Sentiment Analysis On The US Drug Reviews

Faiza Ayoun - 015960139

Heer Bhavesh Parekh - 015270320

Syama Ravi Teja Jerrypothula - 015906098

Vamshi krushna Lakavath - 015351310

Department of Data Analytics, San Jose State University

Data 245: Machine Learning Technologies

Spring 2022

Dr. Shih Yu Chang

May, 18 2022

Abstract

Machine learning is one of the promising and fastest growing fields. In the healthcare industry, the benefits of using machine learning is immeasurable. Using textual analysis decidening the context of a sentence as positive, negative, or neutral is called sentiment analysis. The sentiment analysis on the text dataset has evolved extensively, but if it comes to sensitive data like health care, there is still a lot of demand and necessity for its improvement. So in this project we have collected a large set of customer reviews in the textual dataset format on the drugs being used by some patients in the United States. The dataset has been obtained from the UCI Machine Learning, which includes condition, reviews, ratings, useful count, and the drug name. The ratings in the dataset for each drug varies from 1 to 10. The dataset has a total of 7 columns and 215063 rows. Based on the rating range we have decided the nature of a review as positive if rating is 7-10, neutral if the rating is 4-6, and negative if the rating is 1-3. Based on the reviews and ratings we want to build a sentiment analysis algorithm that provides the analysis on each drug for a medical condition. For that we have performed data engineering tasks like data cleaning, data preprocessing, data transformation, and data preparation which includes splitting data for training and testing purposes. Before building the model we have done natural language processing that involves tokenization for splitting sentences into split words, removing stop-words and punctuations, stemming, and encoding. We have used TextBlob and NLTK library for generating sentiment scores, and also TF-IDF to determine the relevance and importance of different words for all reviews. We have built Naive Bayes, KNN, Logistic Regression, Random Forest, Decision Tree, XGBoost, and Voting classifier models for our project. Using the test data set we have evaluated our models and out of all models the Random Forest has given the best results with the highest accuracy of 89%, precision of 90%, recall of 89%, and F1-Score of 89%.

Table of Contents

Introduction

In the machine learning sector, natural language processing (NLP) is extremely important. It is particularly useful in the realm of medical healthcare, where it is used to analyze medical reviews and literature (Vijayaraghavan & Basu, 2020). Sentiment analysis is a sort of subjectivity analysis that investigates emotion in a textual entity with the goal of determining the sentiment orientations (i.e., positive, negative, or neutral) of people's comments on various subjects (Na, 2015). Sentiment analysis has been conducted on different sectors of textual content (reviews) like product reviews, movie reviews, services reviews, stock market analysis, election news analysis, etc. but very few researchers have conducted studies on sentiment analysis in the field of healthcare and medicine.

There are hundreds of thousands of drugs available in the market. For a single medical condition there are hundreds of drugs in the market manufactured by different companies. The same drug has been provided by different brands using different names, but each of these medicines work differently from others and their effectiveness is a lot different when it comes to curing the disease. Therefore, selecting an effective and affordable drug has always been difficult. In addition to the doctor's recommendation, users want to know more about the medicines they are using from the previous users from their experiences. Today many people choose to analyze the rating and reviews before consuming any product. In the field of health care this trend is rapidly growing. Before buying a drug, patients want to consider the experiences of previous consumers of the same drugs for the same health condition. To come up with a solution for this problem there should be in-depth analysis, review, and rating on each brand drug and it should be publicly available. Also, there is a need to analyze and know the effectiveness of each brand of drug under each circumstance for better use and cure, which can be done through the experimental method, but that can be a very expensive and time consuming process. One of the best alternatives is to analyze the

ratings and reviews provided by the users of these drugs. That can be done by sentiment analysis.

Sentiment analysis is a rapidly growing field of study in user data analysis. It is one of the most promising technologies for analyzing the reviews and ratings of the large set of customers dataset.

There are a lot of researchers conducting research on the product review of customers but a very few researches have been done on the drug reviews. So in this project we want to perform sentiment analysis on the Drug Review and Ratings Dataset. We will build Machine learning models that can perform the Sentiment Analysis on the drug review dataset and help us know the most effective drug for a medical condition. Then we will build a system that will provide sentiment analysis scores for the best available drug for underlying medical conditions based on the previous user experiences and reviews.

For this project we are following CRISP-DM methodology. Which involves Business understanding, data understanding, data preparation, model building, and model evaluation. For this project we need a large set of reviews and ratings dataset, that has been collected from the UCI Machine Learning Repository. This dataset has 7 columns and 215063 rows. The columns are UniqueID, Condition, Drug name, reviews, rating, date, and useful count. Each row is a review and rating provided by users for a drug. Based on the rating range we have decided the nature of a review as positive if rating is 7-10, neutral if the rating is 4-6, and negative if the rating is 1-3.

Based on the reviews and ratings provided in this dataset, we will build a sentiment analysis model that provides the analysis and sentiment report on each drug used for a medical condition. After collecting data will perform data engineering tasks like data cleaning, data preprocessing, data transformation, and data preparation which includes splitting data for training and testing purposes. For this project we will also do natural

language processing that involves tokenization for splitting sentences into split words, We will remove stop-words and punctuations, Stemming, and encoding. We are also choosing to use TextBlob for processing textual data, NLTK library for generating sentiment scores, and TF-IDF to determine the relevance and importance of words.

In this project we will build Naive Bayes, Neural Network, Random Forest, Support vector machine, XGboost, and Ensemble Techniques. We will split the dataset into three parts, for training we use 80% of the data, for testing we use 20% of the data. Using the train data we will train our models, then we will build the models. Moreover, we will evaluate our models using some metrics like accuracy, precision, recall, and f1 score. After evaluating the performances of the models we will deploy the best performing model on the web portal.

Literature Survey

One of the papers we chose is the “Drug Rating Prediction Based on the Drug Review Dataset from UCI Machine Learning Repository " by Ting-Chou Lin , Yufei Mao, and Yen-Yi Wu from the University of San Diego, California. In this paper, many techniques have been used for feature generation : one hot encoding, bag of words, and tf-idf. For tf-Idf, the range of unigrams used are from 3000 to 45000. Moreover, bi-gram and trigrams were used. The selected machine learning algorithms were, Linear Regression, Ridge Regression, Gradient Descent Regression, and Random Forest Regressor.

