UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE POLITICHE, ECONOMICHE E SOCIALI

Master degree in Data Science and Economics

Project Text Mining and Sentimental Analysis

Title: Sentimental analysis of Patient Reviews to Predict Drug Effectiveness Using Machine Learning

Prepared by: Samuel Tsegai

Submitted To: Prof. Ferari

Academic Year: 2023/2024

Date: sep 04, 2023

Abstract

In this project, we apply Text Mining and Sentiment Analysis techniques on a dataset from kaggle provided by a recently acquired pharmaceutical startup. The dataset contains information about various drugs, including their names, customer reviews, popularity, and use cases. The report elaborates on various stages of the project including data preprocessing, text mining, sentiment analysis, model development, evaluation, testing, and concludes with insights for future work.

In summary, This report presents the results of an analysis on a set of drug reviews data, utilizing a combination of Natural Language Processing, Machine Learning ,i.e Unsupervised and supervised Learning techniques.The models that employed: – K-Means ,Random Forest Classifier and Logistic Regression - delivered commendable results. Among these, the Random Forest Classifier emerged with the highest overall accuracy and AUC-ROC score, suggesting it might be the most suitable model for predicting drug effectiveness based on this specific analysis of patient reviews. Nevertheless, the efficacy of each model can vary depending on the particularities of the dataset and problem scenario, highlighting the necessity to contemplate various models during a machine learning investigation.

The models succeeded in classifying the new unseen reviews accurately based on their content, underscoring the models' robustness and the effectiveness of the deployed text mining and sentiment analysis techniques

Contents

1 Introduction

This report presents an exhaustive data analysis,text mining and sentiment analysis with machine learning model development process carried out on a dataset containing drug reviews. The primary objective was to build models capable of predicting the effectiveness of a drug based on its review In the initial phase, rigorous data preprocessing was performed to preserve the data's quality for subsequent analysis. This stage involved text normalization and the removal of non-essential elements like stop words, punctuation, and special characters. Further, a Vader lexicon-based sentiment analysis was employed to yield sentiment scores for each review, enhancing the dataset with an additional feature.

The analysis employed two machine learning models - Logistic Regression and Random Forest - combined with a TfidfVectorizer and review length . This amalgamation transformed the sanitized text data into a matrix of TF-IDF features with bigram feature to capture contextual meaning, prepping it for model training and evaluation.

We used a range of metrics to evaluate the models, including precision, recall, F1-score, and the area under the Receiver Operating Characteristic (ROC) curve (AUC-ROC). The outcomes revealed all models performing significantly well, with Random Forest marginally performing better , demonstrating an average accuracy of 89%, and an AUC-ROC of 96%.and logistic Regression model echoed promising results with accuracies of 86% an AUC-ROC of 93%.

In the concluding phase, the trained models were put to the test on unseen reviews. The models succeeded in classifying the reviews accurately based on their content, underscoring the models' robustness and the effectiveness of the deployed text mining and sentiment analysis techniques.

In summation, this study establishes that text mining and sentiment analysis, in conjunction with machine learning models, can efficaciously predict drug effectiveness. The insights from this investigation can act as a foundation for future research, exploring more advanced text feature extraction techniques or complex models like deep learning algorithms to augment predictive accuracy.

2 Objectives

Primary objectives for this study are as follows:

- To devise an advanced predictive model utilizing Natural Language Processing (NLP) and Machine Learning (ML) techniques that can efficiently understand and categorize textual data.
- To ascertain the most effective drugs for each medical condition and patients review.
- To unravel concealed trends or patterns that could aid in data-guided decision making.

3 Data Loading :Descriptive statistics

The dataset is taken from Kaggle used for this analysis, consisting of 161,297 entries, each containing 7 attributes. These attributes include a unique identifier for each entry, the drug name, the condition it is used to treat, the review given by a user, the rating assigned to the drug, the date of the review, and a count indicating how many people found the review useful.

4 Data Exploaration

The dataset originally contained 3436 unique drug names, 884 unique conditions, and 112,329 unique reviews. Levonorgestrel, used for birth control, was the most frequent drug in the dataset.

Distribution plots were utilized to visually display the distribution of 'rating' and 'usefulCount'. A bar plot illustrated the correlation between ratings and usefulness, revealing a trend that higher-rated drugs are generally perceived as more useful. This valuable insight further guided our data preprocessing strategy.

