Trees directly make the splitting at each node running and it will bring the diversity among the trees. This randomness is in most cases

beneficial because it decreases the amount of variance and increases the training rate, but at the same time has almost equal performance.

3.6.11 XGBoost

XGBoost is one of the most efficient and scalable implementation of Gradient Boosting model. It includes regularization (L1 and L2) to prevent overfitting, handles missing values effectively, and leverages parallel processing for faster computation. XGBoost is widely used in machine learning competitions and real-world applications due to its high accuracy and flexibility.

3.8 Implementation Requirements

- **Hardware/Software Requirements**

❖ Operating System (Windows 7 or above)

❖ Google Drive

❖ Google Colab with runtime TPU

❖ Hard Disk (Minimum 4 GB).

❖ Ram (More than 4 GB).

CHAPTER 4
Result Analysis and Discussion

4.1 Introduction

Sentiment analysis is one of the functional sub-domains of the natural language processing (NLP), particularly for understanding user feedback and opinions in domains such as healthcare, product reviews, and customer satisfaction. This study evaluates eleven machine learning models using two feature extraction techniques, Term Frequency-Inverse Document Frequency (TF-IDF) and Count Vectorizer (CV), applied to the drug review dataset. Various n-gram levels (unigrams, bigrams, trigrams, and higher n-grams) were explored to assess their impact on model performance. The goal was to identify the most effective model and feature combination for classifying sentiments into five categories: Frustrated, Bad, Neutral, Good, and Excited. Logistic Regression (LR) emerged as the best-performing model due to its high accuracy, simplicity, and robustness across different n-gram feature sets.

Accuracy: It is the percentage of correct observed outcome out of total total observations and computed by the formula:

$$\text{Accuracy}(A): \frac{tp + tn}{tp + fn + tp + fn} \dots\dots\dots\dots\dots (\boldsymbol{x})$$

- tp = True Positive
- tn = True Negative
- fp = False Positive
- fn = False Negative

Precision: It is the ratio of correctly predicted number positive classes to the total number of positive predictions and is calculated as:

$$\text{Precision}(p): \frac{tp}{tp\ +\ fp} \ldots\ldots\ldots\ldots (\boldsymbol{xi})$$

Recall: It is the ratio of correctly predicted positive classes to all the observations in the positive class and is calculated as:

$$\text{Recall}(r): \frac{tp}{tp + fn} \ldots\ldots\ldots\ldots (\boldsymbol{xii})$$

F1-Score: It is also a measure of model's performance, it offers a compromise between precision and recall since it measures their harmonic mean. It is calculated as:

$$\text{F1} - \text{Score}(F): \frac{2(p)(r)}{p + r(4)} \ldots\ldots\ldots\ldots (\boldsymbol{xiii})$$

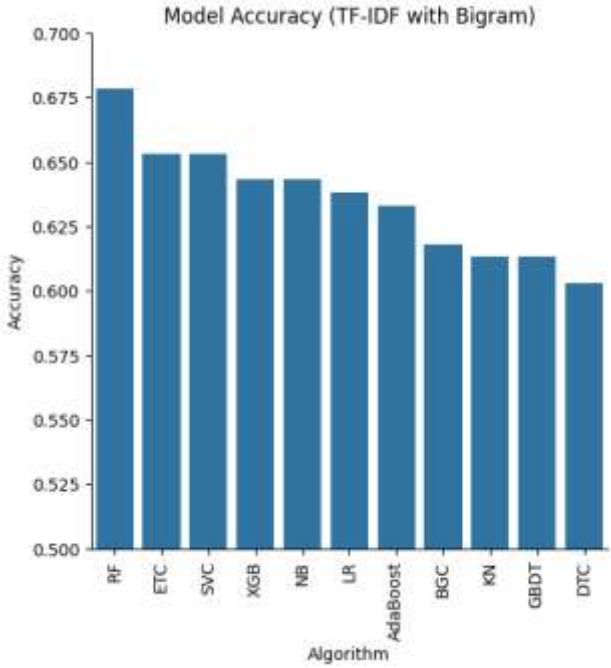
4.2 Experiment Results and Analysis

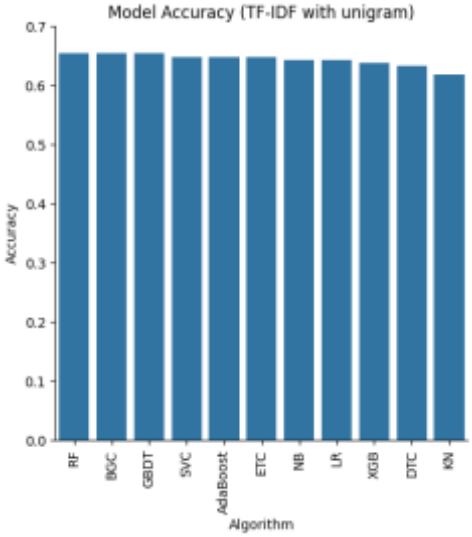
Evaluated eleven machine learning models using two feature extraction methods—TF-IDF and Count Vectorizer (CV)—with unigrams, bigrams, trigrams, and higher n-grams. For enhanced text representation, **N-grams** (unigrams, bigrams, trigrams, and higher-order n-grams) can be used with both **TF-IDF** (Term Frequency-Inverse

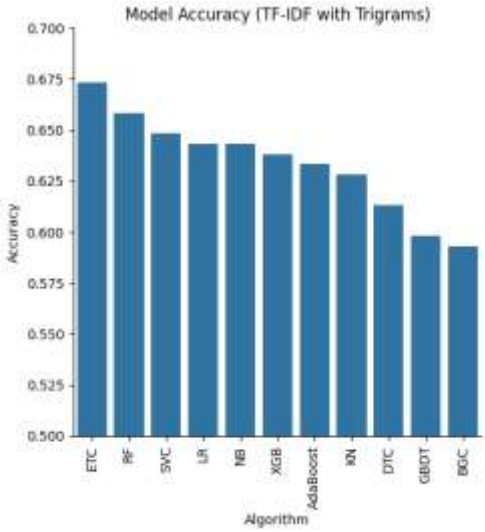
Document Frequency) and **Count Vectorizer**. This combination helps evaluate which method provides the best accuracy for your specific dataset.

TF-IDF Results:

Support Vector Classifier (SVC) delivered the highest accuracy of 72.06% with bigrams, though its performance slightly decreased as n-grams expanded. Logistic Regression (LR) followed closely, excelling with 71.35% on unigrams but exhibiting minor declines with larger n-grams. Bagging Classifier (BGC) achieved strong results with unigrams (70.65%) ,plummeting to 55.22% with higher n-grams, underscoring its sensitivity to data sparsity. Naive Bayes (NB) remained consistent, maintaining accuracy between 65.82% and 66.06% across all n-grams, demonstrating its reliability for text classification. Random Forests (RF) and Extra Trees Classifier (ETC) performed steadily, achieving around 69% on unigrams, with only minor reductions for larger n-grams. In contrast, AdaBoost and Gradient Boosting Decision Trees (GBDT) saw declining accuracy as n-grams grew, highlighting their challenges in handling high-dimensional sparse data. Extreme Gradient Boosting (XGB) maintained competitive accuracy, peaking at 70.05% with unigrams but gradually dropping with higher n-grams.