In the other research Han et al., (2020) proposed an aspect level sentiment analysis is a quite well sentiment analysis task for determining the polarities of a given target in a sentence. They have noticed that this technique is not performed well on a variety of datasets. So they came up with a Pretraining and Multi-task learning model based on Double BiGRU (PM-DBiGRU). Multi-task learning is also used by them to transfer useful domain information from a corpus of small text-level pharmacological reviews. They have also come up with a dataset called SentiDrugs for an aspect-level sentiment categorization of drug reviews whereby each review can include one or more objectives. For this project they have also designed a couple of other models and methined that in their future work they plan to consider more models.

Researchers (Gräßer et al., 2018) have worked on a variety of duties involving drug reviews and data gathered from online pharmacy review websites. They have utilized sentiment analysis to forecast how users feel about general wellbeing, side effects, and drug effectiveness in user reviews. To address the lack of annotated data, they looked at the transferability of trained classification models across domains, such as circumstances and data sources. They focused on the transfer learning methodologies that could be used to leverage cross-domain commonalities and is a viable technique for cross-domain sentiment

analysis. They trained the logistic regression models using lexical features like unigrams, bigrams and trigrams extracted from the reviews. Aside from patient satisfaction, sentiments about efficiency and adverse effects were investigated.

Acceptable classification findings were achieved based on the feature and data source.

Project Objectives

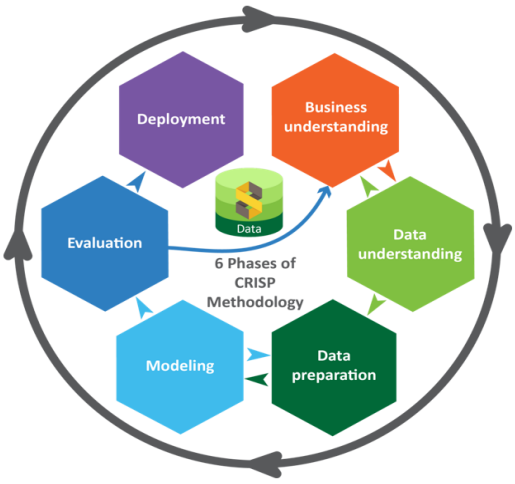
As we know, Natural Language Processing is one of the most important branches in artificial intelligence, interpreters, and computer vision. It deals with the interrelationship of how to design the computers in order to interpret and analyze the massive quantities of natural language data.This technique is used in various industries, specifically in the healthcare field for analysis of medical records, reviews and different texts. We decided to utilize a data set that would be more representative of the medical healthcare industry in our project, and we used sentiment analysis to explore the application of NLP algorithms. Our main goal is to understand the reviews and comments that are given by the consumers on every drug. This will help us in anticipating the real sentiments of the consumers. The terms used in the review, as well as the sentiments expressed, can be utilized to classify an effective medication review. In our project, we intended to see how important the terms in the review are and how they would affect the sentiment prediction of the review and the ability to anticipate the reviewer's ratings. Additionally, we also wanted to check how successful it was to acknowledge the rating classifications along with the sentiments of the reviews. We were confident that the statements in the reviews may create a pivotal impact in defining the tone of the evaluation. The best way to resolve our sentiment recognition difficulty was Natural Language Processing. This machine learning technique will perform the Sentiment Analysis on the drug review dataset and help us know the most effective drug for a medical condition.

Project Management Plan

(Schröer et al., 2021) Data research initiatives can benefit from project planning and development strategies. Agile techniques should support the implementation of a project strategy, and optimization models like CRISP-DM can even enhance it. (Wowczko, 2015) CRISP-DM is a commonly used data mining model. It comprises six basic stages, as shown in Figure 1, that assist in transparently managing the performance of any project. The CRISP-DM project methodology acts as a blueprint for the project's implementation stages. As per the CRISP-DM approach, the project's road map is divided into six stages. The stages include Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

Figure 1

CRISP-DM Methodology



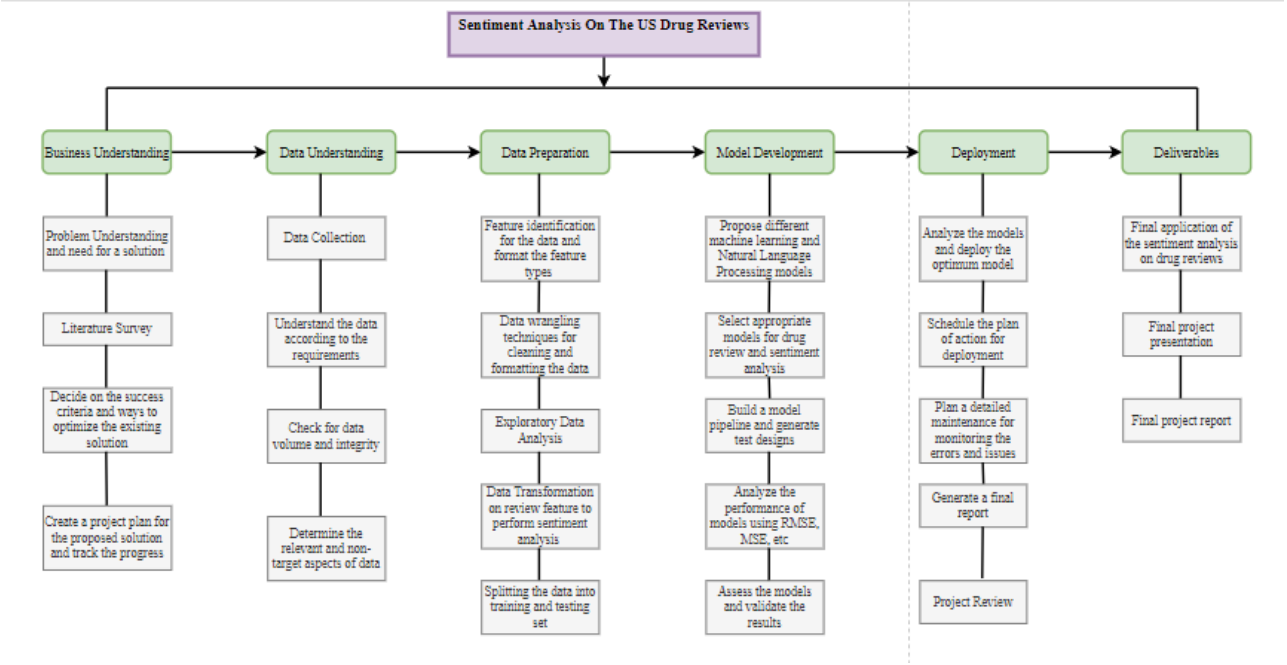
Tare, R. (2021, March 29). *#1 : CRISP DM framework*. Ruchareads. Retrieved from https://ruchareads.wordpress.com/2021/03/29/1-crisp-dm-framework/

