UNIVERSITY OF HERTFORDSHIRE
School of Physics, Engineering and Computer Science
M.Sc. Artificial Intelligence and Robotics

Interim Project Report
7COM1039-0109-2023
Artificial Intelligence and Robotics Master's Project
July 18, 2024

Sentiment analytical system to enhance patient satisfaction on drugs using a fine-tuned Python libraries-based model.

Name: Nneoma Nnorom
Student ID: 21073212
Supervisor: Dr. Kelechi Emerole

TABLE OF CONTENT

TABLE OF FIGURES

SECTION 1: OVERVIEW

1.1 Introduction:

In recent years, there has been a need for a more sufficient way of analysing patient feedback because the amount of patient reviews data increases as time unfolds. Sentiment analysis is a subfield of Natural Language Processing (NLP) that understands people's emotions, attitudes, appraisals, and opinions in unstructured text about topics, issues, entities, events, and products. Sentiment analysis, a critical field in natural language processing, enables the extraction of subjective information from textual data. In the context of healthcare, understanding patient sentiments expressed in drug reviews is invaluable for public health decision-making and patient care (Garg, S., 2021). Considering the various fields of Natural Language Processing (NLP), sentiment analysis can be seen to have evolved into a promising domain, which covers various sectors such as economics, politics, and healthcare, etc. Particularly in the pharmaceutical sector, sentiment analysis can play a vital role because of the ability to access large volumes of user-generated content which gathers detailed information of drug effectiveness and side effects. User sentiments expressed in drug reviews offer a wealth of information about their experiences and preferences. This data can significantly assist medical professionals in making informed decisions, particularly through enhanced monitoring of public health (Hameed et al., 2023).

1.2 Research Questions:

The research questions include:

1. Does leveraging Spacy and Textblob for preprocessing and sentiment analysis of the data increase the algorithm’s performance and accuracy and reduce the model's complexity? This is to find out if using Spacy, a python library used in data cleaning and preprocessing will increase the system performance and accuracy.

2. Was aspect-based sentiment analysis using machine learning able to classify and identify specific aspects of the drug reviews?

1.3 Research Objectives:

1. Develop a sentiment analysis model that utilizes the power of natural language processing to enhance patient satisfaction with drugs.
2. To design and implement an algorithm that uses Spacy Python libraries to preprocess data and understand the data, with the purpose of obtaining accurate analysable results.
3. To design and implement an algorithm that uses sentiment analysis on patient reviews on drugs to categorize the review into positive, negative, and neutral using the TextBlob python library.
4. To split the dataset into train and test using Sklearn.Train_test_split.
5. To design an algorithm that trains the naïve bayes classifier and Random Forest Classifier.
6. To classify the specific aspects of the patient reviews using naïve base.
7. To investigate and evaluate the accuracy, precision and performance of the proposed model.
8. Compare the performance and efficiency of the proposed model with that of other models.

9. To design a User Interface to enable interaction with the model.

1.4 Literature Review:

Recently, the healthcare sector has experienced a significant increase in patient-generated data, specifically in the form of online evaluations and self-reported experiences. These patient accounts offer insightful information about the effectiveness, adverse effects, and general satisfaction of patients with different drugs. With the goal of helping patients and healthcare professionals make informed decisions, researchers have created creative drug recommendation systems by utilising this abundance of real-world data. (Garg, 2021).

As a means to increase patient satisfaction, the use of sentiment analysis to glean insightful data from patient reviews and opinions about healthcare providers is summed up in the literature review. Sentiment analysis of drug reviews has been studied recently to increase patient happiness and better healthcare decision-making.

(Khan and Fatima, 2023) employed machine learning and deep learning algorithms to examine consumer perceptions of medication efficacy, side effects, and convenience of use. (Helae, Ebrahimi, and Alzhouri, 2022) evaluated patient satisfaction and forecasted medical issues based on reviews using random forest and LSTM models.

Sentiment analysis is essential for interpreting and comprehending patients' experiences with drugs, which aids in making informed decisions and improving public health. In this work, sentiment classification of drug reviews using machine learning techniques was investigated. The researcher extracted features using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm. Future initiatives for this research include examining drug efficacy analyses using other algorithms and looking at other aspects of drug evaluations for ways to improve natural language processing. (Hameed, Shaker and Khalaf,2023).

The researchers I have reviewed their works have made significant contributions in this area of research but there is a need to identify the aspects of patients' reviews and conditions. Thus, my proposed system employs Aspect-based sentiment analysis which identifies the various aspects of the patients' feedback. Thereby, solving a problem that previous researchers couldn't solve.

1.5 Methodology:

The methodology that will be used to carry out this project is the CRISP-DM, and this has many phases which includes:

1. Business Understanding: In this project, I focused on leveraging Natural language technique (Sentiment Analysis) to find the best means to ensure that patient's satisfaction on drugs is enhanced. This will be achieved by comparing and evaluating the results of the existing models with my proposed model.
2. Data Understanding: This is the second stage of the CRISP-DM where I understood the type of Data for this project. Having understood the project, I searched for datasets and discovered data from an open source (Kaggle) which I used for this project. This data contains the drug name, the condition, patient's reviews, etc.