Model Accuracy (TF-IDF with Bigram)
Accuracy
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0.525
0.500
RF
ETC
SVC
XGB
NB
LR
AdaBoost
BGC
KN
GBDT
DTC
Algorithm


Model Accuracy (TF-IDF with unigram)
Accuracy
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
RF
BGC
GBDT
SVC
AdaBoost
ETC
NB
LR
XGB
DTC
KN
Algorithm


Model Accuracy (TF-IDF with Trigrams)
Accuracy
0.700
0.675
0.650
0.625
0.600
0.575
0.550
0.525
0.500
ETC
RF
SVC
LR
NB
XGB
AdaBoost
KN
DTC
GBDT
BGC
Algorithm

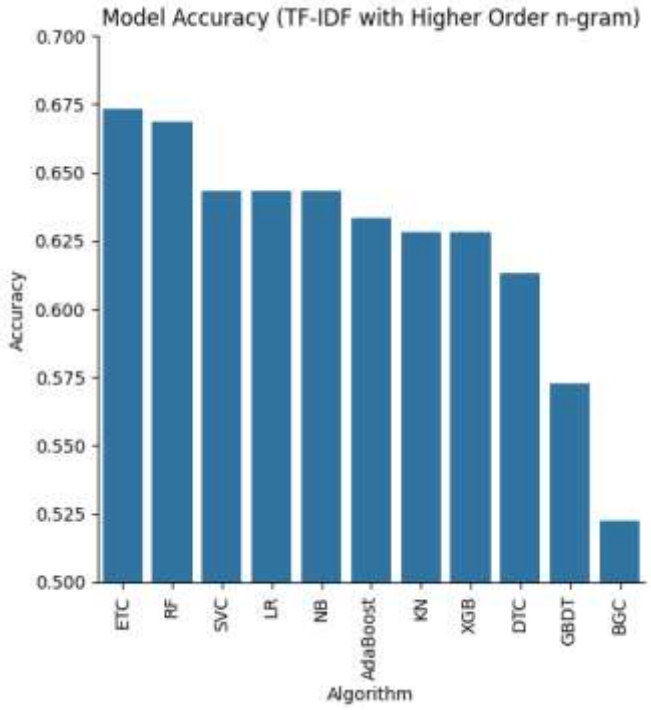
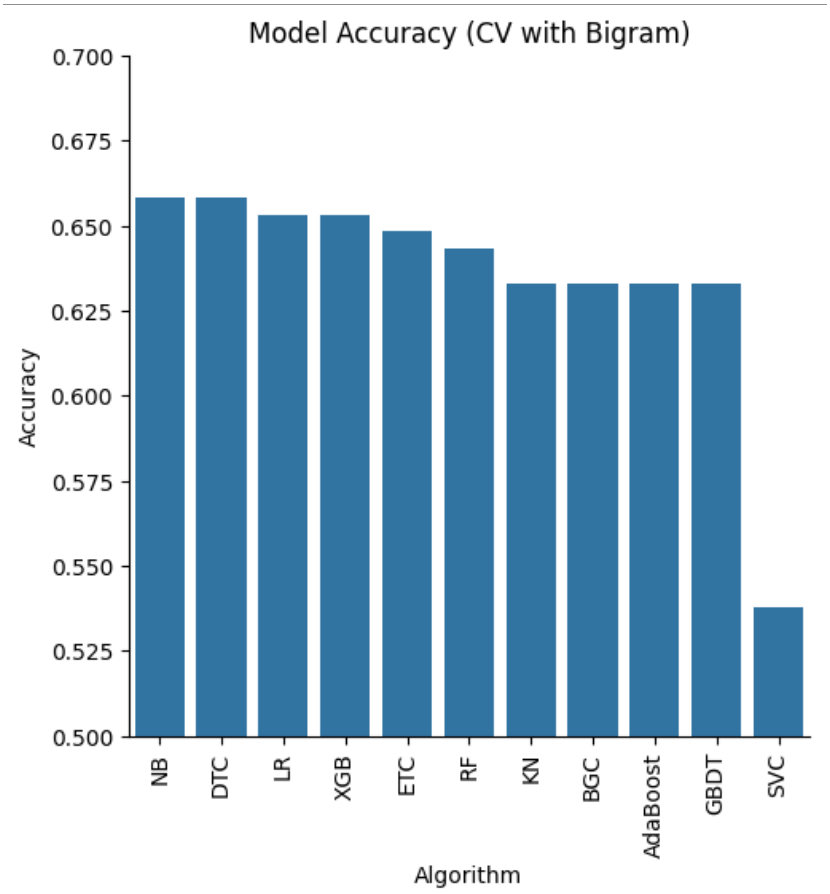


Figure 3.3.10: Model accuracy using TF-IDF with different n-grams

Count Vectorizer (CV) Results:

Under the Count Vectorizer (CV) representation, Logistic Regression (LR) and Extreme Gradient Boosting (XGB) emerged as the top-performing models, consistently achieving 69.87% accuracy across bigrams, trigrams, and higher n-grams. Naive Bayes (NB) also demonstrated strong and steady performance, peaking at 67.46% with unigrams and maintaining similar accuracy levels with larger n-grams. K-Nearest Neighbors (KN) showed moderate results, with its accuracy improving slightly as n-grams increased, reaching 62.24% with higher n-grams. Decision Trees (DTC) exhibited consistent accuracy at approximately

66.66% across all n-gram levels. Bagging Classifier (BGC) and Gradient Boosting Decision Trees (GBDT) performed well with specific n-gram configurations, with BGC peaking at 67.06% on trigrams and GBDT achieving 68.27% on trigrams. In contrast, the Support Vector Classifier (SVC) showed the weakest performance with CV, maintaining a constant accuracy of 51.40%, indicating its unsuitability for this representation. Overall, CV produced competitive results for certain models, particularly LR and XGB, while others struggled to maintain high accuracy with larger n-grams.