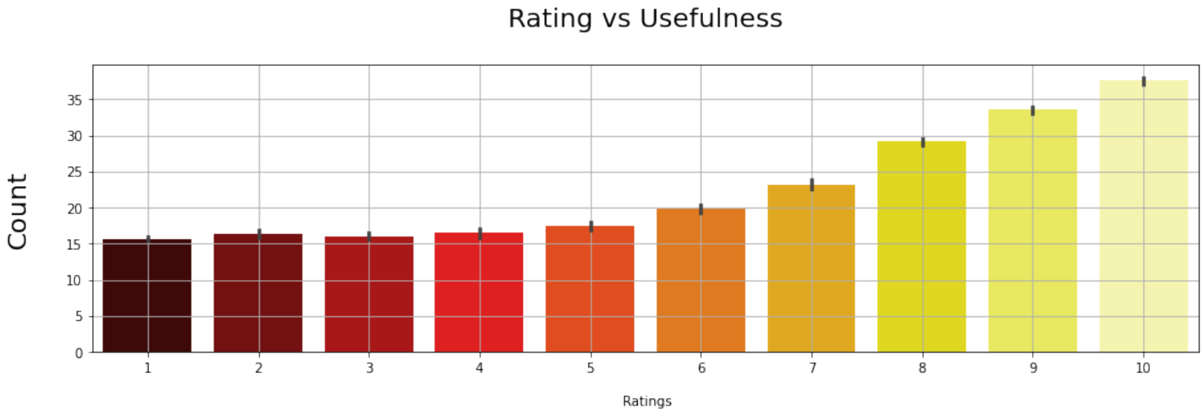


Figure 1: rating vs Useful Count

During our data exploration and preprocessing, we discovered a strong positive linear relationship between the 'rating' and 'usefulCount' features, indicating that drugs with higher ratings are deemed more useful by users.

5 Data Pre-processing

data cleaning : handling Missing Values,removing duplications:The dataset contained 899 missing values in the 'condition' column. This column is crucial for our analysis, thus the decision was made to remove these records. Following this, no missing values remained in the dataset

6 Feature Engineering

Generate a new column named 'len' that captures the character count of each review. Investigate the correlation between the length of reviews and the corresponding ratings. The preliminary assessment hints at a pattern where reviews with lower ratings (from 1 to 3) tend to be more verbose on average, with some stretching to greater lengths. This might imply that when users are unhappy with a medication, they might be inclined to provide more elaborate feedback. Conversely, reviews with higher ratings (spanning 8 to 10) are generally more concisely, suggesting that content users might offer their feedback in a more direct manner. A new feature engineered, termed as 'text length', was constructed to capture the pattern length of individual reviews, speculating it might have a potential correlation with the drug's rating score. plot the distribution of review length to visualize the observed patterns.

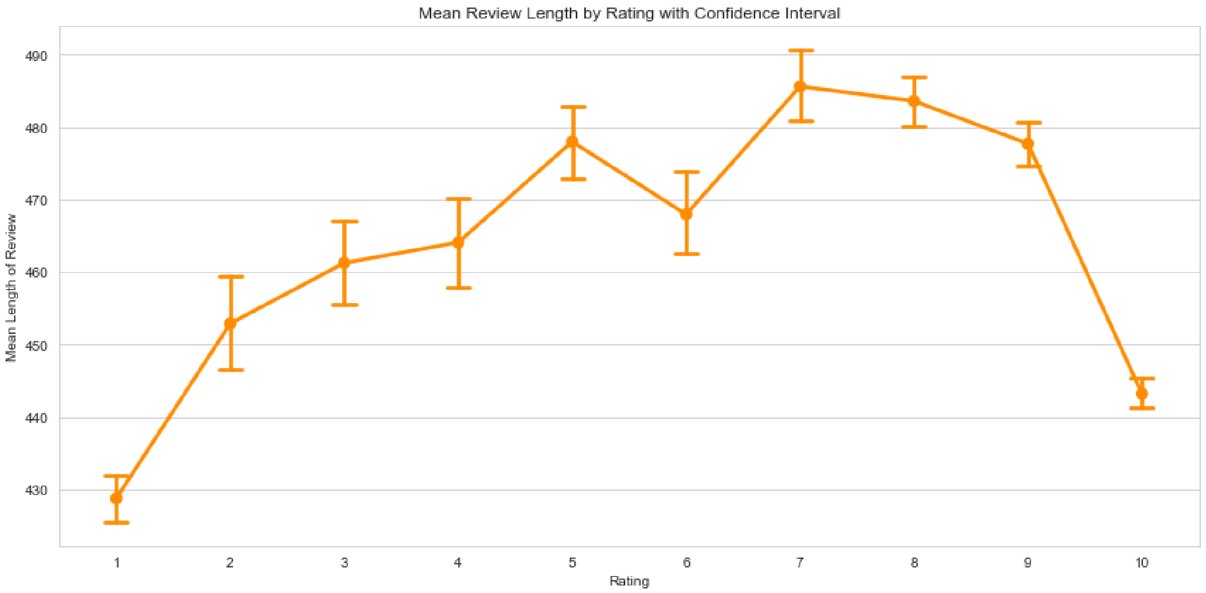


Figure 2: Distribution of review length by rating

7 Text Pre-processing

The source data included user reviews of various drugs. Text mining techniques were used to clean and preprocess the data by converting text to lowercase, removing punctuation, special characters, and stop words. Process the review text data to remove unnecessary elements and to make it ready for analysis.

Data preprocessing is the first and essential step in any data project. It involves preparing and cleaning data for further analysis:

Text Normalization: This includes converting all text to the same case (usually lower case) to maintain uniformity and avoid duplication based on case differences.

Removing unnecessary characters: Text data, especially when collected from the web, often contains a lot of noise, including HTML tags, punctuation marks, numbers, special characters, white spaces, etc. These usually don't contribute to the sentiment of the text and can be removed.

Tokenization: This is the process of breaking down the text into individual words or tokens. This is necessary for the next steps, such as stopword removal and stemming.

Stopword Removal: Stopwords are commonly used words (such as "the", "a", "an", "in") that a search engine has been programmed to ignore. These words don't usually contribute to the sentiment and can be removed.