3. Data Preparation: I carefully prepared the data by cleaning the dataset and ensuring that words that are not needed are erased. I lowered the words, carried out word tokenization and finally explore the data to ensure that it is properly cleaned and prepared because I will be training the dataset with libraries that will give a perfect result when the dataset is properly cleaned.
4. Modelling: At this stage, I applied sentiment analysis on the dataset using Textblob and this returned the polarity which would in turn be classified based on positive, negative, and neutral. I also designed a machine learning model trainer with a naïve bayes classifier which will train the dataset. I will also design an aspect-based algorithm which will classify the various aspects of these patients' reviews. I also trained Naïve Bayes and Random Forest classifiers on the train dataset.
5. Evaluation: This is the fifth stage of CRISP-DM Methodology. I will evaluate the performance of the model based on its accuracy and precision.
6. Deployment: At this stage, I will design a user interface and implement the model on it. This will enable users to easily work with the model.

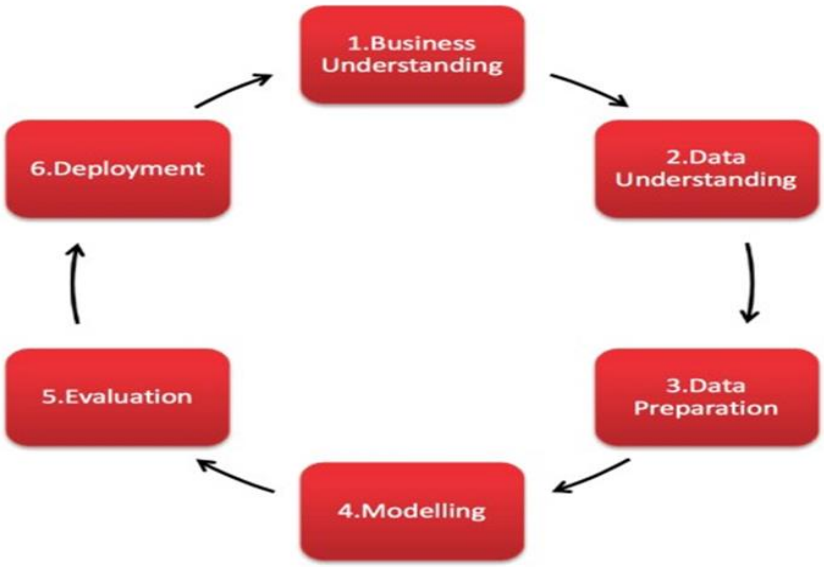


Figure 1.1: CRISP-DM Methodology (Source: https://www.sv-europe.com/crisp-dm-methodology/)

1.6 Practical Investigation:

The technical work involved in this research include:

1. Data Sourcing and Collection: Dataset of drug reviews named UCI ML Drug Review dataset curated from Kaggle (https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018/data) after I reviewed literature to guide my data sourcing process. This dataset is a patient feedback and reviews on drugs. The dataset has 7 features which includes: i) Unique ID which identifies each of the patients records. ii) Name of drug states the name of the drug. iii) Name of condition states the patient's illness. iv)Patients review defines the patient's feedback and emotions about the drug. v) 10-star patient rating vi) Date of review rating vii) Number of users who found review useful. See Appendix 1 for a screenshot of the dataset showing its features and some records. (Gräßer *et al.*, 2018).

The dataset has been split and uploaded with the names drugsComTrain_raw and drugsComTest. These datasets were uploaded using Panda's python package and read with the CSV read method. The datasets were merged for cleaning and preprocessing using the concat method in Panda's package. Go to Appendix 2. The result was stored in a variable which was further used throughout the project.