The first stage is the Business Understanding stage where the main goals of the project are discussed. Here, we determine the sentiments of the drug reviews and know the most effective drug for a medical condition. The Business Understanding stage would be accomplished by creating a project plan that includes performance targets, information about

existing approaches that can be optimized, and systems to assess the progress of the recommended techniques. The next two phases of the methodology are data understanding and data preparation. It refers to gathering relevant data based on a business understanding and requirements in order to preprocess and clean the data before evaluating and analyzing it. As a result, a variety of data sources have been investigated in order to identify the proper dataset with the necessary columns and rows. A better understanding of the data source as well as an assessment of the acquired test data are included in the data description and data exploration methods. Data quality is the final step in the Data Understanding process, and it ensures data integrity by ensuring that inconsistencies in the data are removed and validated data with consistency is assessed for further analysis. Every dataset feature is examined for acceptable and irreplaceable incomplete values, changing missing values, removing features with frequent and missing values, and so on. The data will then be divided into splits, such as train, validate, and test data. The fourth stage is modeling, which defines the data assumptions depending on different modeling types used and the data that is trained. This step also includes model optimization approaches and a testing plan. The evaluation phase, is used to evaluate the precision and the results of multiple models, is the fifth phase. A remodeling stage is required if any errors or faults are detected; else, the best model is transferred to the deployment phase. Deployment is the final step, which comprises deploying the most optimal model. Figure 2 shows the Work Breakdown Structure.

Figure 2

Work Breakdown Structure



Note. The six phases of Work Breakdown Structure

Project Resource Requirement Plan

The resources used for this project are:

- Google Collab - Used for shared access to the notebooks where we wrote our python code.
- LucidChart - To create the WBS and flow of the project.
- Zoom Calls - Used for project discussions
- Tableau Prep Builder - Merge and clean the initial datasets
- Trello - Project Management Tool to divide the tasks.

Multidisciplinary Analysis

Data Understanding

The dataset contains 215063 rows and has seven features : uniqueID, drugName, condition, review, rating, date, and usefulCount.

- The drugName feature has no null values with 3671 unique values. Figure 3, shows the information about the different features.

Figure 3

Drug Review Dataset

```
In [13]: data.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 215063 entries, 206461 to 113712
Data columns (total 6 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   drugName     215063 non-null  object
 1   condition    213869 non-null  object
 2   review       215063 non-null  object
 3   rating       215063 non-null  int64
 4   date         215063 non-null  object
 5   usefulCount  215063 non-null  int64
dtypes: int64(2), object(4)
memory usage: 11.5+ MB
```

Note. Information about the drug review dataset

- The condition feature has many nulls and meaningless values. We found many values in the condition feature that contain the value “ user found this comment helpful”. Instead of dropping all these 1171 rows, we decided to clean it as described in the next section.
- The review feature is the feature that will mostly be our focus in this project. It has reviews from different patients or families of patients who used the medication. The reviews are full text in english, but it has some special characters that were transferred as HTML codes that should be cleaned before we do any further transformation. This review should be aligned with the rating feature, since the rating is the numerical value of the review.

Figure 4 shows the condition feature with the meaningless values and the review with some HTML code.

Figure 4

uniqueID	drugName	condition	review	rating	date	usefulCount
220696	Loestrin 24 Fe	2 users found this comment helpful.	"I'm 16 and I have been on Loestrin 24 f...	3	3-Nov-10	2
67383	Provera	4 users found this comment helpful.	"I'm 24 years old and have always had a p...	1	27-Mar-16	4
81588	Yaz	3 users found this comment helpful.	"I took Yaz for a little over 2 years. From a...	3	1-Jun-10	3
132965	Loestrin 24 Fe	4 users found this comment helpful.	"Took this pill for 1.) Acne and 2.) Birth Con...	2	24-Jun-14	4
91050	Norco	11 users found this comment helpful.	"I have suffered with low back pain - 2 surger...	9	15-Mar-09	11
...	...	...	...	...	...	...
133354	Tri-Sprintec	3 users found this comment helpful.	"I have been taking this pill for less than a ...	8	24-Sep-10	3
149494	Mirena	5 users found this comment helpful.	"I got the Mirena put in last month. And holy ...	7	12-Feb-13	5
91988	Lyrica	21 users found this comment helpful.	"It was a nightmare.I had the worse side effec...	1	14-Apr-15	21
174757	Dulera	28 users found this comment helpful.	"My 10 year old son took Dulera for asthma. I...	1	29-Feb-12	28
37632	Vyvanse	5 users found this comment helpful.	"Wonderful drug, after being on narcotics for ...	9	29-Dec-09	5

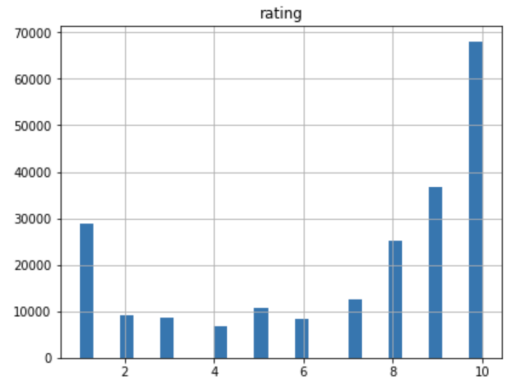
1171 rows × 6 columns

Note. Condition with meaningless values and the review with some HTML code.

- The rating feature doesn't have any null values. This column has a rating from 1 to 10 that can describe the user's opinion about the medication from unsatisfied(1) to very satisfied (10). The value follows a bimodal distribution with the peaks for the highest and lower rating. It seems that the users tend to give extreme ratings whether they are happy or unhappy with the medication. We also noticed that there are more positive than negative ratings with more 10, 9, and 8 ratings (about 60%) than ratings under 7. This results in an overall average rating of 6.99. Figure 5 shows the ratings distribution.