Stemming/Lemmatization: These are techniques to reduce words to their root form. For example, "running" gets reduced to "run". This helps in reducing the dimensionality of the data and grouping similar sentiments together.

8 Sentiment Analysis

Sentiment Analysis identifies emotions in words, helping categorize reviews as positive, negative, or neutral. The Natural Language Toolkit (NLTK) offers VADER, an effective tool for analyzing casual texts like social media. VADER can analyze varying text lengths, is adept at recognizing emojis and informal language, and doesn't require extensive training data. However, like all lexicon-based approaches, VADER may have limitations in capturing the nuances of complex language and context. From the sentiment analysis, four main features were derived: 'neu', 'neg', 'pos', and 'compound'. The 'compound' values range between -1 and 1, with its mean value suggesting balanced reviews. The 'neu' feature, however, indicates most reviews are neutral. A histogram of 'compound' scores was made to visually present the sentiment distribution in drug reviews.

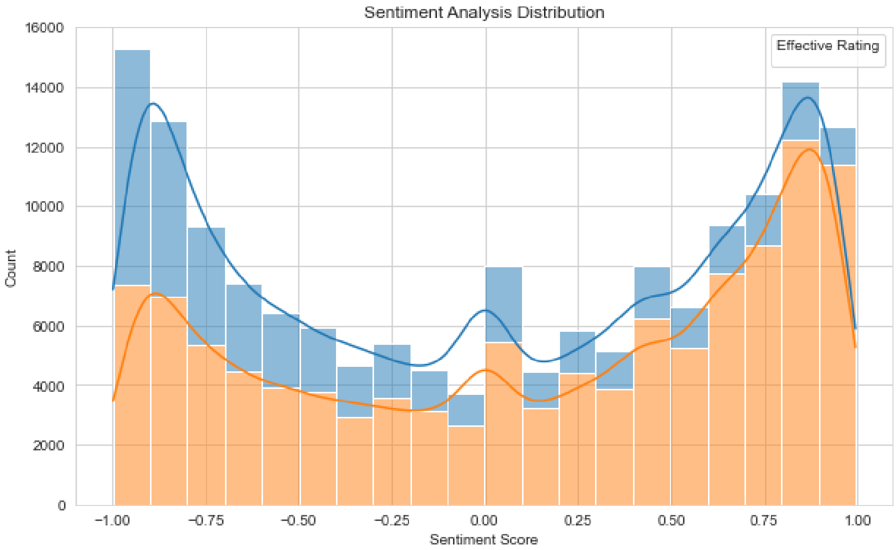


Figure 3: distribution of sentiment scores and effective rating

The data was grouped by 'rating' to study the effect of sentiment on ratings. We observed varied sentiment scores within each rating, from negative to positive. However, on average, higher ratings correlated with more positive sentiments. Simply put, better-rated drugs typically have more positive reviews. This relationship can be useful in predicting drug ratings using review sentiments. This insight could be very valuable in building a model to predict drug ratings or effectiveness from review text.

9 Data Transformation

This provided valuable context for our new engineered feature 'eff score'. The 'eff score' was designed to simplify the effectiveness of the drugs into a binary classification, where 0 stands for less effective and 1 for effective. The creation of 'eff score' respected the observed usefulness threshold in the data and encapsulated the relationship between 'rating' and 'usefulCount', therefore, providing a simplified yet powerful representation for subsequent data analysis and modeling.

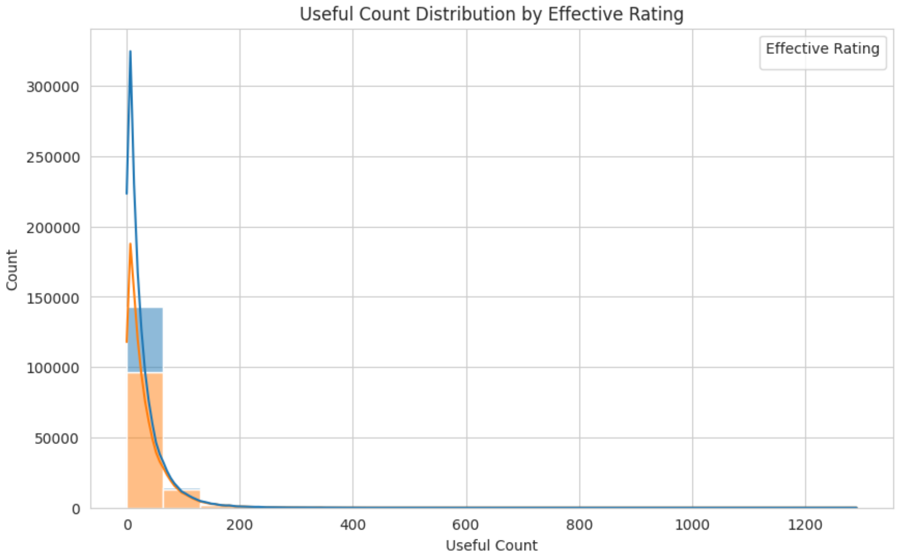


Figure 4: distribution of Usefullcount and effective rating

Transforming the original 'rating' values in a dataset into a new variable named 'eff score'. This transformation involves rescaling the 'rating' values from their original scale (0-10) to a 0-5 scale and then discretizing those into two classes.The code scales the ratings in the "rating" column to a new "Effective Rating" based on a scaling formula and assigns the scaled ratings to the "eff score" column in the dataset. The scaling formula ensures that the "Effective Rating" ranges from 0 to 1. Ratings below 3 are mapped to 0, and ratings of 3 and above are mapped to 1.