2. Data Cleaning and Preprocessing: The cleaning and preprocessing of the dataset. This involves the removal of noise, irrelevant information, stop words, and word tokenization. Spacy Python library was used to achieve this. Spacy Python package was first installed (pip install spacy), the spacy model was loaded and used in the preprocessing of the model. The words were lowered, null values were dropped, stop words and punctuation were also removed. Go to Appendix 3.
3. Sentiment Analysis and aspect analysis: This involves using the Textblob Python library to analyse and interpret the dataset. This involves the following steps:
    a) Performing of sentiment analysis: This analysis is performed on the dataset, which returns polarity and subjectivity, and these features are added to the dataset where the polarity is quantified as sentiment score and ranges between -1 to 1. These sentiment scores are further categorized into positive, negative or neutral.
    b) Categorization of the sentiment: This categorizes the sentiment scores (polarity) into positive, negative, and neutral classes.
    c) Aspect analysis: this involves extracting the various aspects of the patients' reviews based on the stated keyword. See Appendix 10.
4. Model Training: This involves training a Multinomial NB classifier and Random Forest classifier on the training data using scikit-learn. It involves the following steps:
    a) Feature extraction: Transforming of the dataset into vectors (a matrix of rows and columns) that Machine Learning algorithms can easily understand. This was achieved with the TfidfVectorizer method in Sklearn was used to transform the pre-processed patient text reviews into vectors.
    b) Splitting the dataset into training and testing sets. This is to use the training data to train the model. This was achieved with the train_test_split method in sklearn. This is to split the dependent and independent features into train and test dataset. The train data was used in the model training and the test data was used for model evaluation.
    c) Train Multinomial NB classifier on the training data. Multinomial NB is one of the naïve bayes text classification variants. Its distribution is based on vector parameters (number of features/size of vocabulary) and the probability (in text classification, the size of the vocabulary) and the probability of features appearing in a sample belonging to each class. The parameters are estimated by a smoothed version of maximum likelihood.
    d) Train Random Forest classifier on the training data: Random Forest is an ensemble method that improves classification accuracy by combining the prediction of multiple decision trees. It is trained by using labelled data. Each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. Its purpose is to decrease the variance of the forest estimator. Just as other classifiers, forest classifiers are fitted with two arrays: a sparse or dense array X of shape which holds the training samples, and an array Y of shape which holds the target values for the training samples.

See Appendix 12 for model training.

5. Model Evaluation: This involves assessment of the performance of the model by evaluating the classifier using the evaluation metrics accuracy, precision, recall and F1-score.

6. Deployment: I will be integrating this algorithm with a user interface (web) in other for users to easily work with the model.

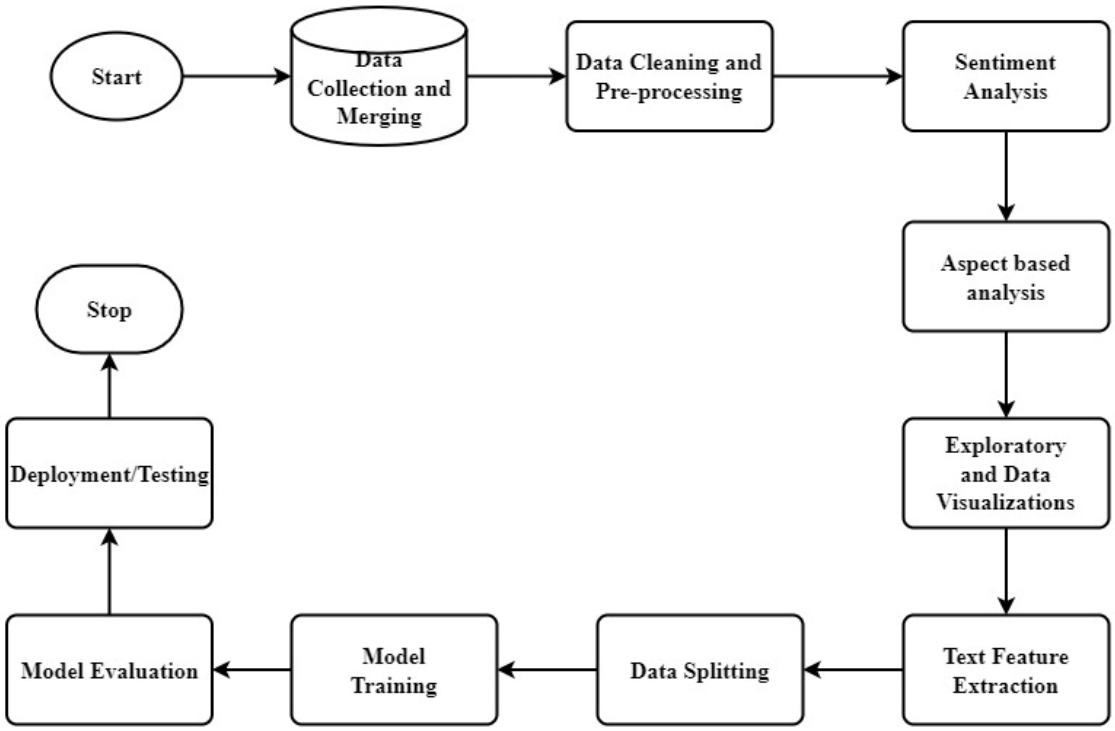


Figure 1.2: Block Diagram of the new system

1.7 Project Milestones:

1. Datasets: A pre-processed dataset of about 215062 records of patients reviews on drugs. Go to Appendix 1 for the sample.
2. Sentiment analysis and aspect extraction output: Additional features of the review's polarity (sentiment score) and subjectivity score, and a further added features of its sentiment which is categorized based on the polarity of the patient review. Also shows additional features of the aspect and its classification.
3. A distribution chart to explore the dataset with the following steps:
    i. Displays the average sentiment score by drug.
    ii. Displays the distribution of sentiment scores.
    iii. Displays the distribution of sentiment categories.
4. Trained Model of the classifier.
5. Evaluation: The following metrics will be printed out:
    i. Accuracy
    ii. Precision, etc.
6. Project reports:

i. Interim Report: a project report with details on the project progress and planned work.
ii. Final Report: a comprehensive report including the following:
a) Methodology
b) Results of the model evaluation
c) Limitations and recommendations for future work.