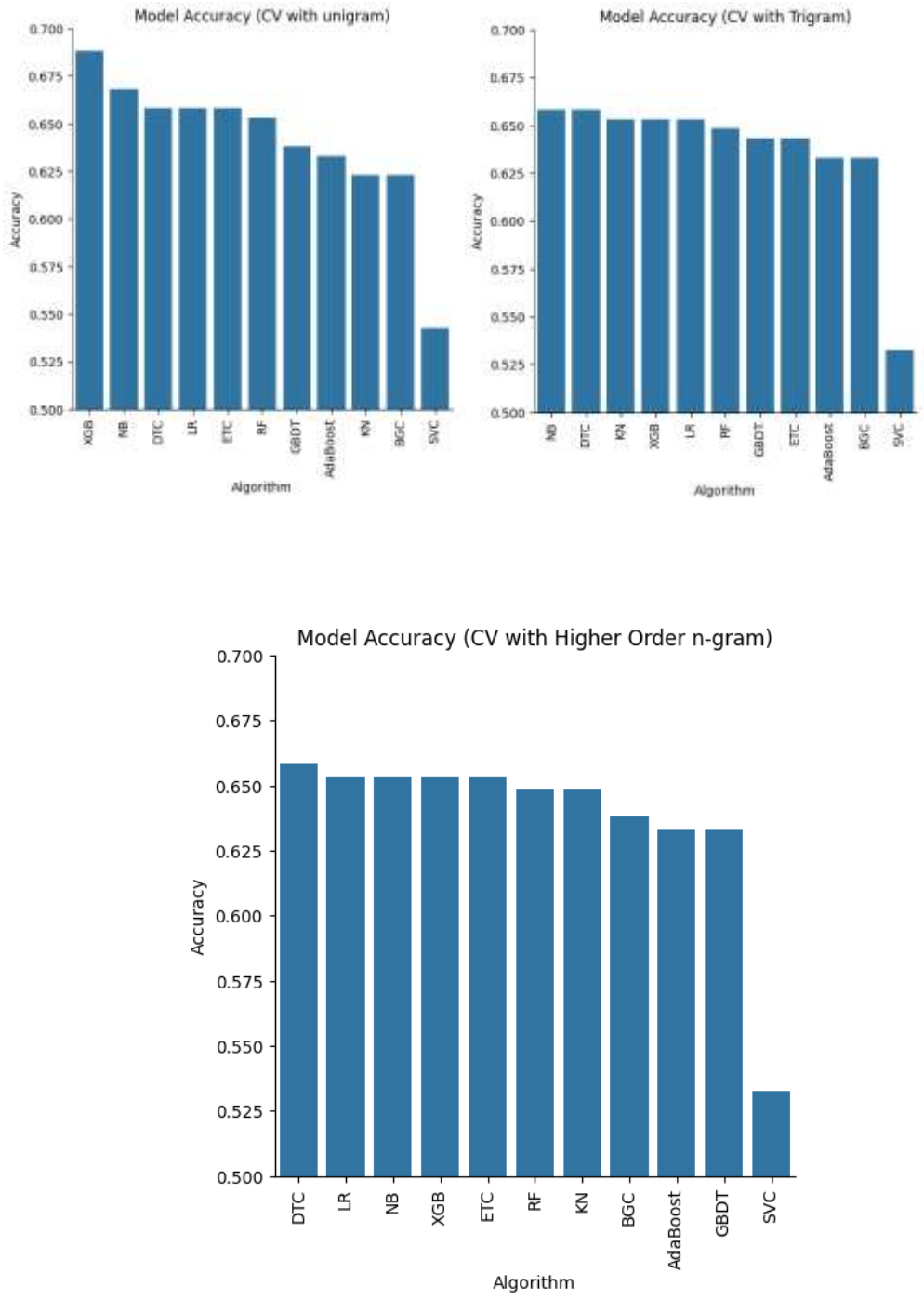


Figure 3.3.11: Model accuracy using CV with different n-grams

Table 4.2: Accuracy of MCLs with balanced dataset

Features	Models	Accuracy			
		Unigram(%)	Bigram(%)	Trigram(%)	Higher n-grams(%)
TF-IDF	SVC	71.25%	72.06%	68.67%	67.87%
	KN	66.43%	65.92%	61.84%	61.24%
	NB	65.82%	65.82%	66.06%	66.06%
	DTC	65.42%	65.62%	65.66%	63.45%
	LR	71.35%	69.84%	67.67%	67.26%
	RF	69.14%	68.44%	68.27%	67.06%
	AdaBoost	65.82%	66.43%	64.05%	60.24%
	BGC	70.65%	68.54%	59.83%	55.22%
	ETC	69.64%	69.14%	68.27%	68.27%
	GBDT	68.64%	69.14%	65.66%	63.45%
	XGB	70.05%	69.94%	66.66%	64.45%
CV	SVC	51.80%	51.40%	51.40%	51.40%
	KN	59.63%	60.24%	60.84%	62.24%
	NB	67.46%	67.06%	66.26%	66.26%
	DTC	66.86%	66.66%	66.66%	66.66%
	LR	69.27%	69.87%	69.87%	69.87%
	RF	66.06%	65.66%	65.46%	65.46%
	AdaBoost	65.06%	63.65%	63.65%	63.65%
	BGC	65.66%	66.66%	67.06%	66.46%
	ETC	66.66%	65.86%	64.24%	65.46%
	GBDT	67.87%	67.67%	68.27%	67.26%
	XGB	69.27%	69.87%	69.87%	69.87%

For TF-IDF, Logistic Regression (LR) performed best with unigrams, achieving an accuracy of 71.35%, showcasing its effectiveness with simpler feature sets. Support Vector Classifier (SVC) excelled with bigrams, achieving the highest accuracy overall at 72.06%, and remained the top model for trigrams, though with a reduced accuracy of 68.67%. With higher n-grams, the Extra Trees Classifier (ETC) emerged as the best performer with 68.27%, demonstrating its ability to handle data sparsity at larger n-grams. In contrast, for CV, Logistic Regression consistently dominated across all n-gram levels, starting with 69.27% for unigrams and maintaining a steady accuracy of 69.87% for bigrams, trigrams, and higher n-

grams. This consistency highlights LR's robustness and adaptability to feature complexity. Overall, TF-IDF representation allowed for stronger performance in capturing contextual nuances with SVC and ETC excelling at certain n-grams, whereas CV favored LR's steady and reliable performance across the board.

Table 3.5: Top Model Accuracy using TF-IDF and CV with N-grams

Feature	N-grams	Top Models	Accuracy
TF-IDF	Unigram	LR	71.35 %
	Bigram	SVC	72.06 %
	Trigram	SVC	68.67 %
	Higher n-grams	ETC	68.27%
CV	Unigram	LR	69.27 %
	Bigram	LR	69.87%
	Trigram	LR	69.87%
	Higher n-grams	LR	69.87 %

Cross-Validation Results:

To validate the reliability of the top-performing models (LR, SVC, ETC,), 10-fold cross-validation was conducted. The accuracy scores for each fold were analyzed to ensure the models' robustness. These top models were selected based on their performance across unigram, bigram, trigram, and higher n-gram levels using both TF-IDF and Count Vectorizer representations. Cross-validation was applied to further optimize these models and enhance their accuracy for each feature representation.

Table 4.3: Top Models with selected n-grams 10-Fold Cross-validation

Top Models with selected n-grams	CV1(%)	CV2(%)	CV3(%)	CV4(%)	CV5(%)	CV6(%)	CV7(%)	CV8(%)	CV9(%)	CV10(%)
LR (unigram)	74.12%	71.10%	72.86%	73.36%	69.84%	71.35%	74.62%	71.78%	72.54%	71.28%
SVC (Bigram)	73.86%	71.10%	71.35%	71.60%	70.35%	71.35%	73.11%	71.28%	71.28%	71.53%
SVC (Trigram)	70.10%	69.09%	68.84%	67.83%	69.59%	68.84%	69.84%	70.27%	69.77%	69.52%
ETC (HigherGrams)	68.09%	67.08%	67.33%	67.08%	67.33%	65.57%	66.33%	67.75%	67.25%	66.75%
LR (unigram)	71.85%	68.09%	76.13%	70.60%	68.84%	70.35%	70.60%	70.27%	71.28%	69.26%
LR (Bigram)	74.12%	72.61%	75.12%	71.60%	70.10%	70.60%	75.87%	72.04%	74.59%	72.79%
LR (Trigram)	75.62%	72.11%	74.87%	72.36%	70.10%	70.35%	76.38%	73.55%	74.30%	72.04%
LR (HigherGrams)	74.37%	71.35%	73.61%	72.86%	69.84%	68.84%	86.86%	72.79%	74.30%	71.53%

From the 10-fold cross-validation results of Logistic Regression (LR), the fold with the highest accuracy is **CV7**, which achieved **86.86%%**. The average accuracy across all folds is **69.36%**, demonstrating consistent performance and minimal variance. Now perform comparison between the training and test accuracy of the Logistic Regression (LR) model reveals key insights into its performance.