10 Unsupervised Model : K-Means

The clustering model helped us group drugs based on their reviews and sentiments, providing valuable insights into user experiences for different conditions. Most drugs received neutral sentiments, but there were variations across conditions. This analysis allows us to better understand user perceptions and satisfaction, which can be helpful in making informed decisions about drug development and marketing strategies

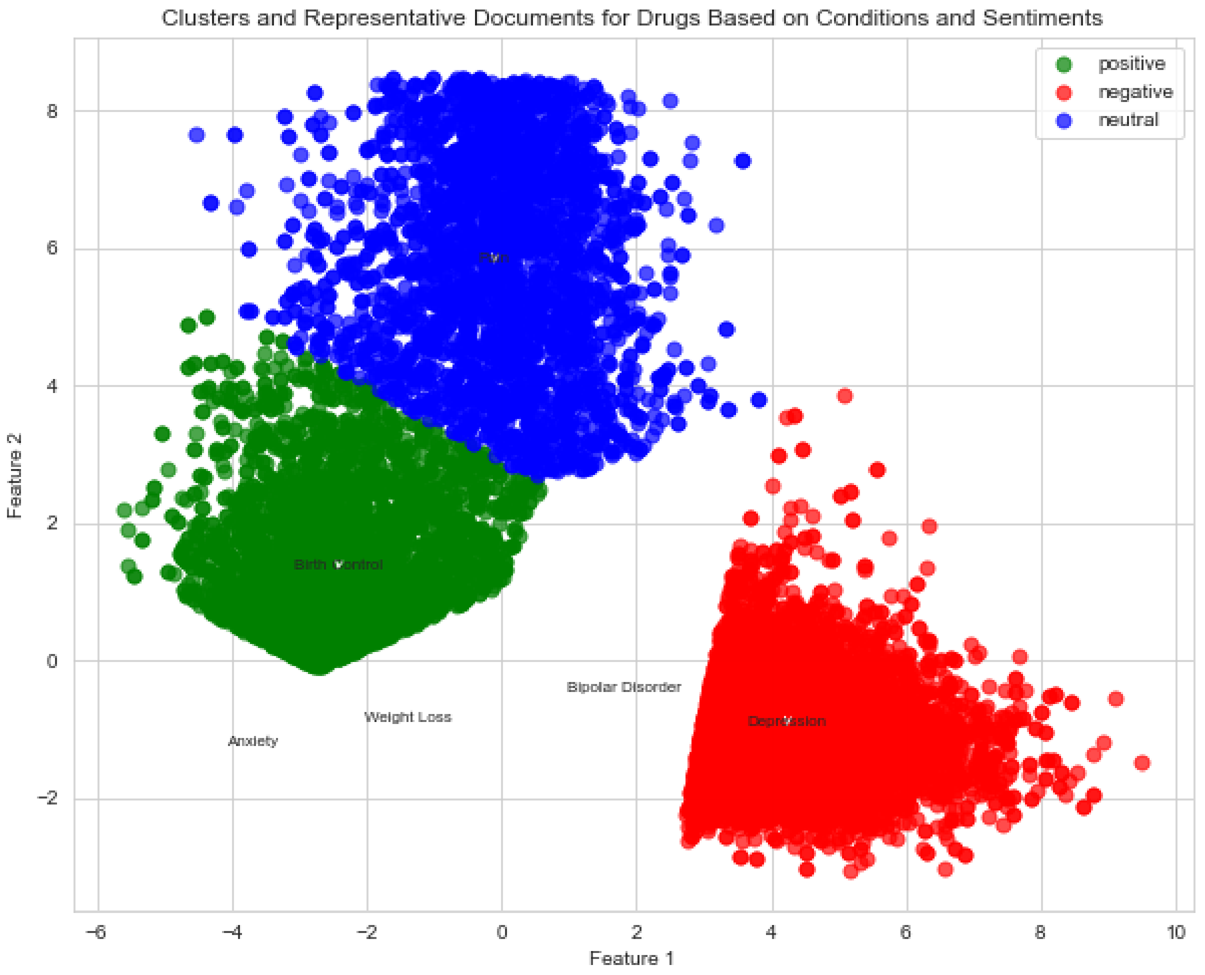


Figure 5: K-Means sentiment clusters

the plot reveals that for most conditions, drugs tend to receive neutral sentiments from users. However, there are variations across conditions, with some conditions having a more balanced distribution of sentiments (e.g.,acne and pain), while others show stronger associations with either negative (e.g., weight loss and depression) or positive (e.g., birth control) sentiments. This analysis provides valuable insights into the overall sentiment and user experiences related to drugs for different conditions, helping us better understand user perceptions and satisfaction.