1.8 Tools and Techniques: The following are the tools and techniques used.

1. Data collection:
   a) Tool: Kaggle (open source). Kaggle is a community platform that provides a collaborative environment that hosts datasets, and notebooks. I obtained the dataset from a 5 years ago hosted project on the platform.
   b) Technique: I researched datasets related to patients' reviews and drug satisfaction. Reviewed the dataset descriptions and features.
2. Data cleaning and preprocessing:
   a) Tool: I used the following python libraries:
      i. Pandas: Pandas is a Python library that is designed for data manipulation and analysis. It provides data structures for working with numerical tables. It's also a powerful tool for handling data in Python when dealing with structured data like spreadsheets or databases. It was installed by running "pip install Pandas" on the terminal. The installed package was imported as np which was then used in handling of the dataset such as loading/reading of the dataset. Version 2.2.2.
      ii. Spacy: spacy is an open-source library for advanced Natural Language Processing (NLP) in Python. It was used to preprocess the dataset. It was installed by running "pip install spacy" on the terminal. The installed package was imported and loaded for further use such as tokenization, lowering of words, etc. Spacy Version 3.7.4.
   b) Technique: I pre-processed the dataset using spacy to lower the words, remove punctuation, tokenize the words, and remove stop words. Go to Appendix 3.
3. Sentiment analysis and aspect extraction:
   a) Tool:
      i. Textblob: TextBlob is a Python library for processing textual data. It was installed by running "pip install Textblob" on the terminal. It returns two metrics: Polarity and Subjectivity, where polarity is a value between -1.0 and 1.0, -1.0 indicates very negative sentiment, 1.0 indicates very positive sentiment, and 0 indicates neutral sentiment. Subjectivity is a value between 0.0 and 1.0, where 0.0 is very objective and 1.0 is very subjective. Textblob Version 0.18.0.post0.
      ii. Spacy: Spacy is an open-source library for advanced Natural Language Processing (NLP) in Python. It was used to preprocess the dataset. It was installed by running "pip install spaCy" on the terminal. The installed package was imported for extraction of aspects. Spacy Version 3.7.4.
   b) Technique: I performed sentiment analysis using text blob, this returned the sentiment score (known as Polarity). Polarity in this context means the measurement of sentiment expressed in a text and this ranges from -1 to 1. I categorized the sentiment score into positive, negative and neutral classes. I will

identify and classify the various specific aspects. Go to Appendix 10 and Appendix 11.

4. Exploratory data analysis/ Data Visualization:
   a) Tool:
      i. Seaborn: is a Python library that is used for statistical plotting. It helps to explore and understand data. It also builds on top of matplotlib. It was installed by running "pip install seaborn on the terminal and imported as sns which was used for the statistical plotting. Version installed was 0.13.2.
      ii. Matplotlib: it is a python library that is used to create interactive visualizations. It provides tools to represent data graphically. It creates various types of plots, charts, and graphs. Matplotlib version 3.9.0.
   b) Technique: It displays the distribution of sentiment scores, sentiment categories. I plotted a histogram distribution to display the average sentiment score by drug, I plotted a histogram distribution to display the sentiment scores, and I plotted a histogram distribution to display the sentiment categories. Go to Appendix 6, 7, 8, and 9.
5. Model Training: to train a Naïve Bayes classifier and Random Forest classifier on the training dataset:
   a) Tool: Scikit-learn is a Python library used for machine learning tasks such as model evaluation and sentiment analysis. It provides tools to build and evaluate machine learning classifiers and sentiment analysis models. Such as naïve bayes, model_selection, feature extraction, etc. It was installed by running "pip install Textblob" on the terminal. Scikit-learn Version 1.5.0.
   b) Technique: I carried out text feature extraction using TF-IDF (Term Frequency-Inverse Document Frequency) which is part of the scikit-learn library that converts these patients feedback into a matrix of TF-IDF (Term Frequency-Inverse Document Frequency) features. After the text feature extraction, I split the dataset into training and testing sets. I trained Multinomial NB which is a Naïve Bayes classifier on the train data. I also trained a Random Forest Classifier on the train set to compare the classifiers. Go to Appendix 12.
6. Model Evaluation:
   a) Tool: Scikit-learn is a Python library used for machine learning tasks such as model evaluation and sentiment analysis. It provides tools to build and evaluate sentiment analysis models. Such as naïve bayes, model_selection, feature extraction, etc. It was installed by running "pip install Textblob" on the terminal. Version 1.5.0.
   b) Technique: I will evaluate the model and print the accuracy.