Figure 5

Ratings Distribution

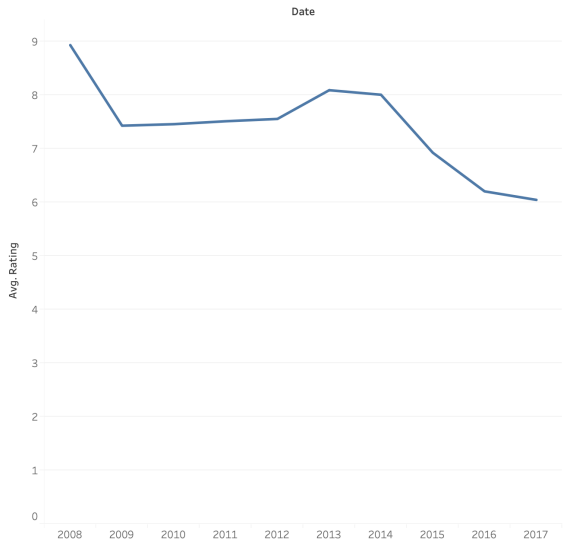


Note. Distribution of ratings.

- The date feature contains the date when the review was published. There are no null values. Figure 6, shows the average rating declining starting from 2014, while figure 7 shows the count of reviews per year. The review count reached a peak in 2016.

Figure 6

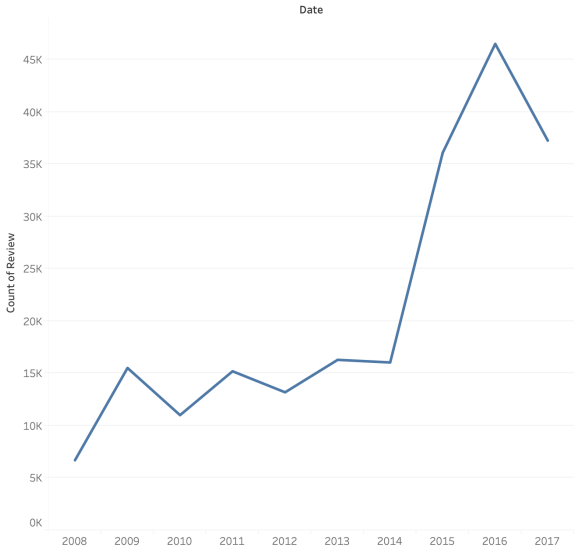
Average Ratings



Note. Average Reviews per Year

Figure 7

Review Count

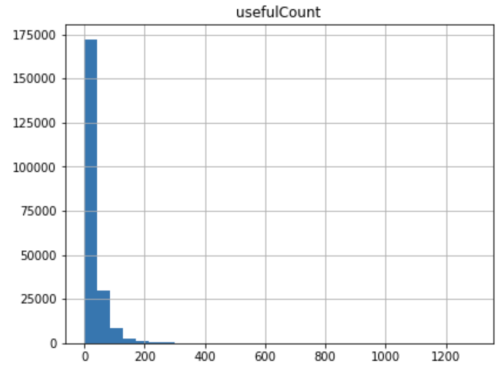


Note. Reviews Count per Year

- The userCount feature contains the number of users who found the review useful. There are no null values, and the values have an exponential distribution with most values (around 95%) being less than 100. Then the distribution drops quickly with fewer rows with higher userfulCounts. We also notice that more than 99% are less than 200, but the value goes to 1291 which makes the distribution squeeze to the left as there are many outliers. We are going to take care of the outliers in our next section. Figure 8 and Figure 9 shows the distribution before and after cleaning respectively.

Figure 8

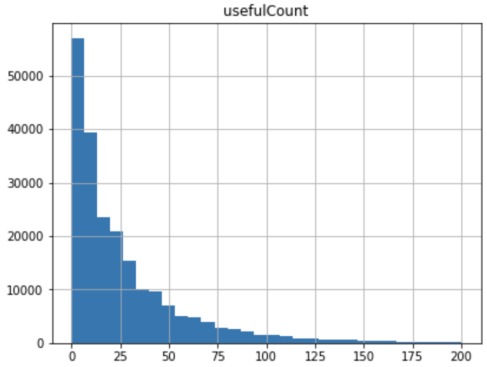
Useful count Distribution



Note. Distribution of UsefulCount before cleaning

Figure 9

Useful Count Distribution



Distribution of Useful Count after cleaning

Data Preprocessing

Cleaning data

We found several anomalies in the dataset that should be taken into consideration. Our dataset had to go through the cleaning process described below:

1. **Replace rows containing a number followed by " user found this comment helpful" in the condition feature**: 1171 rows had a meaningless value of condition, so we replaced all these values with null to be treated later when we fix the null values in other rows.

2. **Fill up null values in condition**: We decided to replace the null values in condition with the most frequent condition having the same drug name as the record with the null condition.
3. **Replace HTML symbols with correspondant characters :** on the reviews we noticed that there are many symbols that are HTML encoded. These characters probably got transferred encoded when the dataset was collected. We decided to manually replace those special codes with the corresponding characters.
4. **Drop the remaining rows with special characters:** after all the symbols replacement above there are 195 rows that still have special characters. We decided to drop these rows which is approximately 0.1% of the whole dataset.
5. **Remove punctuations and stopwords, and perform stemming on the review feature:** to be able to do sentiment analysis, the review column should be cleaned from any distraction to the essential words used for the message sentiment. First, we remove the stop words, which are words that don't add any value to the sentence. Words like this can be was, am, and, and are. Next, we remove the punctuations of all the reviews. Finally, we perform stemming. Stemming is the process of reducing the word to its minimal stem. For example, drive, drive, driving, would become “driv”. We used NLTK word_tokenize, stopwords, and PorterStemmer.
6. **Remove extreme outliers from the usefulCount feature** : a rule of thumb should be, anything that is 1.5 IQR above the third quartile or 1.5 IQR below the first quartile is an outlier. In our case we wanted to keep maximum records so we enlarged this magin to reach 99% of the dataset.

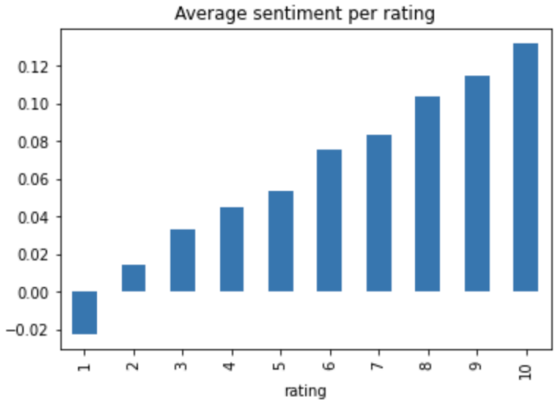
Exploratory Data Analysis

In addition to the distribution histograms we presented in our data understanding section above, we created other visualizations after we cleaned the dataset.