10.1 Discovering Underlying Topics

In the context of this text mining task, K-Means clustering is used to group together similar drug reviews. The benefit of this approach is multifold:It can help identify the underlying topics that people are talking about in their reviews. Each cluster can be seen as a specific topic that people mention when discussing the drug. By identifying these topics, you can better understand the key areas of concern or interest for patients

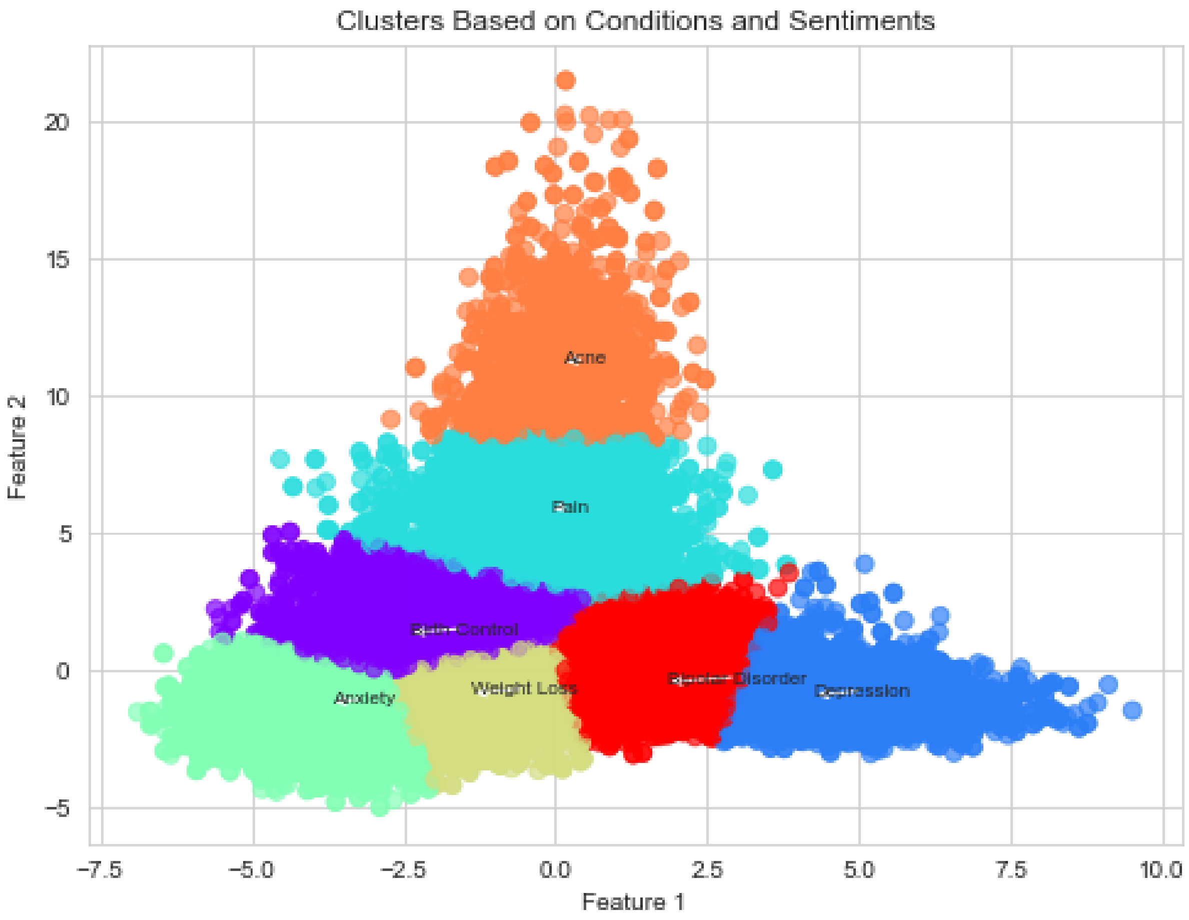


Figure 6: K-Means review clusters

11 Supervised Machine Learning: Predictive Models Development

After the sentiment analysis, the model that performs the best is selected and deployed. It's trained on the entire dataset and can then be used to predict sentiments of new drug reviews. This aids in continually keeping track of the effectiveness of drugs, which could be subject to change due to factors like changes in manufacturing process, user demography changes, etc.

The model can be deployed in a real-time system, which allows for sentiment analysis to be performed on-the-fly as new reviews come in, allowing the company to react quickly to changes in public perception of their drugs.

Each model was used in combination with the TfidfVectorizer in a pipeline, transforming the cleaned text data into a matrix of TF-IDF features and review length. The response variable was binary, labeling the drugs as effective (1) or ineffective (0). The data was divided into a 80-20 split for training and testing purposes with SMOTE used to rectify class imbalance This resulted in two distinct pipelines for: Logistic Regression, for Random Forest.

$$\text{TF(t,d)} = \text{No.of times term t appears in doct d}_{\overline{Totalnumberoftermsindocumentd}} \qquad \text{IDF(t,D)} = \log \frac{Totalnumberofdocuments}{Numofdocswithtermtinit}$$

$$\text{TF-IDF(t,d,D)} = \text{TF(t,d)} \times IDF(t,D)$$

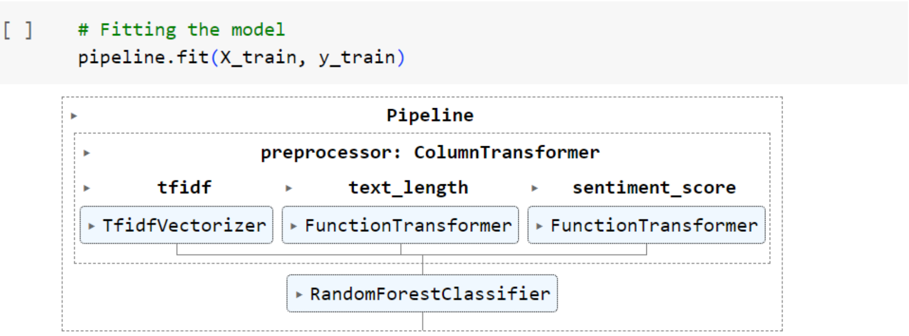
11.1 Random Forest Model

Random Forest model created by employing a blend of TF-IDF Vectorizer , review length as features and sentiment score .