1.9 Consideration of Ethical, Legal, Professional and Social Issues:

I searched and selected a dataset that was relevant and ethically sourced. The sequential procedures involved searching open sources such as Kaggle for datasets relating to patient reviews and drug satisfaction, reviewing dataset descriptions, and ensuring they are ethically sourced with patient consent, downloading the dataset, and examining the data fields for an understanding of its structure and content. The dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Which allows for the sharing and adaptation of the datasets for any purpose, provided that the appropriate credit is given. Thus, the project will not need ethical approval because

the study does not involve human participants. On the course of my research, I discovered that the dataset license is not on Kaggle, I had to send an email to the dataset owner requesting a guarantee that the dataset is compliant with the research protocols of my university (See Appendix 13). The owner on Kaggle sent me the link to where the license is (https://archive.ics.uci.edu/dataset/462/drug+review+dataset+drugs+com). I have also attached the screenshot of the license details in Appendix 14. Therefore, there is no need to obtain ethical clearance from the relevant authority. ECDA (Ethics Committee with Delegated Authority which functions as a sub-committee of the Ethics Committee) in the University of Hertfordshire.

1.10 Meeting with my supervisor:

This sub-section involves the discussions I had with my supervisor during some of our weekly meetings and the action points I have implemented. Some of the action points are:

a) Addition of background and introduction to the DPP. This was to separate the aim which I have written together with the introduction. This was to easily introduce the topic with a background on the project topic.
b) Rephrasing of the research questions. The phrase “To what extent” does not clearly define the research questions, thereby giving it a different meaning. I rephrased it as my supervisor advised.
c) Providing more details on the problem statement. I gave more details on why the existing algorithms are complex to perform. This was to clearly state the problem I am proposing to solve.
d) Rephrasing of some of the objectives and clearly stating the machine learning method I will be using to write the model. This was to clearly state the objectives of my project.
e) Addition of resources. The need to identify the resources needed to carry out my project as it will help me to get every material required ready and available.
f) Addition of how I would access ethical and social issues of my application. This was to ensure that I am fully aware of the ethical and social issues of my application and how to access it. These corrections were completed before the final submission on canvas. Research was also carried out on the existing work on applying sentiment analysis to patient satisfaction on drugs.

SECTION 2: PROGRESS TO DATE

2.1 Project Setup and Development Environment setup: I setup the development environment which is the Python environment on Visual Studio Code (VS Code) by installing the necessary packages such spacy, textblob, sklearn, matplotlib, pandas, etc,. and importing the required libraries. The python version used is 3.11.9. I also outlined the project plan, project scope, research questions, and project objectivities.

The deliverable of this task is a well-set python integrated development environment with the necessary python libraries installed.

2.2 Data Collection: I obtained the dataset of drug reviews named UCI ML Drug Review dataset curated from Kaggle which is an open source community-driven platform that hosts datasets and notebook (https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018/data). This dataset is a drug patient review data which is to be used for this project. (Gräßer *et al.*, 2018).

The deliverable of this task is a drug review dataset.

2.3 Data cleaning and preprocessing: the dataset was cleaned and pre-processed using Spacy which is a python library to remove stop words, perform word tokenization, lower the words, and to get the dataset ready for further analysis and training.

The deliverables of this task are:

a) pre-processed dataset which is ready to be used for model training and applying sentiment analysis.
b) An added feature of a cleaned feedback which will be used for analysis and model training.

2.4 Exploratory data analysis and sentiment analysis: The dataset was explored after performing sentiment analysis. Statistical description was also carried out in the EDA, this was to get a sense of the data before performing sentiment analysis. Example: displays the distribution of sentiment scores, sentiment categories. After performing the analysis, I plotted a histogram distribution to display the average sentiment score by drug, I plotted a histogram distribution to display the sentiment scores, and I plotted a histogram distribution to display the sentiment categories. shows well detailed information about the dataset. Sentiment analysis was performed on the dataset in two stages:

a) Sentiment analysis was performed, and it returned polarity and subjectivity. The polarity is also known as sentiment score.
b) The sentiment score was categorized into positive, negative, and neutral.

The sentiment analysis on this dataset uses Text blob python library to understand and interpret the patient's feedback on drugs.

The deliverable of this task includes:

i. Polarity and subjectivity features added to the dataset.
ii. Sentiment feature of either positive or negative or neutral added to the dataset.

2.5 Model training and Sentiment classification: The following task were performed during the model training:

a) I did feature extraction on the data to transform the text into vectors.
b) The dataset was split into train and test sets using train_test_split from sklearn .model_selection.
c) I trained the Multinomial NB on the train dataset using sklearn.naive_bayes.
d) I also trained Random Forest Classifier on the train dataset using sklearn. ensemble

SECTION 3: PLANNED WORK

3.1 Major tasks to be completed:

3.1.1 Aspect based sentiment analysis (Week 7: ongoing): The task is to understand various aspects of the patients review with the keywords. It is performed with the following steps:

a) Extraction of the aspects using Spacy python library.
b) Classification of the extracted aspects using the Text blob library. Go to Appendix 15.

The deliverable is an additional feature of the extracted aspects and the classified aspects. See Appendix 16.

