Drugs Sentiment Analysis Using R

Somanshu Gupta
AID:A20499077
sgupta87@hawk.iit.edu

Kajal Dalai
AID:A20528697
kdalai@hawk.iit.edu

Pranjal Patil
AID:A20546612
ppatil22@hawk.iit.edu

December 4, 2023

1 Problem Statement

Social web contains a large amount of information with user sentiment and opinions across different fields. For example, drugs.com provides users' textual review and numeric ratings of drugs. However, text reviews may not always be consistent with the numeric ratings. Since covid-19 pandemic has shown up, shortage of healthcare workers, lack of medicines is at its peak. Due to unavailability , individuals started taking medicines without appropriate consultation making entire medical fraternity in distress. We aim to answer following questions.

- How to use Sentiment Analysis to recommend the drugs?
- To discover the most influential age group and gender of the target audience for a chosen drug?
- What is the emotional inclination of users towards a chosen drug ?

2 Introduction

There has been an expansion in the number of individuals worried about their well-being and finding a diagnosis online. Internet has opened up new unique pathways to obtain information about consumers' drug reviews through websites on drug reviews. Such reviews have contained an extensive amount of user sentiment related to a particular condition which could be used to detect the side effects and efficacy of the drugs. User reviews of drugs in online forms are unconventional and generally most reviewers lack of medical knowledge which may be barriers for good and reliable extraction of information. Also numerical rating can be misleading in quantifying a sentiment as biases can lead to different understanding of why constituted a high score versus what constituted a low score. Web based reviews can be a good source of information but ratings could be biased , For example , addictive drugs may be highly rated in comparison to other drugs which treated the same condition. The drug reviews are much more complicated as they contain medical terminologies like chemical used in drugs, medical health conditions etc. In this project, we will be evaluating the feasibility of leveraging machine learning and natural language processing to classify user ratings based on their textual review to identify the locations of contingency and model will also be used to identify overtly positive or negative scores. These scores are user ratings which was incorrectly classified by the model to identify some generalizations. We will also try to use computational technique to highlight the phrases in the text that have positive or negative impact on the classification results.

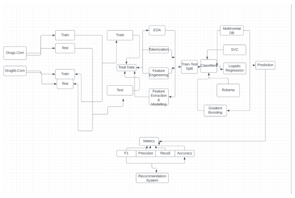
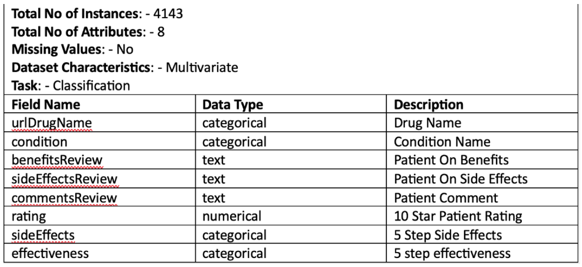


Figure 1: **Drugs Sentiment Analysis System Workflow**

3 Methods

3.1 Dataset Overview

We obtained the dataset from the UCI Machine Learning Repository. These instances were collected [1] from Drugs.com and Druglib.com , the dataset consists of user drug reviews, drug names, related medical conditions, and a 10-point rating. In total, the dataset consists of 215,063 instances and 9 attributes.



Total No of Instances: - 4143
Total No of Attributes: - 8
Missing Values: - No
Dataset Characteristics: - Multivariate
Task: - Classification

Field Name	Data Type	Description
urlDrugName	categorical	Drug Name
condition	categorical	Condition Name
benefitsReview	text	Patient On Benefits
sideEffectsReview	text	Patient On Side Effects
commentsReview	text	Patient Comment
rating	numerical	10 Star Patient Rating
sideEffects	categorical	5 Step Side Effects
effectiveness	categorical	5 step effectiveness