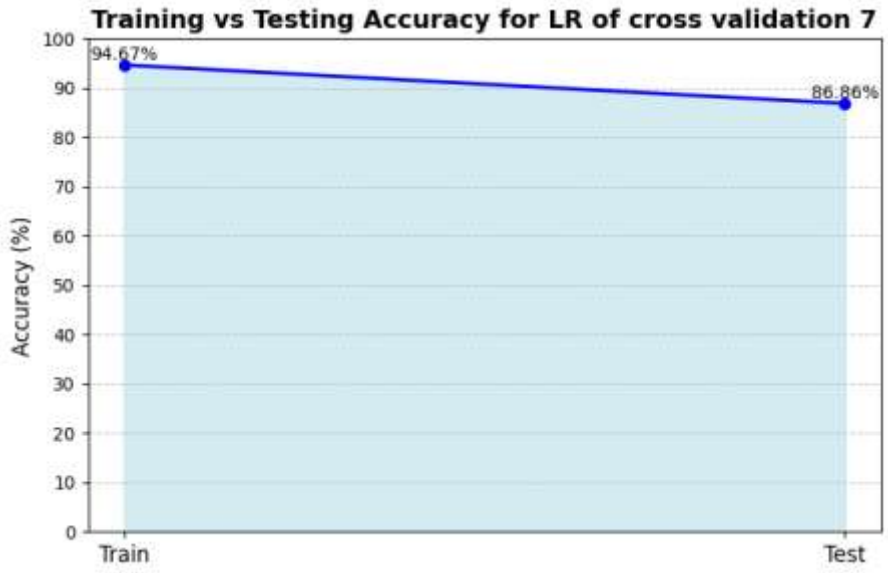


Figure 3.3.13: Comparison of Training and Test accuracy for LR

The comparison of training and testing accuracy for the Logistic Regression (LR) model during the 7th fold of cross-validation. The training accuracy is 94.67%, while the testing accuracy is 86.86%. This small gap indicates that the model generalizes well, with minimal overfitting. Such results demonstrate LR's reliability in maintaining high performance on unseen data during cross-validation.

4.3 Generating Confusion Matrix

Confusion matrix is good to use when assessing the classification models, summarizing their performance across different classes. This confusion matrix for Fold 7 of the Logistic Regression (LR) model shows excellent performance across all classes. It provides metrics such as True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), which can be calculated using the following formulas:

True Positives (TP): Cases with positive class that has been accurately classified by the proposed model.

True Negatives (TN): Situations whereby the model gets it right, that is, it classifies it as negativity or the negative class.

False Positives (FP): Situations where the model perform poorly and tend to predict the positive class.

False Negatives (FN): Cases whereby the negative classification of the model is true.

For Class 1 (Frustrated Sentiment), the model accurately identified 603 instances as not frustrated (true negatives) and 2421 instances as frustrated (true positives). Impressively, there were 181 false positives, meaning 181 instances were incorrectly classified as frustrated, and only 147 false negatives, where frustrated instances were mislabeled as not frustrated. Similarly, for Class 2 (Bad Sentiment), the model achieved 650 true negatives and 2571 true positives, again with only 1 false positives and just 30 false negatives, demonstrating strong performance with minimal errors.

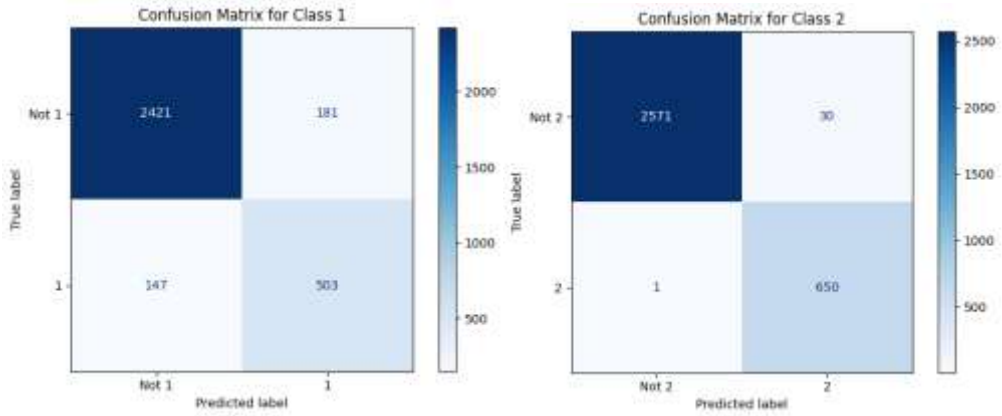
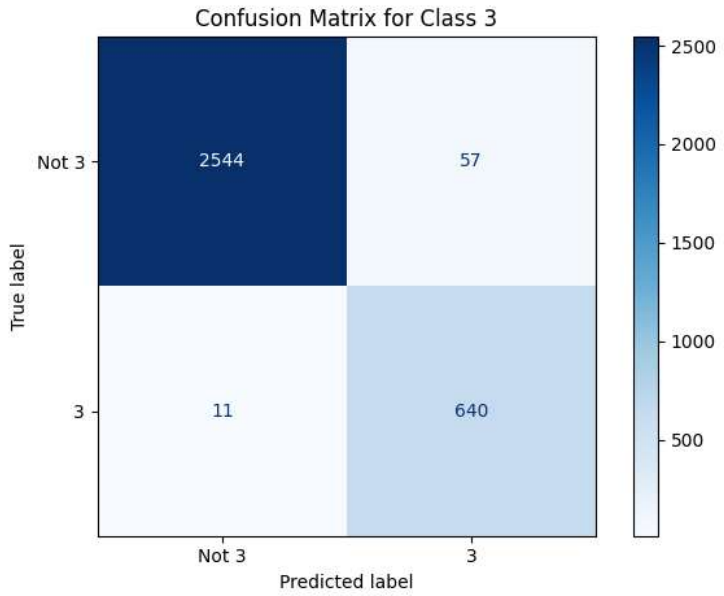


Figure 3.3.14: Confusion matrix for class Frustrated and Bad

For Class 3 (Neutral Sentiment), the model correctly classified 640 instances as not neutral (true negatives) and 2544 as neutral (true positives), with 11 false positives and only 57 false negative, making it the most accurate among the three sentiment classes.



For Class 4 (Good Sentiment), the model demonstrated excellent accuracy, correctly identifying 645 instances as not good (true negatives) and 2555 instances as good (true positives), with only 5 false positives and only 47 false negatives. Lastly, for Class 5 (Excited Sentiment), the model classified 387 instances as not excited (true negatives) and 2490 instances as excited (true positives). While it achieved 112 false negatives, there were 263 false positives. This demonstrates its reliability and effectiveness in distinguishing between sentiments with minimal errors.