3.1.2 Data visualization (Week 8): This is the continuation of the exploratory data. The analysed data will be visualized using the seaborn and matplotlib python libraries to show the relationships and distribution between the output features of the sentiment analysis, its deliverables will be:

a) Histogram chart distribution of the sentiment scores (polarity)
b) Histogram chart distribution of the sentiment categories (sentiment)

3.1.3 Evaluation of the classifier (Week 9): The trained classifier model will be evaluated using the Scikit-learn and its deliverables will be:

a) The accuracy results.
b) Confusion matrix
c) Precision result, etc.

3.1.4 Model integration with the User-friendly interface environment (Week 10):

A user-friendly environment will be designed and interfaced with the model. The deliverable will be a user-friendly web app that will show the designed model.

3.1.5 Documentation and final reporting presentation (Week 12):

Compile all my findings, literature review, methodology, results, and discussions into a thesis document.
The Deliverable will be a well-documented thesis outlining the project's objectives, methodologies, results, recommendation, and suggestions for future research.

By following these tasks and target completion dates, the project aims to build a fine- tuned model that will enhance patient satisfaction on drugs and also recommend on the best model to use.

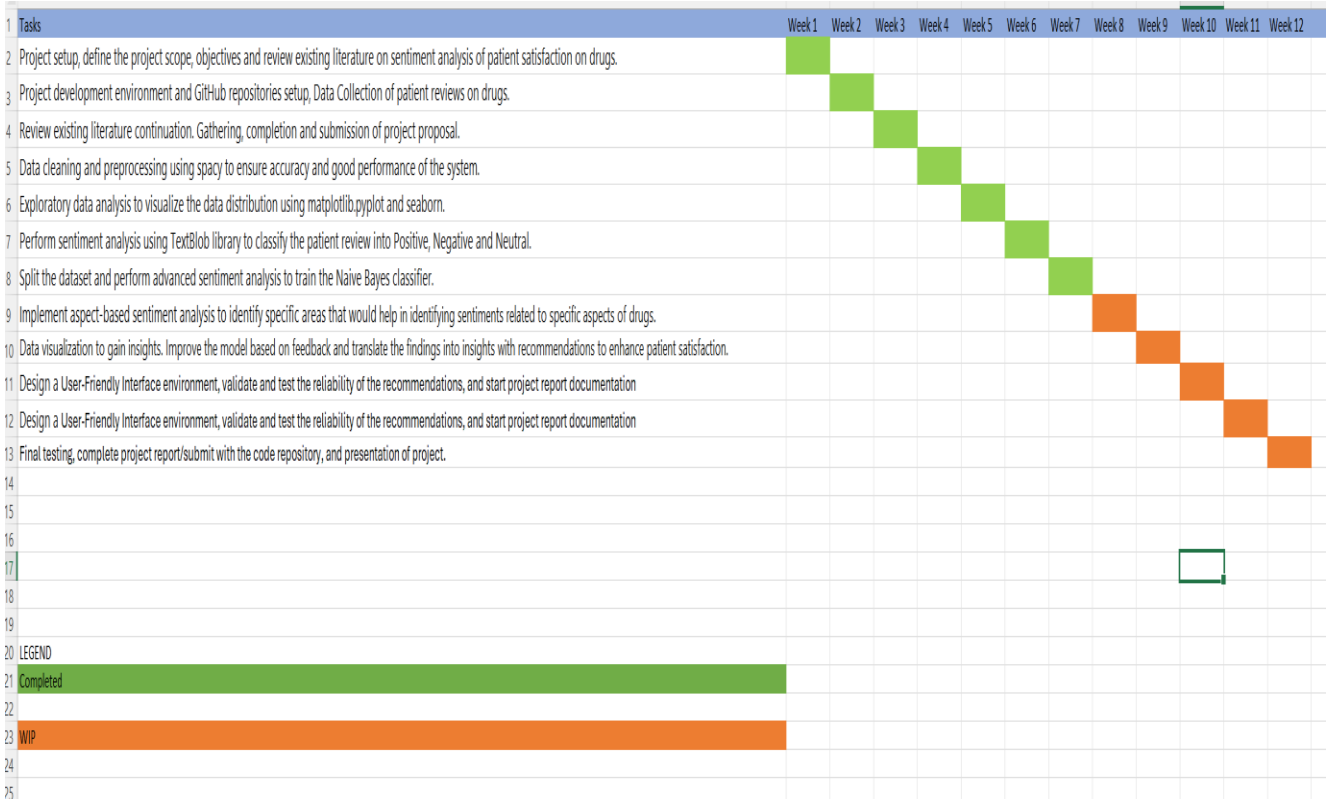
3.2 Testing

1. After loading the dataset with the pandas' package, the dataset was explored by checking its shape using shape method. See Appendix 2.

2. The result of the dataset merging was tested by calling the head, info and describe method to show the number of rows (datapoints) on the dataset. See Appendix 5 for some of the results.

3. After performing sentiment analysis, the output was viewed to test the result by running the info (), head () methods. These shows the added features. See Appendix 5.