Figure 10 and Figure 11 shows the average sentiment scores per rating distributions. We used TextBlob and NKTK to measure the sentiment for each review, then we plotted the results. The NLTK gives better results as the patients' sentiments follow about the same distribution as the ratings distribution shown on Figure 5.

Figure 10

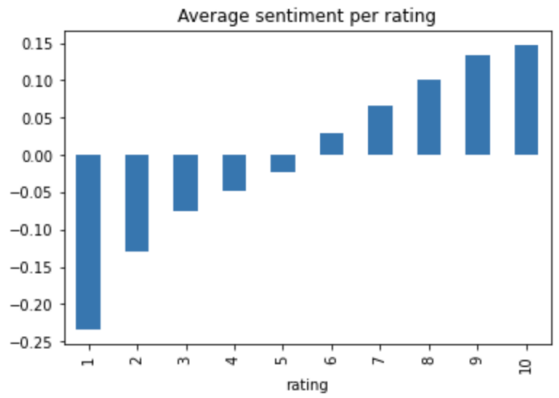
TextBlob Sentiment Distribution



Note. Sentiment Distribution using TextBlob

Figure 11

NLTK Sentiment Distribution

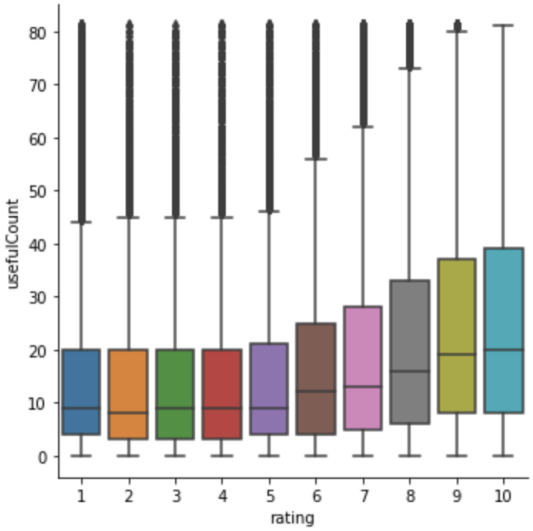


Note. Sentiment Distribution using NLTK

Also, we noticed that the distribution of the usefulCount feature is higher when the rating is higher and vice versa. Figure 12 shows this observation. This means that the higher the rating is, the more it is considered useful.

Figure 12

Useful Count Ratings



Note. UsefulCount per rating

Figure 13

Word Cloud

Note. Word Cloud for the “review” feature

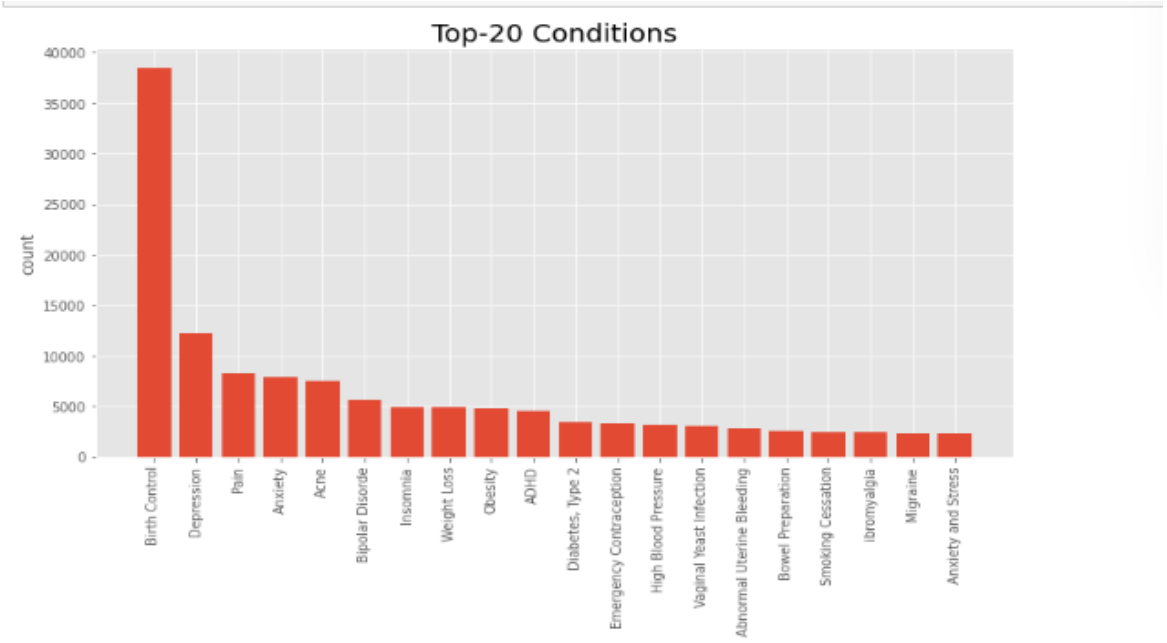
In the word cloud shown in Figure 13, we noticed that many words make sense; they are very frequent and will help our sentiment analysis. Words like side effect, mood swing,

weight gain will certainly be predictive in our models. That is why we decided to perform a TF-IDF transformation to the review feature explained in the next section.

Figure 14 and Figure 15 shows the top 20 conditions and their corresponding drug name that is used by the patients. From Figure 14, we can see that birth control is twice as big as any other disease with a count of 38,000. Figure 15 depicts the drugName Levonorgestrel as the recommended drug. The top 3 drugName has count around 4000 and above. Most of the drugName counts are around 1500 if we look at the top 20.