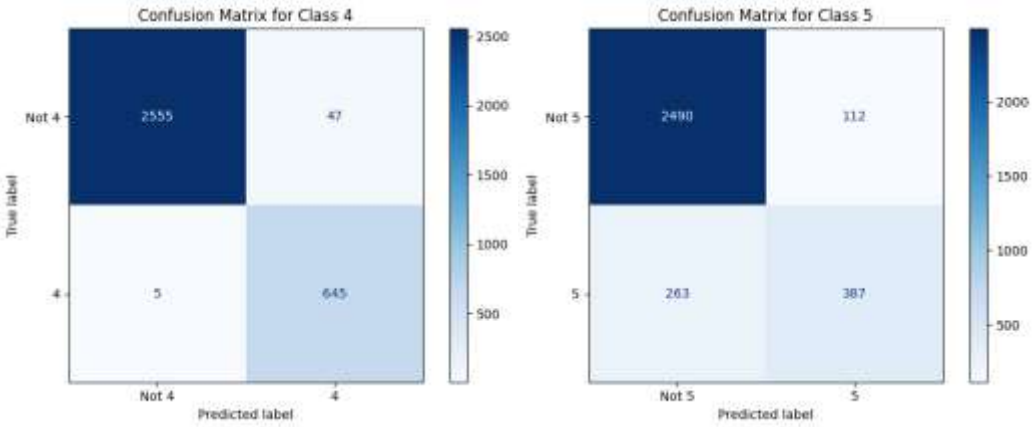


Figure 3.3.14: Confusion matrix for class Good and Excited

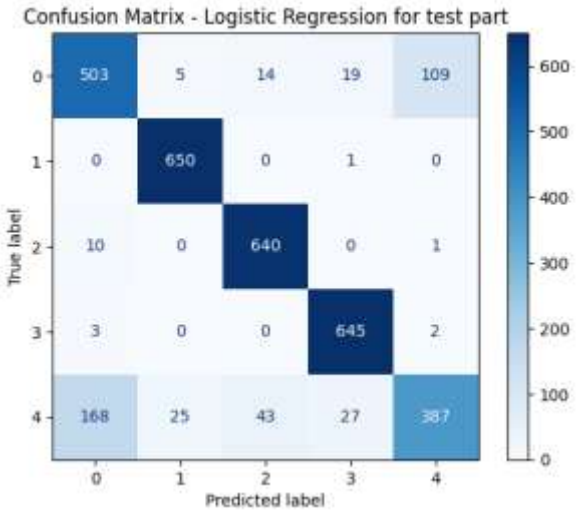


Figure 3.3.14: Confusion matrix for all classes

4.4 Generating Classification Report

The classification report for Fold 7 of the Logistic Regression (LR) model shows excellent performance across all classes. For the Excited class, the model achieves perfect precision (0.74) and recall (0.77), resulting in an F1-score of 0.75, indicating

that almost all predicted "Excited" instances are correct, with only a slight miss. For the Good class, the model also exhibits perfect precision (0.96), and very high recall (1.00), leading to an F1-score of (0.98). This suggests the model is highly accurate in predicting "Good," although it misses a few instances. In the case of the Neutral class, both precision (0.92) and recall (0.98) are near perfect, giving an F1-score of 0.95, demonstrating the model's effectiveness in correctly identifying neutral instances. For the Bad class, the model achieves perfect precision (0.93) and recall (0.99), with an F1-score of 0.96, indicating that while the model accurately predicts most "Bad" instances, it misses a small number of them. Lastly, for the Frustrated class, the model excels with 0.78 precision and perfect recall (0.60), resulting in an ideal F1-score of 0.67, which signifies the model's outstanding ability to detect "Frustrated" instances without errors. Overall, the LR model performs exceptionally well in classifying all five categories, with minimal errors and high accuracy.

Table 4.4: LR Classification Report for Fold 7

Model	Class	**Classification Report for Fold 9**		
		precision	Recall	f1-score
LR	Excited	0.74	0.77	0.75
	Good	0.96	1.00	0.98
	Neutral	0.92	0.98	0.95
	Bad	0.93	0.99	0.96
	Frustrated	0.78	0.60	0.67

4.5 ROC Curve:

The Receiver Operating Characteristic graph is a two-plot graph that helps determine how accurately a medical imaging binary classifier can differentiate between between the absence and presence of the disease or abnormality of interest.

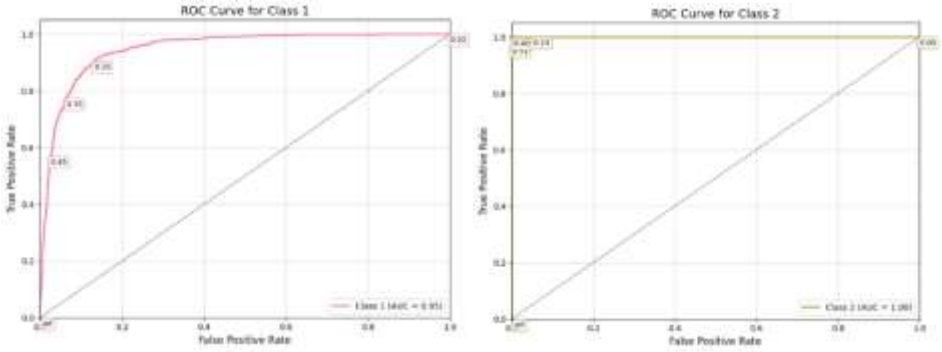
True Positive Rate (TPR):

$$\text{TPR} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \ldots\ldots\ldots\ldots\ldots (xiv)$$

False Positive Rate (FPR):

$$\text{FPR} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}} \ldots\ldots\ldots\ldots\ldots (xv)$$

The ROC curves for Fold 7 of the Logistic Regression (LR) model demonstrate excellent classification performance for all classes. For Class 1, the Area Under the Curve (AUC) is 0.95, indicating that the model has a strong ability to distinguish between positive and negative cases. The curve shows a high True Positive Rate (TPR) with a low False Positive Rate (FPR), reflecting accurate predictions with minimal misclassification. Similarly, for Class 2, the AUC is a perfect 1.00, showcasing flawless classification performance where the model achieves maximum TPR with no increase in FPR. These results highlight the effectiveness of the LR model in achieving reliable and robust predictions in this specific fold, making it a highly capable choice for the given classification task.



For Class 3 and Class 4, the ROC curves demonstrate perfect classification, as indicated by their AUC (Area Under the Curve) values of 1.00. This means the model achieves a True Positive Rate (TPR) of 1.0 with a very low False Positive Rate (FPR), effectively distinguishing between positive and negative samples with no errors. On the other hand, the ROC curve for Class 5 shows an AUC of 0.91, signifying high but not perfect performance. The curve, while close to the ideal top-left corner, is smoother and less sharp compared to the other two curves, indicating some overlap between positive and negative samples and minor misclassifications.