Figure 2: **Dataset Attributes**

3.2 Proposed Methodology

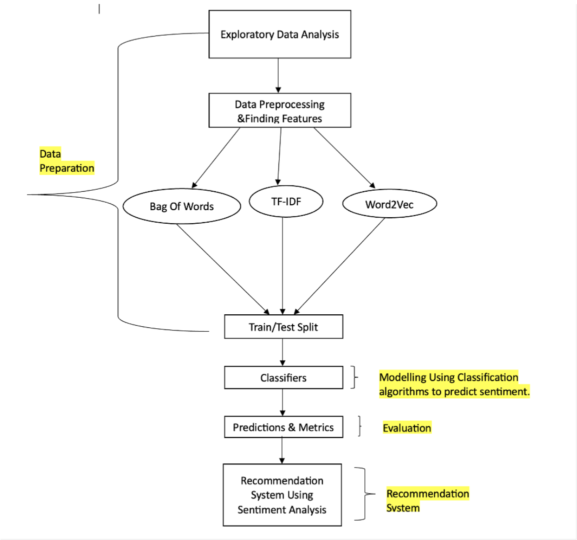

Exploratory Data Analysis
Data Preprocessing &Finding Features
Data Preparation
Bag Of Words
TF-IDF
Word2Vec
Train/Test Split
Classifiers
Modelling Using Classification algorithms to predict sentiment.
Predictions & Metrics
Evaluation
Recommendation System Using Sentiment Analysis
Recommendation System

Figure 3: **Execution Steps**

4 Exploratory Data Analysis

4.1 Data Cleaning

- Removed erroneous text and NA from review field.

- Renamed Column urldrugname to drugname for better readability.

- Dropped following fields X, benefitsReview,sideEffectsReview,commentsReview, sideEffects,effectiveness.

- Removing erroneous words from reviews which might not help in prediction.

4.2 Data Visualization

4.2.1 Drugs Rating Count

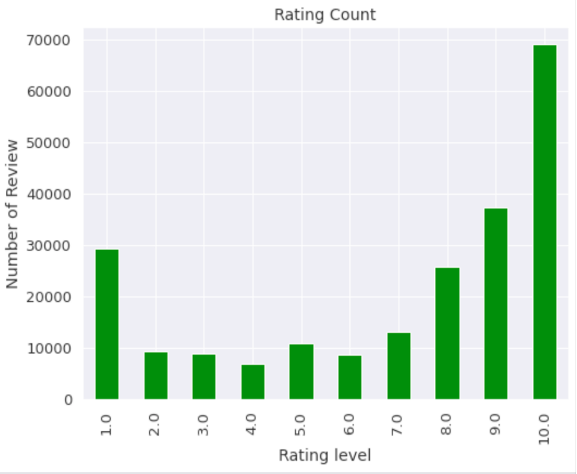


Figure 4: **Drug Rating Count Based On No. Of Reviews**

[1] https://archive.ics.uci.edu/dataset/461/drug+review+dataset+druglib+com

4.2.2 Sentiments Rating Distribution

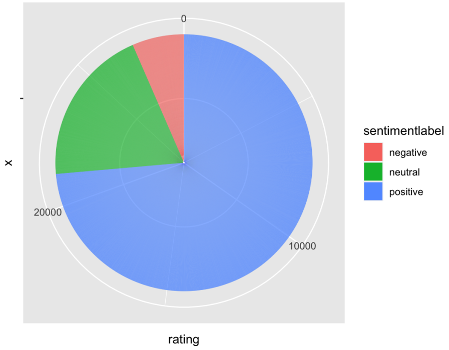


Figure 5: **Positive sentiments ratings are high followed by negative sentiments and neutral**

4.2.3 Most Common Drugs

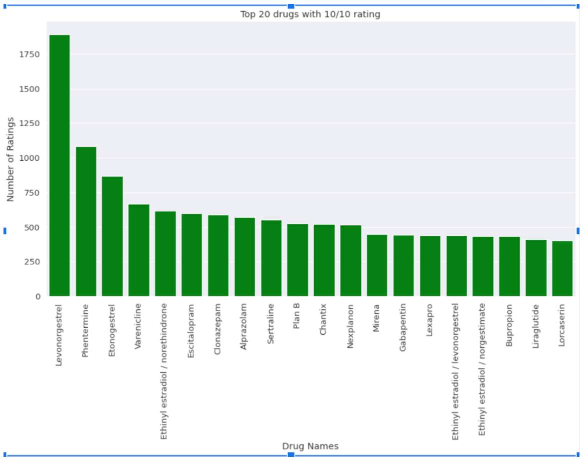


Figure 6: **Top 20 conditions that have maximum number of drugs available**

4.2.4 Most Common Conditions

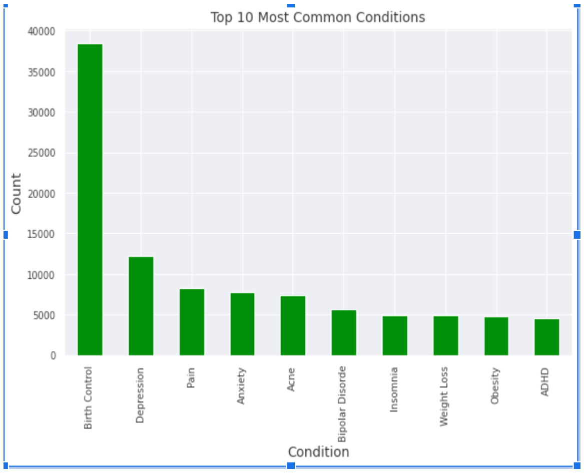


Figure 7: **Birth control is the most common condition in the review followed by depression.**