Details on Feature Development:

TF-IDF Vectorization: This approach converts the raw review text into a numerical format, capturing both individual words and bi-grams. The resulting measure underscores the significance of words within the review, effectively highlighting both common and distinct terms, thus offering a comprehensive understanding of the review's essence.

Extraction of Review Length: Based on our earlier observation, we included review length as an additional feature that computes the number of words in each review, offering insight into the extent of the user's feedback.



```
[ ]  # Fitting the model
     pipeline.fit(X_train, y_train)
```

Figure 7: fitting random forest model

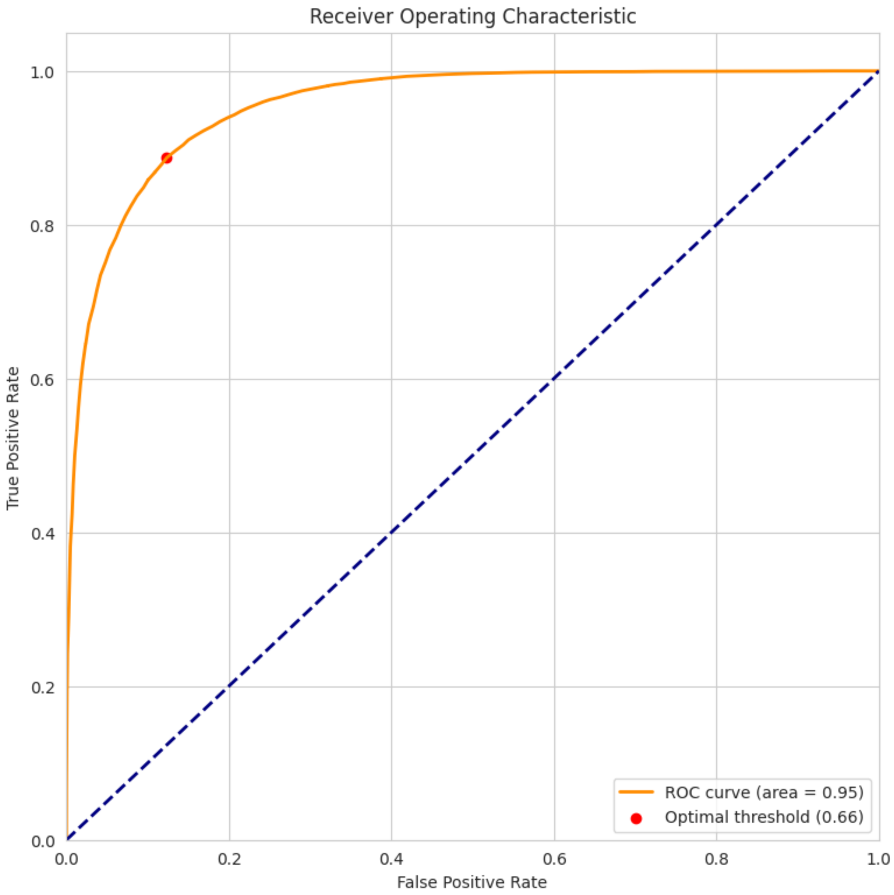


Figure 8: ROC curve of random forest model

The model had an overall accuracy of 90%. In terms of precision and recall, for '0' (less effective), the precision was 91% and the recall was 71%. For '1' (more effective), the precision was 89% and the recall was a substantial 97%. The F1 scores were 81% for '0' and 93% for '1'. The AUC-ROC score was 0.95.

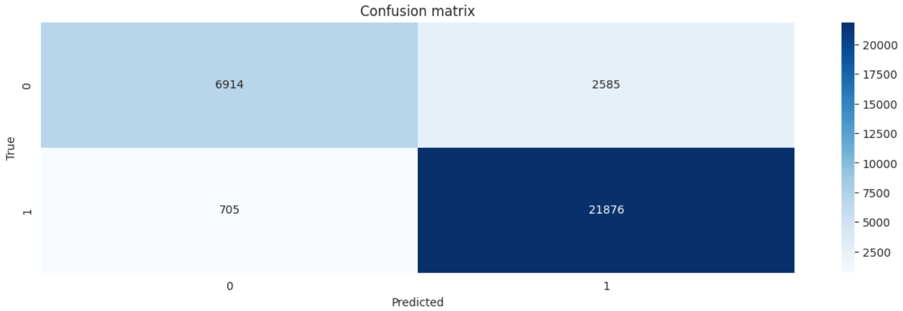


Figure 9: Confusion Matrix of random forest model

11.2 2. Logistic Regression Model

Lastly, we utilized a Logistic Regression model. Logistic regression is a statistical model that uses a logistic function to model a binary dependent variable.