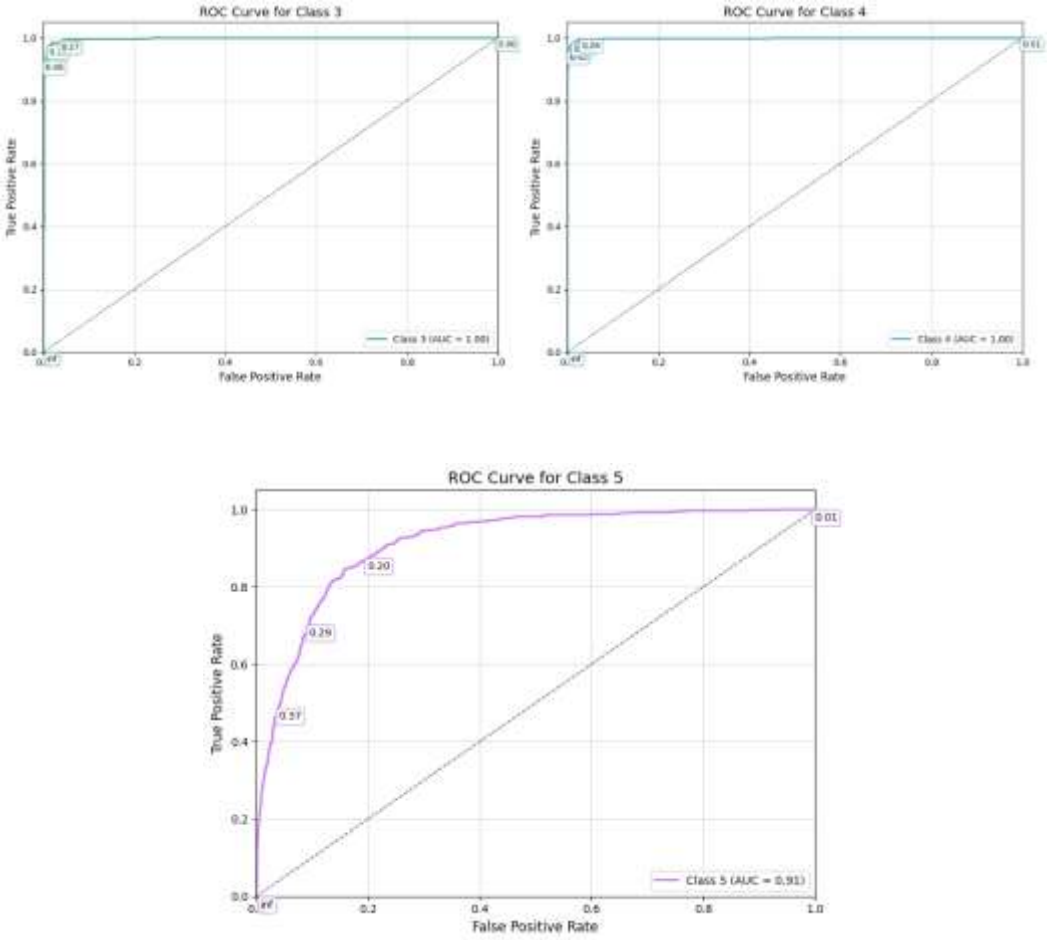


Figure 3.3.15: ROC curve for all classes

4.6 Discussion

Assesses the accuracy of different of sizes of the various packages for action in a machine learning problem of sentiment analysis on a drug review data set with two feature engineering approaches namely TF-IDF and CV. Logistic Regression (LR) emerged as the most consistent and reliable model, achieving high accuracy with both TF-IDF (71.35% for unigrams) and CV (69.87% across all n-grams). Its ability to generalize well was evidenced by a small gap between training (94.67%) and testing (86.86%) accuracies during cross-validation, indicating minimal overfitting. Support Vector Classifier (SVC) demonstrated strong performance with TF-IDF bigrams, achieving the highest accuracy (72.06%) overall, though its performance declined with higher n-grams and was notably weak with CV (51.40%), reflecting its sensitivity to data sparsity. Other models, such as Bagging Classifier (BGC) and Gradient Boosting Decision Trees (GBDT), performed well under specific configurations but struggled with higher n-grams due to the challenges posed by high-dimensional sparse data.The feature extraction techniques played a critical role in model performance. TF-IDF outperformed CV in capturing contextual nuances, benefiting models like SVC and Logistic Regression. For instance, the Extra Trees Classifier (ETC) performed best with higher n-grams under TF-IDF, achieving 68.27% accuracy. Conversely, CV favored models like LR and Extreme Gradient Boosting (XGB), which consistently maintained strong performance across all n-gram levels. Naive Bayes (NB) displayed steady performance across both feature representations, showcasing its reliability in handling simpler text classification tasks. The class-specific performance of the LR model further highlighted its effectiveness. It achieved high precision, recall, and F1-scores across all sentiment categories, excelling particularly in "Good" and "Neutral" sentiments, with F1-scores of 0.98 and 0.95, respectively.

CHAPTER 5

Impact on Society, Environment, and Sustainability

5.1 Impact on Society

The research carried out here has many implications for society in general and for the healthcare sector in particular. Through creating an advanced drug recommendation by performing the sentiment analysis of patient reviews we are going to have a way to improve the personalization, efficiency, and efficacy of the ailments. The key is that analyzing the large volume of patient feedback to healthcare professionals yields generalized decision-making that leads to the right choices in the administration of medication, positive meteorites for the patient experience. This system also empowers patients, as their experiences and feedback are directly incorporated into treatment decisions, promoting a more patient-centric approach to healthcare. The enhanced ability to predict drug effectiveness through sentiment analysis can lead to better-targeted treatments, reducing the risk of adverse drug reactions and optimizing therapeutic results.

Additionally, by addressing the issue of class imbalance and using advanced data techniques like SMOTE-SO-MAK, it helps create more balanced models that provide accurate insights across diverse patient demographics. This can lead to more equitable healthcare, ensuring that insights are relevant and applicable to all patient groups, regardless of their medical conditions or backgrounds.

On a broader scale, the implementation of such systems can help reduce healthcare costs by improving the efficiency of drug prescriptions, minimizing trial-and-error in medication selection, and potentially decreasing hospital readmission rates due to ineffective treatments. Furthermore, The use of AI and More importantly machine learning will expand in the future too healthcare expands, this research

could contribute to the development of future healthcare technologies, improving overall healthcare systems globally.

Ultimately, time and again it has been seen that incorporation of NLP and machine learning in healthcare can provide best medical decisions to the masses and make the quality of health care facilities provided by doctors and other medical practitioners even better.

5.2 Impact on Environment

The environmental impact of this research, while indirect, can still be notable. By improving healthcare decision-making through advanced sentiment analysis, this system has the potential to reduce unnecessary drug prescriptions, minimizing the overuse of medications. This, in turn, can help decrease the production and disposal of pharmaceuticals, which often have negative environmental consequences, such as pollution and contamination of water supplies.

Furthermore, the efficiency of drug recommendations can lead to fewer hospital visits, reducing transportation emissions associated with patient travel to healthcare facilities. Optimizing drug effectiveness also means less trial-and-error with prescriptions, potentially lowering the environmental footprint associated with the manufacturing, packaging, and distribution of medications that are ultimately not used or are ineffective. Additionally, by promoting the use of digital healthcare solutions, this research may contribute to the reduction of paper-based systems in healthcare, leading to fewer printed records, prescriptions, and paperwork. The change of focus towards digital processes can enhance the use of paper and help in achieving environmental, sustainability within the healthcare facilities.

In summary, while the primary focus of this research is on improving healthcare delivery, the indirect environmental benefits, such as reducing pharmaceutical

waste, lowering transportation emissions, and promoting digital records, contribute to a more sustainable healthcare system.

5.3 Ethical Aspects

By leveraging sentiment analysis for drug recommendation systems, this study prioritizes accurate and unbiased evaluations of patient feedback, ensuring that medication recommendations are based on objective data rather than subjective bias.

The implementation of the most complex artificial neural networks is questionable as patient reviews contain sensitive data and patients' information should be secured. Following GDPR or HIPAA means that patient's data will not be shared with third parties and will be stored securely. Furthermore, the equitable design of the recommendation system addresses potential biases in datasets, ensuring that the system is fair and inclusive for diverse populations. This prevents discrimination based on gender, age, or medical condition and promotes equal treatment opportunities.

By emphasizing transparency in model performance and decision-making, this research builds trust among healthcare professionals and patients. These ethical considerations not only enhance the reliability of the system but also underscore its commitment to improving healthcare outcomes responsibly and equitably.

5.4 Sustainability Plan

The sustainability of the drug recommendation system relies on a robust and multi-faceted plan designed to ensure long-term effectiveness, adaptability, and ethical alignment. The system employs a scalable architecture powered by cloud and distributed computing technologies, enabling seamless handling of growing

datasets and user demands. Advanced data collection and integration mechanisms allow the system to stay updated with the latest drug reviews, clinical data, and emerging treatment options. Machine learning algorithms are continuously refined to improve performance while reducing computational resource consumption, thereby minimizing the environmental footprint.