4.2.5 Rating Distribution

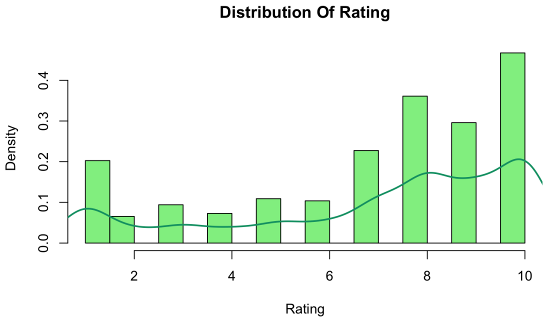


Figure 8: **Distribution Of Rating With Density**

5 Data Preprocessing Results

We build the corpus for reviews field and performed below steps to clean the raw data.

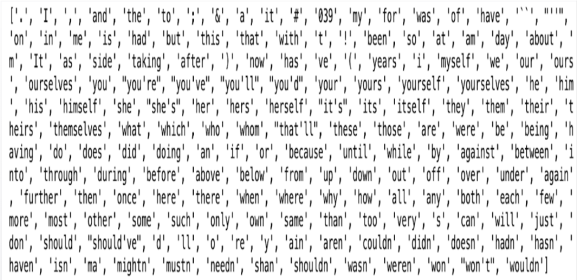
- Removed HTML tags, special characters, punctuations, quotes, URLs, etc.

- Cleaned reviews were lowercased to avoid duplication, and tokenization was performed for converting the texts into small pieces called tokens.

- Dropped following fields X, benefitsReview,sideEffectsReview,commentsReview, sideEffects,effectiveness.

- Stopwords like ’a’, ’were’, ’with’ were removed from corpus.

5.0.1 Data Corpus



```
['.', 'I', ',', 'and', 'the', 'to', ';', '&', 'a', 'it', '#', '039', 'my', 'for', 'was', 'of', 'have', '``', "''",
'on', 'in', 'me', 'is', 'had', 'but', 'this', 'that', 'with', 't', '!', 'been', 'so', 'at', 'am', 'day', 'about', '
m', 'It', 'as', 'side', 'taking', 'after', ')', 'now', 'has', 've', '(', 'years', 'i', 'myself', 'we', 'our', 'ours
', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him
', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', "it's", 'its', 'itself', 'they', 'them', 'their', 't
heirs', 'themselves', 'what', 'which', 'who', 'whom', "that'll", 'these', 'those', 'are', 'were', 'be', 'being', 'h
aving', 'do', 'does', 'did', 'doing', 'an', 'if', 'or', 'because', 'until', 'while', 'by', 'against', 'between', 'i
nto', 'through', 'during', 'before', 'above', 'below', 'from', 'up', 'down', 'out', 'off', 'over', 'under', 'again'
, 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', '
more', 'most', 'other', 'some', 'such', 'only', 'own', 'same', 'than', 'too', 'very', 's', 'can', 'will', 'just', '
don', 'should', "should've", 'd', 'll', 'o', 're', 'y', 'ain', 'aren', 'couldn', 'didn', 'doesn', 'hadn', 'hasn', '
haven', 'isn', 'ma', 'mightn', 'mustn', 'needn', 'shan', 'shouldn', 'wasn', 'weren', 'won', "won't", 'wouldn']
```