The model presented an overall accuracy of 81%. The precision and recall for '0' (less effective) were 64% and 80% respectively, and for '1' (more effective), these were 91% and 81%. The F1 scores were 71%

for ’0’ and 85% for ’1’. average accuracy is 81% with The AUC-ROC score of 88%, indicating a strong performance.

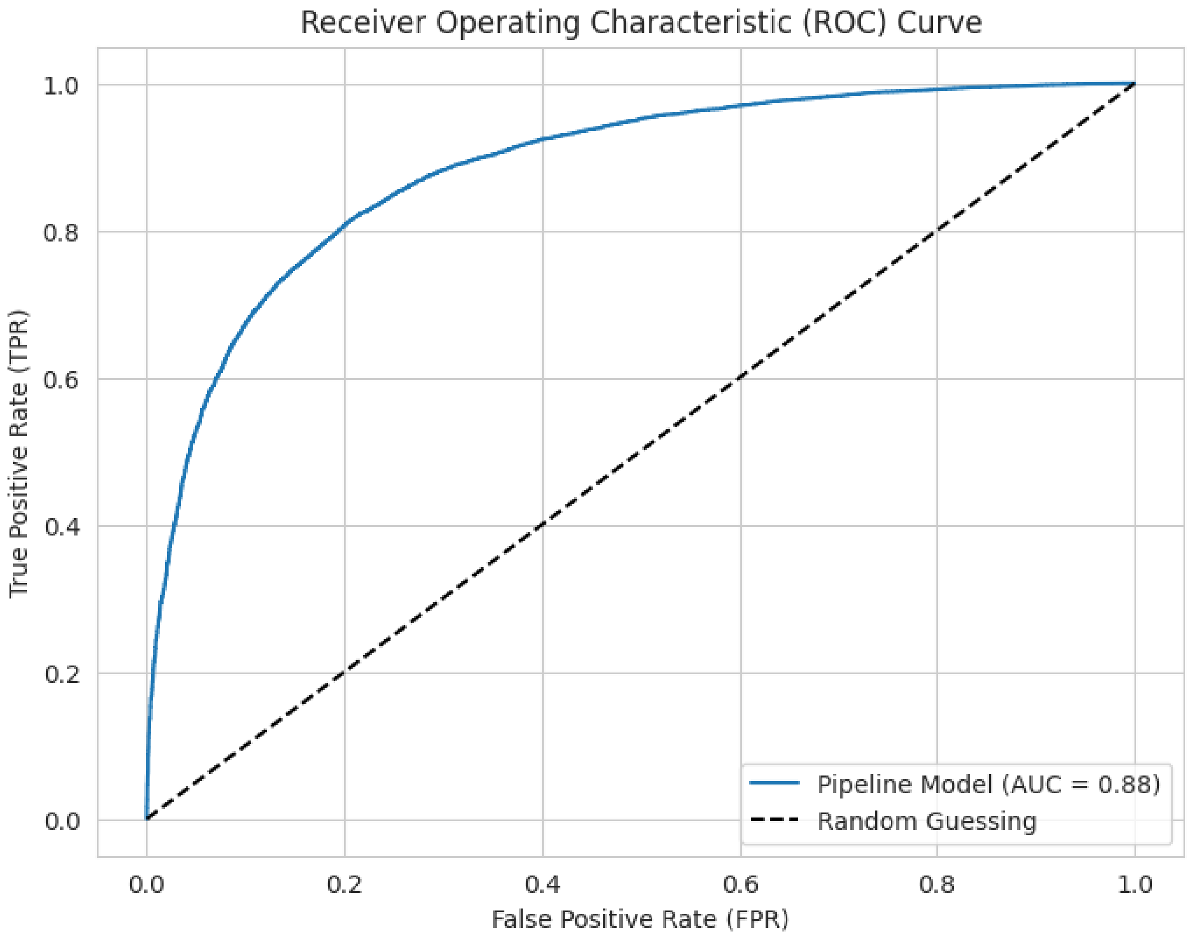


Figure 10: ROC curve of Logistic Regression Model

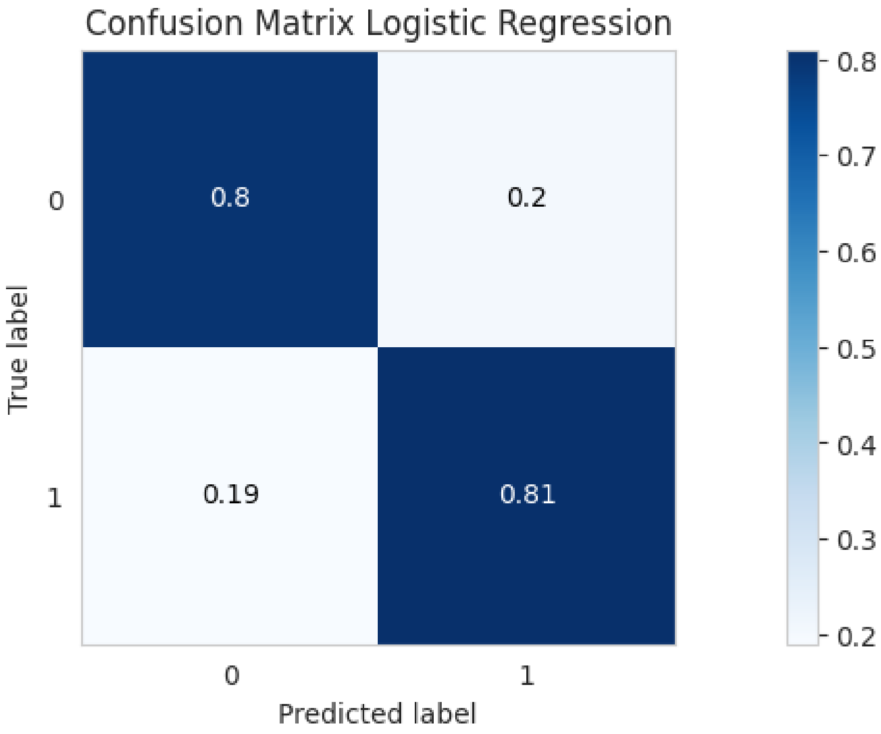
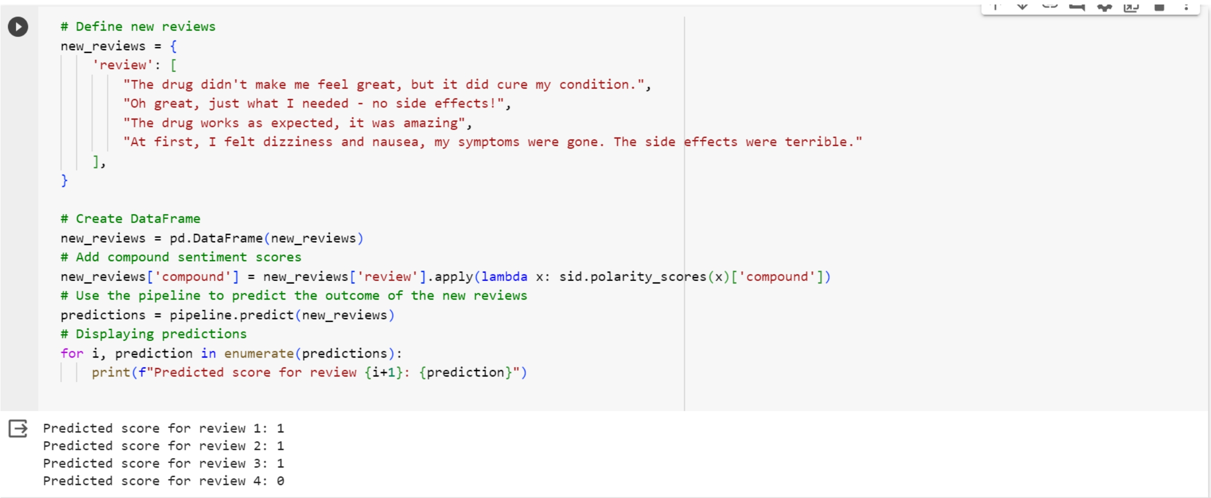


Figure 11: Confusion Matrix

By using different models, we were able to compare the effectiveness of each approach in predicting drug effectiveness based on patient reviews. Despite using the same feature extraction method (TF-IDF Vectorizer) and review length, different classifiers demonstrated varying performance. The models' performance suggests valuable insights can be gleaned from unstructured text data like patient reviews.

12 Model Testing

On applying these models to new unseen reviews, each generated predictive scores (1 for effective and 0 for ineffective). On the whole, all models showed an appreciable performance, with accurate classification of the reviews based on their content.(example below taken from random forest prediction on unseen reviews)ù



```
# Define new reviews
new_reviews = {
    'review': [
        "The drug didn't make me feel great, but it did cure my condition.",
        "Oh great, just what I needed - no side effects!",
        "The drug works as expected, it was amazing",
        "At first, I felt dizziness and nausea, my symptoms were gone. The side effects were terrible."
    ],
}

# Create DataFrame
new_reviews = pd.DataFrame(new_reviews)