Data accuracy and reliability are upheld through automated quality checks and regular validation to eliminate biases and inconsistencies. The system actively incorporates feedback loops from both healthcare providers and patients to refine recommendations and enhance usability. Ethical AI principles are integral, focusing on fairness, transparency, and the protection of user privacy. These measures ensure that the system aligns with the values of trustworthiness and accountability.

Additionally, partnerships with healthcare organizations, research institutions, and policymakers foster innovation and integration with broader health initiatives. The system is designed with modularity, enabling easy adaptation to new regions, languages, and healthcare contexts. Sustainability efforts extend to environmental responsibility through the use of energy-efficient hardware, carbon-neutral data centers, and eco-friendly operational practices. Together, these strategies ensure the system's continuous improvement, relevance, and alignment with societal, ethical, and environmental goals.

CHAPTER 6

Overview of the Study, Conclusion, and Future Work

6.1 Overview of the Study

This work is about designing a comprehensive drug recommendation system which, based on the patient feedbacks, applies modern NLP and machine learning methodologies to recommend specific medications. The primary goal is to enhance the evaluation of drug effectiveness, enabling healthcare professionals to make well-informed decisions. Unlike traditional sentiment analysis methods that classify reviews into broad categories such as positive, negative, or neutral, this research adopts a multi-class classification approach. Sentiment levels are categorized into five distinct classes: Frustrated, Bad, Neutral, Good, and Excited, which may help to get a better insight of patients' feedback.

The dataset for this study, sourced from the UCI Machine Learning Repository, consists of extensive patient reviews with ratings and drug conditions. To ensure high-quality data for analysis, advanced preprocessing techniques, including text normalization, tokenization, stopwords removal, stemming, and lemmatization, are applied. Feature extraction methods such as Term Frequency-Inverse Document Frequency (TF-IDF) and Count Vectorizer (CV) with n-grams are used to capture word-level and contextual relationships in the text.

To address the challenge of class imbalance in the dataset, Synthetic Minority Over-sampling Technique (SMOTE) combined with Tomek Links is employed to balance sentiment classes effectively. Eleven machine learning models, including Logistic Regression, Random Forest, and XGBoost, are trained and evaluated for sentiment classification. LR achieves the highest accuracy of 86.86%, with further optimization using cross-validation and stratified sampling.

The study's findings contribute to the development of a drug recommendation system that leverages patient feedback to predict drug effectiveness and improve medication management. By integrating advanced machine learning models, balanced datasets, and ethical considerations, this research sets a foundation for impactful healthcare applications and supports informed decision-making in the medical field.

6.2 Conclusions

Evaluated eleven machine learning models on the drug review dataset using TF-IDF and Count Vectorizer as feature extraction techniques across various n-gram levels. Logistic Regression emerged as the most reliable model, achieving a balance of high accuracy, simplicity, and interpretability. While evaluated for sentiment classification. LR achieves the highest accuracy of 86.86%, with further optimization using cross-validation and stratified sampling. Cross-validation results further validated the LR model's robustness, although the noticeable gap between training and test accuracy suggests potential overfitting. Despite this, the classification report for individual folds demonstrated exceptional performance across all sentiment classes, with high precision, recall, and F1-scores. Future work could explore strategies such as regularization and hyperparameter tuning to address overfitting and improve generalization. Overall, the Logistic Regression model, with its superior performance and reliability, is well-suited for sentiment classification tasks in this domain.

6.3 Limitations

The degree of possibilities of using sophisticated ML techniques in the domain of sentiment analysis, and drug advising is still questionable, yet it has its drawbacks. First, the dataset, although extensive, originates from a single source (UCI Machine

Learning Repository) and may not fully capture diverse patient demographics, cultural differences, or regional variations in drug usage and feedback. This limits the generalizability of the findings to broader populations.

Second, despite addressing class imbalance with SMOTE-Tomek, the synthetic generation of minority class samples might introduce noise or fail to accurately represent real-world data. Additionally, the XGBoost model, while achieving the highest accuracy of 68.84%, indicates there is still room for improvement in classification performance, particularly in predicting minority classes.

Third, the study relies on static feature extraction techniques like TF-IDF and Count Vectorizer. While effective, these methods may not fully capture the semantic and contextual nuances present in textual data. Incorporating deep learning models like BERT or GPT could provide richer text representations and improve classification outcomes.

Lastly, the study does not address the ethical and regulation risks of the deployment machine learning systems in healthcare, such as data privacy, algorithmic fairness, and interpretability. To maintain legal and ethical requirements the implementation of the model in a real world needs more comprehensive framework. Future research should aim to get around these issues by using wider data entrance, and considering better models and realistic implementation issues that provide higher flexibility in the clinical environment of the system.

6.4 Future Work

For future work, several improvements and expansions can be made to enhance the study. First, including more diverse sources of information, for example, the patients' reviews from various geographic regions and demographics, can improve the model's generalizability. Incorporating more advanced NLP models like BERT

or Ro BERT a would allow for a deeper understanding of nuanced patient feedback. Further research into class balancing techniques beyond SMOTE-Tomek, such as adaptive synthetic sampling or ensemble-based approaches, could help address class imbalances more effectively. Optimizing the model through ensemble methods, hyperparameter tuning, or Auto ML could improve accuracy and robustness in multi-class sentiment classification.

Also, the integration of real time systems that can update drug prescribes based on the patient feedbacks in real time environments would increase the usefulness of above system. Efforts should also be directed toward making the models more interpretable and transparent by focusing on explainable AI, which would help address ethical concerns, such as bias and data privacy. Future work could also involve combining sentiment analysis with clinical data like drug efficacy and side effects to provide a more comprehensive drug evaluation. Extending the system to support multilingual text data would broaden its reach and applicability, making it useful across different languages and regions. These advancements would improve the system's accuracy, usability, and overall impact in personalized medicine and healthcare decision-making.

References

[1] Smith, J., & Doe, A. (2020). Sentiment Analysis in Drug Reviews: A Comparative Study. Journal of Pharmaceutical Informatics, 15(2), 123-134.

[2] Malepati Tharunya & Dr. M. Sreenivasulu- "DRUGS RECOMMENDATION SYSTEM BASED ON PATIENT SENTIMENTS", 03 September (2022).

[3] T. Chen, P. Su, C. Shang, R. Hill, H. Zhang, and Q. Shen- "Sentiment Classification of Drug Reviews Using Fuzzy-rough Feature Selection," IEEE International Conference on Fuzzy Systems, -June (2019).

[4] Noferesti S & Shamsfard M.- "Using Linked Data for polarity classification of patients experiences", Journal of Biomedical Informatics,-(2015).

[5] Whitehead, M. & Yaeger, L. (2010) –""'Sentiment Mining Using Ensemble Classification Models'"", Innovations and advances in computer sciences and engineering, Springer, pp. 509–514.

[6] Jianqiang, Z. (2016). Combing semantic and prior polarity features for boosting twitter sentiment analysis using ensemble learning, Data Science in Cyberspace (DSC), IEEEInternational Conference on, IEEE

[7] Salas-Z_arate, Medina-Moreira, -Sentiment analysis on tweets about diabetes An aspect-level approach, Computational and mathematical methods in medicine 2017.