Figure 9: **Corpus Generated From Reviews**

5.0.2 Words Frequency

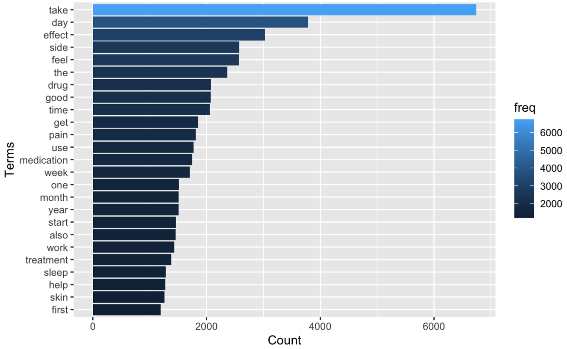


Figure 10: **Top Terms Based on Term Document Matrix**

5.0.3 Reviews Word Cloud 1

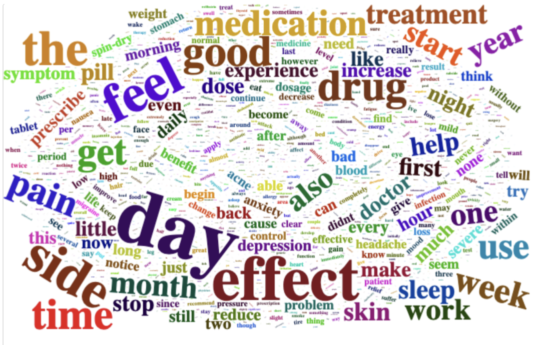


Figure 11: **Tokenized Word Cloud 1 For Reviews**

5.0.4 Reviews Word Cloud 2

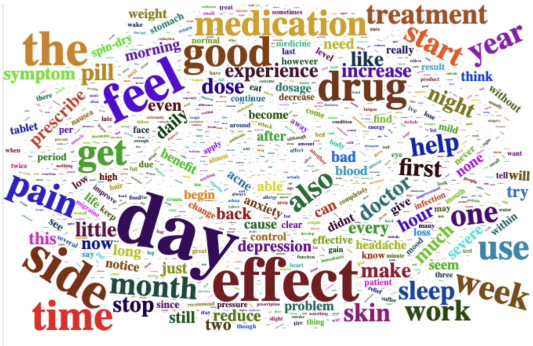


Figure 12: **words are daily words like ”first”,”given”, which are not predictive for rating,some words looks predictive for rating like ”lim- ited”,”quick” medical words like ”antidepressants” and meaningless words like ”ve”**

5.0.5 Sentiments Scores

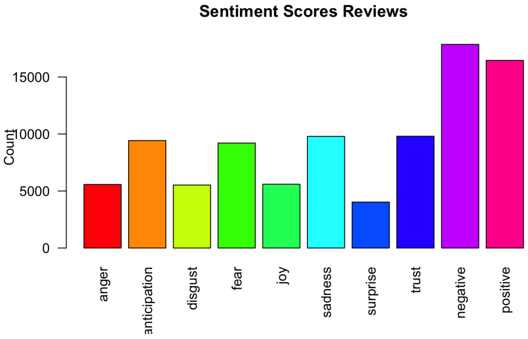


Figure 13: **Sentiment Scores Of Reviews**

6 Feature Generation

After data prepossessing , Machine learning algorithms can't work with text straightforwardly; it should be changed over into numerical format. We used below strategies for feature extraction

- Bag Of Words.
- TF-IDF.

6.0.1 Bag Of Words

It is used in natural language processing responsible for counting the number of times of all the tokens in review or document. Most fre-quent n words are related with a unique ID, wordID, where n varies from 3000 to 45000. Number of each word in the text are counted. We vectorized these numbers by their wordIDs, and put this list into feature vector. We counted uni-gram, bi-gram, tri-gram.

6.1 Union-Gram Results

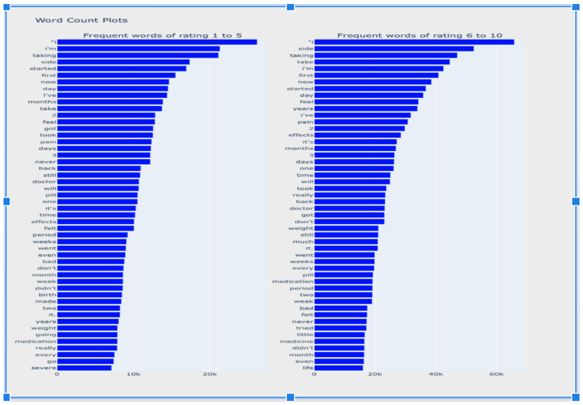


Figure 14: **Combination Of 1 Word**

6.2 Bi-Gram Results

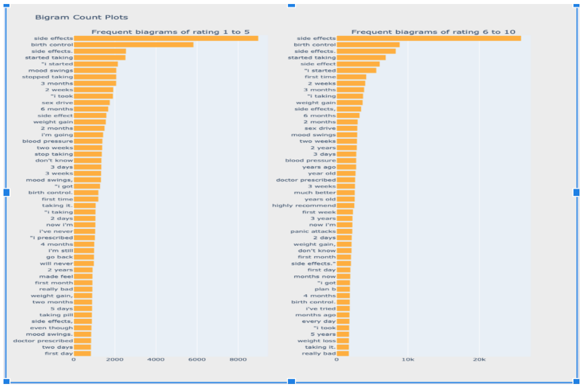
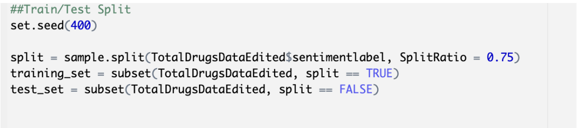