Figure 14

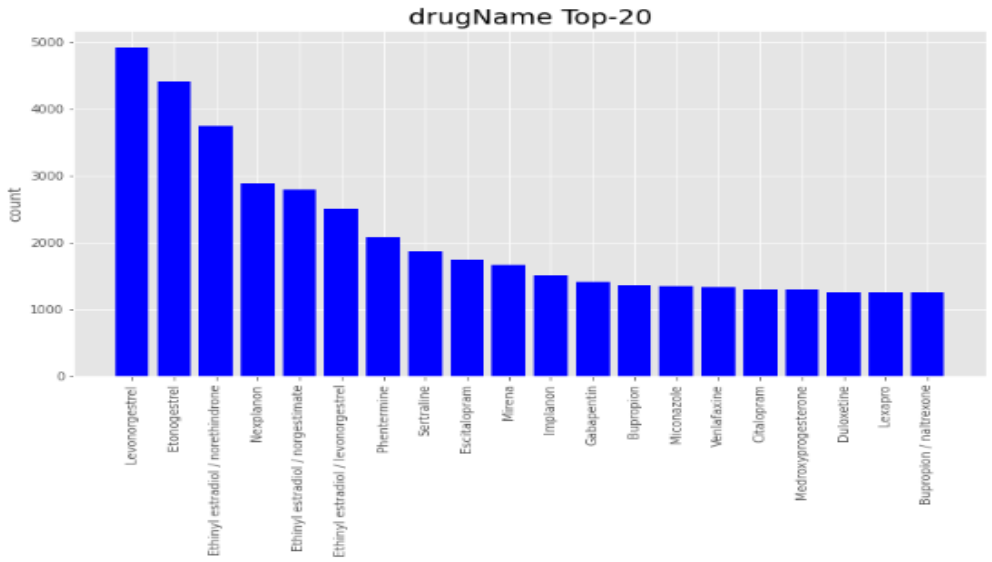
Top- 20 Conditions



Note. Histogram showing the top-20 conditions

Figure 15

Top- 20 Drug Names

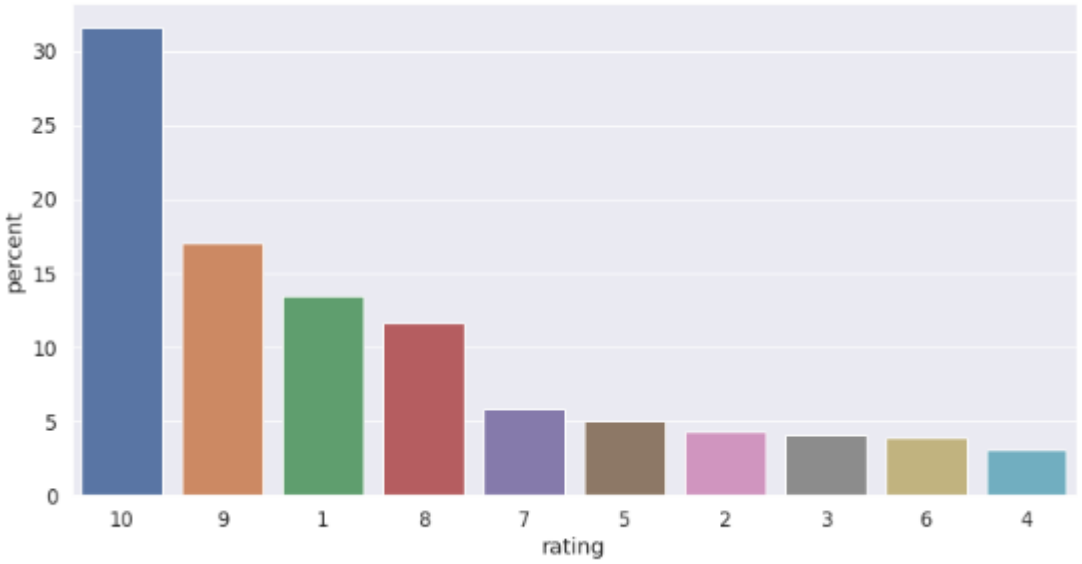


Note. Histogram showing the top-20 drugs.

Next insight that we can predict from the dataset is for the ratings distribution. From Figure 16 we can notice that most of the ratings are high with ratings 10 and 9. rating 1 is also high which shows the extreme ratings of the users. We can say that the users mostly prefer to rate when the drugs are either very useful to them or the drugs fail, or there are some side effects. About 70% of the values have ratings greater than 7. Figure 17 shows most of the drugs are taken for pain, birth control and high blood pressure which are the most common conditions faced by the patients. Additionally, in top- 20 each condition has above 50 drugs.

Figure 16

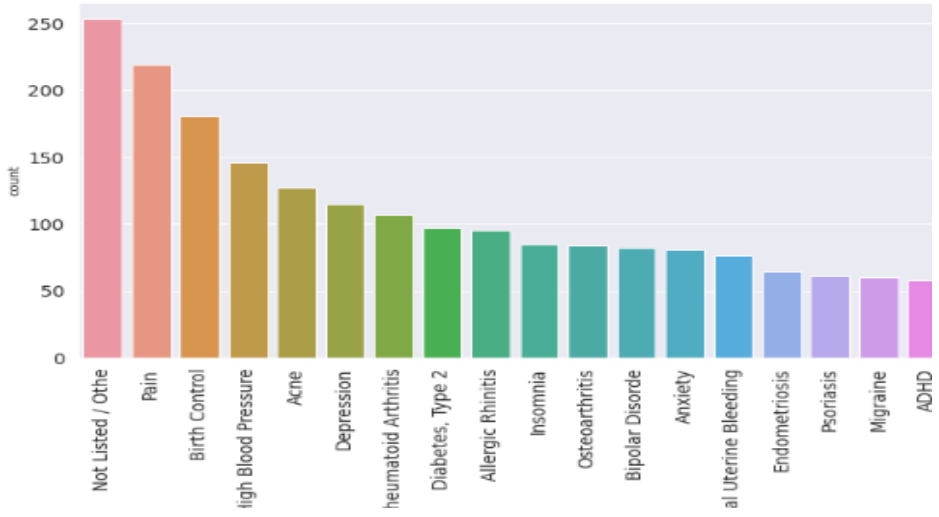
Ratings Percent



Note. Rating percent given by the patients and users.

Figure 17

Top-20 Number of Drugs per condition



Note. The most common conditions and its drug count.

Data Transformation

After better knowing our data and after making a deep cleaning, we proceeded to transform the review feature to perform the sentiment analysis. To be able to develop a model for NLP, we performed many transformations to the review feature before training our models.

1. **Tokenization :** which is splitting a text document into individual words called token. We performed tokenization to all the reviews to be able to remove the stop words and punctuations
2. **Remove stop-words and punctuations:** Stop words are any words that doesn't add any sentiment to a sentence like was, is, because, once, and there. We removed any word in the nltk.corpus stopwords list from all the reviews.
3. **Stemming:** After removing all the strop words, only words that add a sentiment to the review are left. We then transform each of the words to their stem. For example, "ask", "asked", and "asking" are the same verb, but in different tenses, and these three words have the same stem "ask".
4. **TF-IDF:** stands for term frequency-inverse document frequency. This weight measures the relevance of a word to a document within a corpus (collection of documents). Here are the formulas to calculate these measures :

$$\text{tf(t)} = \frac{Nbre\ of\ times\ term\ t\ appears\ in\ a\ doc}{Nbre\ of\ all\ terms\ in\ the\ doc}$$

$$\text{idf(t)} = log\ (\frac{Number\ of\ Documents}{Number\ of\ Documents\ with\ term\ t\ in\ it})$$

The final matrix in our caes will have 3000 features, for the most frequent words in all reviews, and 215063 rows, where each row corresponds to a review from our original dataset. The values for each column i and row j will have the product of the $tf_{i,j}$ and $idf_{i,j}$. The formula for each weight is as follows :

$$w_{i,j} = tf_{i,j} * Log \frac{N}{df_{i,J}}$$

where N is the number of documents, in our case N = 215063

For that we used TfidfVectorizer from sklearn.feature_extraction.text with max_features = 3000.