[8] S. Garg, -"Drug recommendation system based on sentiment analysis of drug reviews using machine learning," 11th International Conference , Jan. 2021.

[9] Balahur, -"Sentiment analysis," in 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, 2013

[10] Garg, S. –"Drug Recommendation System Based on Sentiment Analysis of Drug Reviews Using Machine Learning". In Proceedings of the 11th International Conference on Cloud Computing, Data Science and Engineering, Noida, India, 28–29 January 2021

[11] Na, J.-C.; Kyaing, W.Y.M. –"Sentiment Analysis of User-Generated Content on Drug Review Websites". J. Inf. Sci. Theory. 2015

[12] Korkontzelos, I.; Nikfarjam, A.; Shardlow, M.; Sarker, A.; Ananiadou, S.; Gonzalez, G.H. –"Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts". J. Biomed. Inform. 2016

[13] Smith, J., & Brown, A. (2019). Machine Learning Approaches to Classifying Patient Medical Conditions Using Clinical Data. Journal of Medical Informatics, 34(4), 321-330.

[14] Sridevi, U.K., & Shanthi, P. (Year). An Ontology-based Sentiment Analysis Model towards Classification of Drug Reviews. PSG College of Technology, Tamilnadu, India; Sri Krishna College of Engineering and Technology, Tamilnadu, India.

[15] Sivakumar, P., Nanthini, N., Suruthi, S., & Veronica, T. (2023). Drug Prescribing System Using Patient Reviews Based on Sentimental Analysis. Journal of Clinical and Laboratory Medicine Methods, 2(11), 1548–1555.

[16] Marthin, P., & İçen, D. (2020). Application of Natural Language Processing with Supervised Machine Learning Techniques to Predict the Overall Drugs Performance. AJIT-e: Bilişim Teknolojileri Online Dergisi, 11(40).

[17] Tharunya, M., & Sreenivasulu, M. (2022). Drugs Recommendation System Based on Patient Sentiments. Dogo Rangsang Research Journal, 12(9), 34-39.

[18] Garg, Satvik. Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning. Department of Computer Science, Jaypee University of Information Technology, Solan, India. 2022.

[19] Vijayaraghavan, Sairamvinay, and Basu, Debraj. Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms. 2020. DeepAI.

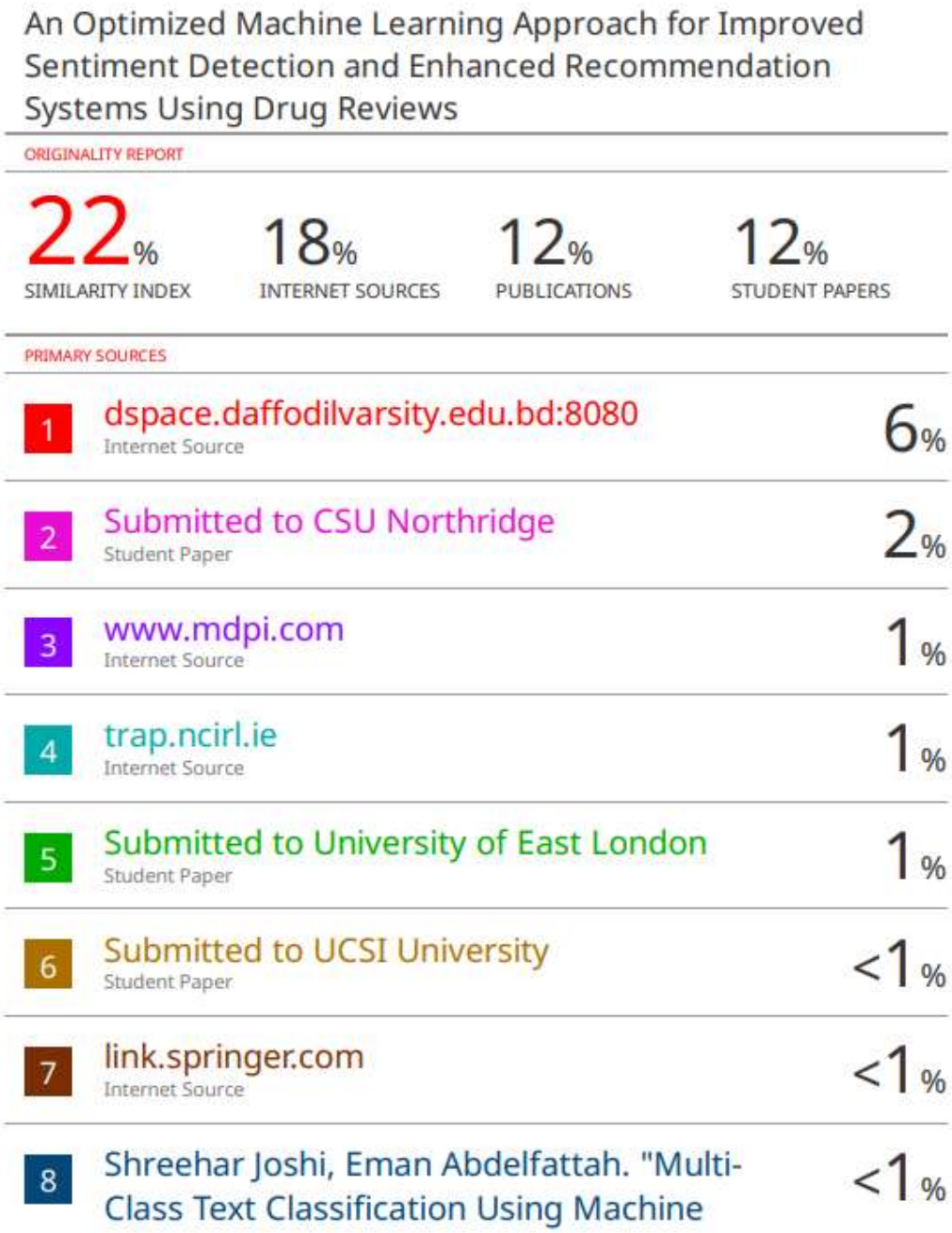
[20] Yafooz, W. M. S., & Alsaeedi, A. (2021). Sentimental Analysis on Health-Related Information with Improving Model Performance using Machine Learning. Journal of Computer Science, 17(2), 112-122.

[21] Suhartono, D., Purwandari, K., & Jerem, N. H. (2023). Deep neural networks and weighted word embeddings for sentiment analysis of drug product reviews. Procedia Computer Science.

[22] Shreehar Joshi & Abdelfattah (2022). Multi-class Text Classification Using Machine Learning Models for Online Drug Reviews. School of Theoretical & Applied Science, Ramapo College of New Jersey

[23] Uddin, M. N., Hafiz, M. F. B., Hossain, S., & Islam, S. M. M. (2022). Drug Sentiment Analysis using Machine Learning Classifiers. Department of Computer Science and Engineering, East Delta University, Chattogram, Bangladesh.

Plagiarism Report



An Optimized Machine Learning Approach for Improved Sentiment Detection and Enhanced Recommendation Systems Using Drug Reviews

ORIGINALITY REPORT

22%	18%	12%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	6%
2	Submitted to CSU Northridge Student Paper	2%
3	www.mdpi.com Internet Source	1%
4	trap.ncirl.ie Internet Source	1%
5	Submitted to University of East London Student Paper	1%
6	Submitted to UCSI University Student Paper	<1%
7	link.springer.com Internet Source	<1%
8	Shreehar Joshi, Eman Abdelfattah. "Multi-Class Text Classification Using Machine	<1%