Figure 15: **Combination Of Multiple Words**

6.2.1 TF-IDF

Words are offered with weight not count. The principle was to give low importance to the terms that often appear in the dataset, which implies TF-IDF estimates relevance, not a recurrence. Term frequency (TF) can be called the likelihood of locating a word in a document.

7 Train/Test Split



```
##Train/Test Split
set.seed(400)

split = sample.split(TotalDrugsDataEdited$sentimentlabel, SplitRatio = 0.75)
training_set = subset(TotalDrugsDataEdited, split == TRUE)
test_set = subset(TotalDrugsDataEdited, split == FALSE)
```

Figure 16: **Split Ratio 0.75**

8 Classifiers

We introduced three models and also present comparisons between these models.A significant problem with this dataset is around 210K reviews, which takes substantial computational power. We selected those machine learning classification algorithms only that reduces the training time and give faster predictions.

- Multinomial Naive Bayes
- Support Vector Machine
- Logistic Regression

8.1 Classifier Performance Metrics

The Sentiment will be measured using 5 metrics. Here P is precision , R is recall.

- Precision

$$Precision = Tp/Tp + Fp \tag{1}$$

- Recall

$$Recall = Tp/Tp + Fn \tag{2}$$

- Accuracy

$$Accuracy = Tp + Tn/Tp + Tn + Fp + Fn \tag{3}$$

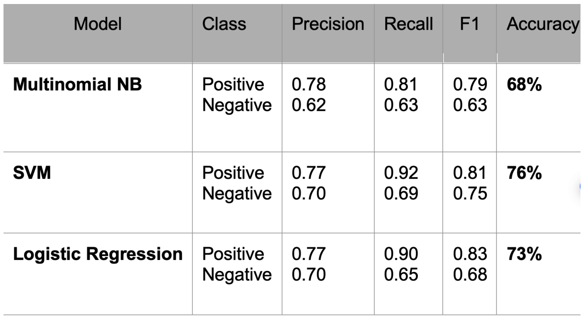
- F1 Score

$$F1Score = 2.(P.R/P + R) \tag{4}$$

9 Comparision Results

- Naive Bayes doesn't require as much training data. It handles both continuous and discrete data. It is highly scalable with the number of predictors and data points. It is fast and can be used to make real-time predictions.
- The advantages of logistic regressions are their properties that make them easy to interpret and deploy; while the weakness is that they only model relationships between dependent and independent variables that are linear.
- Linear SVM tries to finds the "best" margin (distance between the line and the support vectors) that separates the classes and this reduces the risk of error on the data, while logistic regression does not, instead it can have different decision boundaries with different weights that are near the optimal point.

9.0.1 Svm V/S Naive Bayes V/S Logistic Regression



Model	Class	Precision	Recall	F1	Accuracy
Multinomial NB	Positive Negative	0.78 0.62	0.81 0.63	0.79 0.63	**68%**
SVM	Positive Negative	0.77 0.70	0.92 0.69	0.81 0.75	**76%**
Logistic Regression	Positive Negative	0.77 0.70	0.90 0.65	0.83 0.68	**73%**

Figure 17: **Linear SVM Performed Well Followed By Logistic Regression and Multinomial Naive Bayes**

10 Existing Work

Some of them proposed an implementation for automatic sentiment classification of drug reviews employing fuzzy rough feature selection. Fuzzy-rough feature selection provided a means by which discrete or real-valued noisy data (or a mixture of both) can be effectively reduced without the need for usersupplied information. Their objective is to improve accessability of relevant information via modern social media. They presented a deep learning strategy based on the idea of a tweet's footprint to improve search and navigation in social media platforms and an efficient searching algorithm. Previous works employed were focussed on sentiment analysis to predict user satisfaction.