5. **Change ratings values to positive (1), Negative (-1), and neutral (0)** : our target feature, rating, has a rating range from 1 to 10. For our sentiment analysis we only need three levels for our ternary classification. We decided to replace any ratings less than 4 to a negative number, "-1". Also, ratings between 4 inclusive and 7 exclusive will be neutral, "0". Finally, we will change ratings having a value of more or equal to 7 to a positive number, "1".
6. **Split the data to 80% for training and 20% for testing and evaluation :** we used the train_test_split function from sklearn

Models

Modeling is one of the important steps in the process of project development. This step consists of two main sub-tasks, model selection and model development. Model selection is about selecting the appropriate model for the data based on the working of the model with respective prepared data. While selecting the model, we need to know the working mechanism of the algorithms with the advantages and disadvantages to cross-check with the data and examine the suitability of the model for our data. Keeping all these parameters in mind we have selected some of the algorithms for this project which are listed below.

Logistic Regression

Logistic regression is supervised machine learning technique. In our dataset the target feature has categorical values, which are negative, positive, and neutral. Data does not need to be linear to apply this algorithm because it estimates the most likelihood (probability) using a logistic function. For predicting the target values there will be a logit function which specifies the range of values for each class of the target attribute. The below is the equation indicating calculating Log likelihood of the target values. Where x is the target feature. Therefore this machine learning technique is good to apply. Apart from that, this technique helps us to understand the basic performance of the ML model on this data.

$$L(X|P) = \sum_{i=1,y=1}^{N} logP(x_i) + \sum_{i=1,y=1}^{N} log(1 - P(x_i))$$

K- Nearest Neighbor

K- Nearest Neighbor (KNN) is one of the classification techniques in machine learning. The basic working principle of this technique is considering physical distance between data points. One of the formulas for calculating distance between the data points is shown in the below equation. For instance, if x and y are two instances in the dataset, and p is

the degree of the distance we are calculating. If p is equal to 2, the distance is called the euclidean distance. If it is 1 then it is called the Manhattan distance. p can also be greater than 2 in some cases. After calculating them depending on the k value, the algorithm will consider k training data vectors and classify the unseen vector on a mode basis. Generally it is a better technique than directly predicting algorithms but sometimes due to k values and number of features algorithms may become too complex to perform. As our data consists of several features this model may develop with more complexity.

$$d(x, y) = \left(\sum_{i=1}^{n} \left| x_i - y_i \right|^p \right)^{\frac{1}{p}}$$

Naive Bayes

Naive Bayes is a probabilistic machine learning model which works on bayes theorem. This model is not only a simple model but also fast and accurate. Naive Bayes calculates the likelihood of the target value but the important thing is that this algorithm assumes all the features are independent. The below equations show the calculation probability of target value. It is taken from Bayes theorem. In our dataset we have thousands of features so it is better to neglect the relation between them as it increases more complexity than required.

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Where P(y|X) is posteriori

P(X|y) is likelihood

P(y) is prior

P(X) is predictor prior

Random Forest

Random forest is an ensemble technique where the base learner is the decision tree. In this algorithm, multiple decision trees are combined by utilizing bagging techniques. Decision trees are classification techniques in supervised learning. It is an information learning technique from the statistics applied on the data. In the decision tree, the algorithm calculates entropy and Information Gain (IG) to select the feature to split the dataset. Entropy calculates the variety in the class. Depending on entropy, IG will be calculated for a feature. A feature with highest IG is selected to split the dataset. Again the process to select the next splitting feature continues until the algorithm reaches leaf node/target value. Initially these decision trees are efficient on smaller datasets but later different algorithms are developed like C4.5, cart etc which are efficient to apply on larger datasets. In a random forest, multiple decision trees are created and stored in memory. When an instance is given to the model then target value will be predicted by each decision tree in the random forest. After getting values, the algorithm aggregates all the target values and finalizes the value with highest votes or in other words mode of the array will be selected as final value. As we are considering, there is a high chance of getting good accuracy using this algorithm.

Voting Classifier (Ensemble Technique)

Ensemble technique implies considering several base models to predict the target feature. This technique is highly flexible because of custom design of implementation. But this technique consists of more computing time and complexity. Generally, we will include a variety of algorithms to improve the performance of the model. There are so many variations or types in ensemble techniques. In this project we are using a voting classifier which is one of them. Voting classifier is an ensemble learning method where we can include an array of estimators of different models. The target value will be predicted by aggregating all the results of individual models in the array. In predicting the results there are two ways to vote

for the target value. One is hard voting and the other is soft voting. Hard voting is used for categorical target values. It will consider the mean of the data from the array of results. The main advantage of this method is that we can select any kind of algorithm and combine them.

Evaluation and Reflection

After building the model by applying an algorithm, it is mandated to check the performance with suitable evaluation metrics. Evaluation metrics are the parameters/measures of the model in terms of the performance of the model. In the process of that the most useful thing is the confusion matrix of the tested dataset. In general, a confusion matrix is a 2 by 2 matrix consisting of the numerical values and the description of those values are below.

True Positive (TP) is the number of predictions where the classifier correctly predicts the positive class as positive.

True Neutral (TNe) is the number of predictions where the classifier correctly predicts the Neutral class as Neutral.

True Negative (TN) is the number of predictions where the classifier correctly predicts the negative class as negative.

False Positive (FP) is the number of predictions where the classifier incorrectly predicts the negative or neutral class as positive.

False Neutral (FNe) is the number of predictions where the classifier incorrectly predicts the negative or positive class as Neutral.

False Negative (FN) is the number of predictions where the classifier incorrectly predicts the positive or neutral class as negative.