3.3 Gantt chart showing the planned work:



BIBLIOGRAPHY

Gräßer, F. *et al.* (2018) 'Aspect-Based Sentiment Analysis of Drug Reviews Applying Cross-Domain and Cross-Data Learning,' *Digital Humanities Conference* [Preprint]. https://doi.org/10.1145/3194658.3194677.

Hameed, M.A.-A.A., Shaker, K. and Khalaf, H.A. (2023) 'Sentiment Classification of Drug Reviews Using Machine Learning Techniques,' *2023 15th International Conference on Developments in eSystems Engineering (DeSE)* [Preprint]. https://doi.org/10.1109/dese58274.2023.10099735.

Garg, S. (2021) 'Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning,' *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* [Preprint]. https://doi.org/10.1109/confluence51648.2021.9377188.

Khan, H.F. and Fatima, N. (2023) 'Analyzing Customer Sentiment in Drug Reviews Using Natural Language Processing,' *International Conference on Electronics* [Preprint]. https://doi.org/10.1109/iementech60402.2023.10423529.

Helae, M.Q.Y., Ebrahimi, D. and Alzhouri, F. (2022) 'Data Analytics in the pharmacology domain,' *International Journal of Big Data and Analytics in Healthcare*, 7(1), pp. 1–16. https://doi.org/10.4018/ijbdah.314229.

C, G. et al. (2023) 'Deep Learning Based Sentiment Analysis on Drug Reviews,' IEEE [Preprint]. https://doi.org/10.1109/icidea59866.2023.10295255.

Yadav, A. and Vishwakarma, D.K. (2020) 'A Weighted Text Representation framework for Sentiment Analysis of Medical Drug Reviews,' IEEE [Preprint]. https://doi.org/10.1109/bigmm50055.2020.00057.

Mishra, A., Malviya, A. and Aggarwal, S. (2015) 'Towards Automatic Pharmacovigilance: Analysing Patient Reviews and Sentiment on Oncological Drugs,' IEEE [Preprint]. https://doi.org/10.1109/icdmw.2015.230.

Haque, R. et al. (2023) 'Improving Drug Review Categorization Using Sentiment Analysis and Machine Learning,' IEEE [Preprint]. https://doi.org/10.1109/icdsns58469.2023.10245841.

Thorat, S. et al. (2022) 'Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning,' International Journal for Research in Applied Science and Engineering Technology, 10(11), pp. 1317–1322. https://doi.org/10.22214/ijraset.2022.47474.

Sreedhar, K.C. et al. (2024) 'Drug Recommendation System Based on Sentiment Analysis of Drug Reviews using Machine Learning,' International Journal for Multidisciplinary Research, 6(3). https://doi.org/10.36948/ijfmr.2024.v06i03.18767.

Wankhade, M., Rao, A.C.S. and Kulkarni, C. (2022) 'A survey on sentiment analysis methods, applications, and challenges,' Artificial Intelligence Review, 55(7), pp. 5731–5780. https://doi.org/10.1007/s10462-022-10144-1.

Garg, S. (2021) 'Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning,' International Conference on Cloud Computing, Data Science & Engineering (Confluence) (Pp. 175-181). IEEE. [Preprint]. https://doi.org/10.1109/confluence51648.2021.9377188.

Available at: https://www.sv-europe.com/crisp-dm-methodology/ [Accessed on 04/06/2024].

Available at: https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018 [Accessed on 01/05/2024].

APPENDICES

Appendix 1 – Sample of patient's reviews on drugs with the seven (7) features.