11 Summary

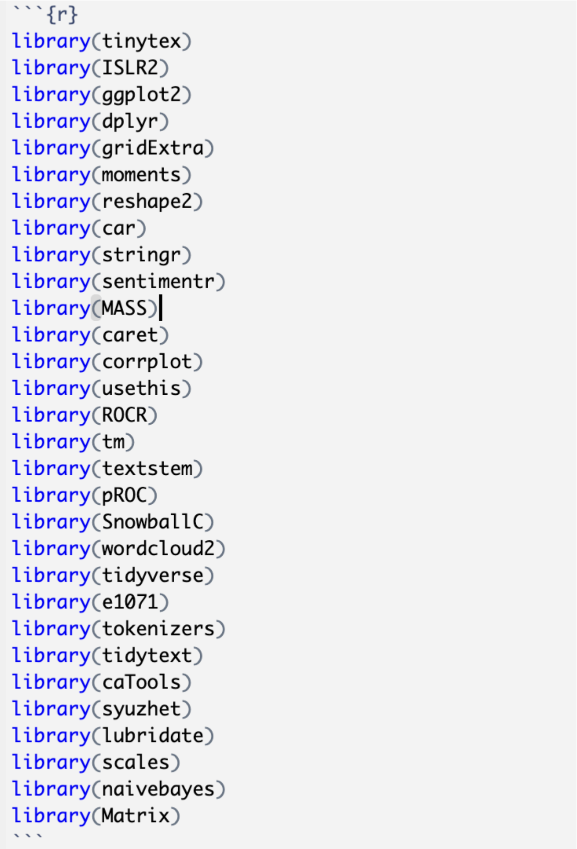
- We tried to predict the exact rating based on user reviews and other features such as useful count on the reviews.
- More Deep learning approach would be best and optimal way to reach to learn categories through its hidden layer architecture.
- The results procured from each of the four methods are good, yet that doesn't show that the recommender framework is ready for real-life applications.

12 Need Improvements

- Predicted results show that the difference between the positive and negative class metrics indicates that the training data should be appropriately balanced.
- Applying more deep learning approaches can further improve the solution.
- Proper hyperparameter optimization is also required for classification algorithms to improve the accuracy of the model.

13 SourceCode/Dependencies/ Libraries

13.1 Libraries



~~~
```{r}
library(tinytex)
library(ISLR2)
library(ggplot2)
library(dplyr)
library(gridExtra)
library(moments)
library(reshape2)
library(car)
library(stringr)
library(sentimentr)
library(MASS)
library(caret)
library(corrplot)
library(usethis)
library(ROCR)
library(tm)
library(textstem)
library(pROC)
library(SnowballC)
library(wordcloud2)
library(tidyverse)
library(e1071)
library(tokenizers)
library(tidytext)
library(caTools)
library(syuzhet)
library(lubridate)
library(scales)
library(naivebayes)
library(Matrix)
```
~~~

Figure 18: **Following Libraries Used**

13.2 Source Code

https://github.com/somanshumatata123/CSP571_Fall23_Final_Project

References

[1] Confusion matrix. https://www.sciencedirect.com/topics/engineering/confusion-matrix.

[2] Sentiment analysis. https://monkeylearn.com/sentiment-analysis/.

[3] Kamal Al-Barznji and Atanas Atanassov. Big data sentiment analysis using machine learning algorithms. In *Proceedings of 26th International Symposium" Control of Energy, Industrial and Ecological Systems, Bankia, Bulgaria*, 2018.

[4] Rezaul Haque, Saddam Hossain Laskar, Katura Gania Khushbu, Md Junayed Hasan, and Jia Uddin. Data-driven solution to identify sentiments from online drug reviews. *Computers*, 12(4):87, 2023.

[5] Surya Kallumadi and Felix Grer. Drug Review Dataset (Druglib.com). UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C55G6J.

[6] P Karthika, R Murugeswari, and R Manoranjithem. Sentiment analysis of social media network using random forest algorithm. In *2019 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS)*, pages 1–5. IEEE, 2019.

[7] Ioannis Korkontzelos, Azadeh Nikfarjam, Matthew Shardlow, Abeed Sarker, Sophia Ananiadou, and Graciela H Gonzalez. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of biomedical informatics*, 62:148–158, 2016.

[8] Akhil Shiju and Zhe He. Classifying drug ratings using user reviews with transformer-based language models. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pages 163–169. IEEE, 2022.

[9] Sairamvinay Vijayaraghavan and Debraj Basu. Sentiment analysis in drug reviews using supervised machine learning algorithms. *arXiv preprint arXiv:2003.11643*, 2020.