Table 1

Confusion matrix for multi-class ML

Predicted Values	**Actual Values**		
	Positive	**Neutral**	**Negative**
Positive	TP	FNe	FN
Neutral	FP	TNe	FN
Negative	FP	FNe	TN

Note. Confusion matrix formulas for multi-class machine learning models, obtained from *Confusion matrix for your multi-class machine learning model, by* Mohajon, J. (2021, July 24), Medium. https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826, copyright by Medium 2021.

Accuracy

Accuracy is the measure of correctness in the prediction of the target value. In other words it is the ratio of correctly predicted samples with total number of samples. It is a basic measure of performance of a model.

Formula

$$Accuracy = \frac{TP + TNe + TN}{Total\ no.\ of\ samples}$$

Misclassification Rate

This is a measure which is completely opposite to accuracy. I.e., the ratio of incorrectly predicted samples with the total number of samples.

Formula

$$Misclassification\ Rate = \frac{FP + FNe + FN}{Total\ no.\ of\ samples} = 1 - accuracy$$

Precision

It is the fraction of predictions as a positive class that were actually positive. Similarly with the other two classes neutral and negative.

$$Precision\ for\ positive\ class\ = \frac{TP}{TP + FP}$$

Recall

It is the ratio of positive samples that were correctly predicted as positive by the classifier. It is also known as True Positive Rate (TPR). Similarly there will be a True Negative Rate (TNR), True Neutral Rate (TNeR).

$$Recall\ for\ positive\ class\ = \frac{TP}{TP + FN}$$

F1-Score

F1-score is a combination of precision and recall. It is calculated as a harmonic mean of them. As the values of recall and precision increases the value of f1-score also increases.

$$F1\ -\ score\ for\ positive\ class\ =\ 2\ X \frac{precision\ X\ recall}{precision + recall}$$

Results

The above models are applied on the training data and developed models. After developing those models, evaluate them using evaluation metrics which are mentioned earlier. The results of all those are listed in the table below.

Table 2

Model Evaluation and Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Naive Bayes	70%	70%	71%	63%
Logistic Regression	77%	73%	77%	64%
KNN	86%	88%	86%	85%
Decision Tree	85%	85%	85%	85%
Random Forest	**89%**	**90%**	**89%**	**89%**
XGBoost	77%	75%	77%	73%
Voting Classifier	**89%**	**90%**	**89%**	**89%**

Note. The highlighted models have performed better than others

As shown in the above table, each model is listed with the accuracy and evolution matrics results. Naive Bayes is the model which performed with least performance accuracy of 70%. F1-score for that model is 63% which we felt is not up to the mark. Then we applied another model which is logistic regression. Performance of this model is better than naive bayes but not significantly. The accuracy of the model is 77% but F1-score is the same as naive bayes that is 64% therefore the overall model performance is almost similar with naive bayes. We applied the k-nearest neighbor algorithm, taking the k parameter as 1, where we can see the overall improvement in the performance in accuracy and other evaluation metrics. The accuracy of that model is 86% which is better than the previous two models. F1 score is

also far better with 85%. We also applied a decision tree algorithm where the performance of this model is very similar to the knn model.

In addition to these models, we applied a variety of ensemble learning methods on the dataset. In that, the first applied model is random forest. The developed model is with the best accuracy until now with 89%. Even the F1-score of this model is 89% which is better than any other applied models until now. The second applied ensemble learning method is XGBoost. Surprisingly the performance of this model is lower than the random forest with accuracy 77% and F1-score is 73%. Even though it is ensemble technique it is lacking in the performance. The last but not least, voting classifier is applied and develops a model with accuracy 89% and f1-score 89%. As mentioned previously we included three base learners i.e., K-nearest neighbor, Decision tree and logistic regression. As those algorithms are diverse the accuracy of models increases. But voting classifier and random forest are with similar performance. So, when we are considering other factors like resources needed, time factor, complexity, random forest is better than this voting classifier method even though both are same in performance.

Conclusion

Ratings and reviews have become an indispensable part of everyday lives; whether we go shopping, buying online, or eat at a restaurant, we generally trust the reviews beforehand to make the best decision. For this reason, sentiment analysis of drug evaluations was investigated in this project in order to develop a decision support system employing several varieties of machine learning classifiers, such as Naive Bayes, Logistic Regression, KNN, and different ensemble techniques like Random Forest, XGBoost and Max Voting, which were applied after the TF-IDF vectorizer.

From our model evaluation phase, we can conclude that out of all the models trained, Random forest and Max Voting Classifier have performed the best with an accuracy score of 89%. Future research will consider multiple strategies to use different n-gram values, and optimize the models. Additionally, use the Deep Learning models for classification. We would also like to create an end to end pipeline for drug classification and review in real time. Also, manage an interface to help the doctors in selecting the best drugs based on the reviews and ratings from the patients.

References

Gräßer, F., Kallumadi, S., Malberg, H., & Zaunseder, S. (2018). Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. *Proceedings of the 2018 International Conference on Digital Health*. https://doi.org/10.1145/3194658.3194677

Han, Y., Liu, M., & Jing, W. (2020). Aspect-level drug reviews sentiment analysis based on double bigru and knowledge transfer. *IEEE Access*, *8*, 21314–21325. https://doi.org/10.1109/access.2020.2969473

Mohajon, J. (2021, July 24). *Confusion matrix for your multi-class machine learning model*. Medium. Retrieved May 17, 2022, from https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

Na, J.-C. K. (2015, March 13). *Sentiment analysis of user-generated content on drug review websites*. Journal of Information Science Theory and Practice. Retrieved May 13, 2022, from https://doi.org/10.1633/JISTaP.2015.3.1.1

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science, 181, 526-534. https://doi.org/10.1016/j.procs.2021.01.199

Tare, R. (2021, March 29). *#1 : CRISP DM framework*. Ruchareads. Retrieved from https://ruchareads.wordpress.com/2021/03/29/1-crisp-dm-framework/

Vijayaraghavan, S., & Basu, D. (2020, March 21). *Sentiment Analysis in drug reviews using supervised machine learning algorithms*. arXiv.org. Retrieved May 13, 2022, from https://doi.org/10.48550/arXiv.2003.11643

Wowczko, I.A. (2015). A Case Study of Evaluating Job Readiness with Data Mining Tools and CRISP-DM Methodology. International Journal for Infonomics, 8, 1066-1070. https://www.semanticscholar.org/paper/A-Case-Study-of-Evaluating-Job-Readinesswith-Data-Wowczko/36771993f16b665c7a1c66f503a5c7d996b95803/figure/4.