# Add compound sentiment scores
new_reviews['compound'] = new_reviews['review'].apply(lambda x: sid.polarity_scores(x)['compound'])
# Use the pipeline to predict the outcome of the new reviews
predictions = pipeline.predict(new_reviews)
# Displaying predictions
for i, prediction in enumerate(predictions):
    print(f"Predicted score for review {i+1}: {prediction}")
```

```
Predicted score for review 1: 1
Predicted score for review 2: 1
Predicted score for review 3: 1
Predicted score for review 4: 0
```

Figure 12: Model testing results of unseen reviews

13 Conclusion

The study demonstrates the power of text mining and sentiment analysis in conjunction with machine learning models in predicting drug effectiveness. Through a combination of data preprocessing, feature engineering, and algorithm tuning, we developed models that can predict drug review effectiveness with considerable accuracy. These insights can assist pharmaceutical firms and healthcare providers in understanding patient feedback and making informed decisions

Future work could consider the inclusion of other text feature extraction techniques or the application of more complex models, such as deep learning algorithms, to further enhance predictive accuracy.

Reference

1. dataset source from kaggle : https://www.kaggle.com/datasets/sahilfaizal/drug-prescription-based-on-consumer-reviews

2. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 1-135.

3. Hutto, C. J., & Gilbert, E. E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14).*

4. Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

5. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513-523.