uniqueID	drugName	condition	review	rating	date	usefulCount
206461	Valsartan	Left Ventricular Dysfunction	"It has no side effe	9	20-May-12	27
95260	Guanfacine	ADHD	"My son is	8	27-Apr-10	192
92703	Lybrel	Birth Control	"I used to take	5	14-Dec-09	17
138000	Ortho Evra	Birth Control	"This is my first tir	8	03-Nov-15	10
35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has cor	9	27-Nov-16	37
155963	Cialis	Benign Prostatic Hyperplasia	"2nd day on 5mg s	2	28-Nov-15	43
165907	Levonorgestrel	Emergency Contraception	"He pulled out, bu	1	07-Mar-17	5
102654	Aripiprazole	Bipolar Disorde	"Abilify changed m	10	14-Mar-15	32
74811	Keppra	Epilepsy	" I Ve had nothing	1	09-Aug-16	11
48928	Ethinyl estradiol / levonor	Birth Control	"I had been on the	8	08-Dec-16	1
29607	Topiramate	Migraine Prevention	"I have been on th	9	01-Jan-15	19
75612	L-methylfolate	Depression	"I have taken anti-	10	09-Mar-17	54
191290	Pentasa	Crohn's Disease	"I had Crohn&#03	4	06-Jul-13	8
221320	Dextromethorphan	Cough	"Have a little bit o	4	07-Sep-17	1
98494	Nexplanon	Birth Control	"Started	3	07-Aug-14	10
81890	Liraglutide	Obesity	"I have been takin	9	19-Jan-17	20
48188	Trimethoprim	Urinary Tract Infection	"This drug worked	9	22-Sep-17	0
219869	Amitriptyline	ibromyalgia	"I've been t	9	15-Mar-17	39
212077	Lamotrigine	Bipolar Disorde	"I've been c	10	09-Nov-14	18
119705	Nilotinib	Chronic Myelogenous Leuke	"I have been on Ta	10	01-Sep-15	11
12372	Atripla	HIV Infection	"Spring of 2008 I v	8	09-Jul-10	11
231466	Trazodone	Insomnia	"I have insomnia, i	10	03-Apr-16	43
227020	Etonogestrel	Birth Control	"Nexplanon does i	9	11-Aug-14	11
41928	Etanercept	Rheumatoid Arthritis	"I live in Western A	10	16-Sep-17	4
213649	Tioconazole	Vaginal Yeast Infection	"Do not use the cr	1	17-Apr-17	7
51215	Azithromycin	Chlamydia Infection	"Was prescribed o	7	14-Dec-15	7
206180	Eflornithine	Hirsutism	"I'm writing	10	11-May-14	99
78563	Daytrana	ADHD	"Hi all, My son wh	10	12-Jan-17	11
132258	Ativan	Panic Disorde	"Honestly, I have b	6	01-Jun-15	47
27339	Imitrex	Migraine	"At first I suffered	8	16-Oct-12	6
51452	Azithromycin		"Very good respon	10	18-Aug-10	1

Appendix 2 – Reading and merging of the train and test datasets.

```
#%%

# Load spaCy model
nlp = spacy.load('en_core_web_sm')
Run Cell | Run Above | Debug Cell
#%%

# Load patient feedback data
df_train_data = pd.read_csv("C:/Users/andre/Downloads/drugsComTrain_raw.csv")
df_test_data = pd.read_csv("C:/Users/andre/Downloads/drugsComTest_raw.csv")
Run Cell | Run Above | Debug Cell
#%%
merge = [df_train_data,df_test_data]
df_review_data= pd.concat(merge,ignore_index=True)

df_review_data.shape    #check the shape of merged_data
Run Cell | Run Above | Debug Cell | Go to [48]
#%%
```

Appendix 3 – Data cleaning and preprocessing python script.

```
Run Cell | Run Above | Debug Cell | Go to [54]
#%%
#check isnull values and features with the highest number of isnull
df_review_data.isnull().sum()

Run Cell | Run Above | Debug Cell | Go to [56]
#%%
# dropping the null values
df_review_data.dropna(inplace=True, axis=0)

Run Cell | Run Above | Debug Cell | Go to [59]
#%%
# function to preprocess the review
def preprocess_text(review):
    doc = nlp(review.lower())
    return ' '.join(token.text for token in doc if not token.is_stop and not token.is_p

Run Cell | Run Above | Debug Cell | Go to [60]
#%%
df_review_data['cleaned_feedback'] = df_review_data['review'].apply(preprocess_text)


Run Cell | Run Above | Debug Cell
#%%
df_review_data.head(10)
Run Cell | Run Above | Debug Cell | Go to [7]
#%%
```

Appendix 4: Data cleaning and preprocessing deliverable dataset with the added feature (cleaned feedback).

```
[62]  df_review_data.head(10)
```

	uniqueID	drugName	condition	review	rating	date	usefulCount	cleaned_feedback
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27	effect combination bystolic mg fish oil
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192	son halfway fourth week intuniv concerned bega...
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17	oral contraceptive pill cycle light periods ma...
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	3-Nov-15	10	time form birth control glad went patch months...
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27-Nov-16	37	suboxone completely turned life feel healthier...
5	155963	Cialis	Benign Prostatic Hyperplasia	"2nd day on 5mg started to work with rock hard...	2	28-Nov-15	43	day mg started work rock hard erections experi...
6	165907	Levonorgestrel	Emergency Contraception	"He pulled out, but he cummed a bit in me. I t...	1	7-Mar-17	5	pulled cummed bit took plan b hours later took...
7	102654	Aripiprazole	Bipolar Disorde	"Abilify changed my life.	10	14-Mar	32	abilify changed life hope

Appendix 5: Testing

```
    df_review_data.head()
    df_review_data.describe()
    df_review_data.info()
    df_review_data.count()
```

[53] ✓ 0.1s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 215063 entries, 0 to 215062
Data columns (total 7 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   uniqueID     215063 non-null  int64
 1   drugName     215063 non-null  object
 2   condition    213869 non-null  object
 3   review       215063 non-null  object
 4   rating       215063 non-null  int64
 5   date         215063 non-null  object
 6   usefulCount  215063 non-null  int64
dtypes: int64(3), object(4)
memory usage: 11.5+ MB

uniqueID       215063
drugName       215063
condition      213869
review         215063
rating         215063
date           215063
usefulCount    215063
dtype: int64
```

Press Shift + Enter to execute.

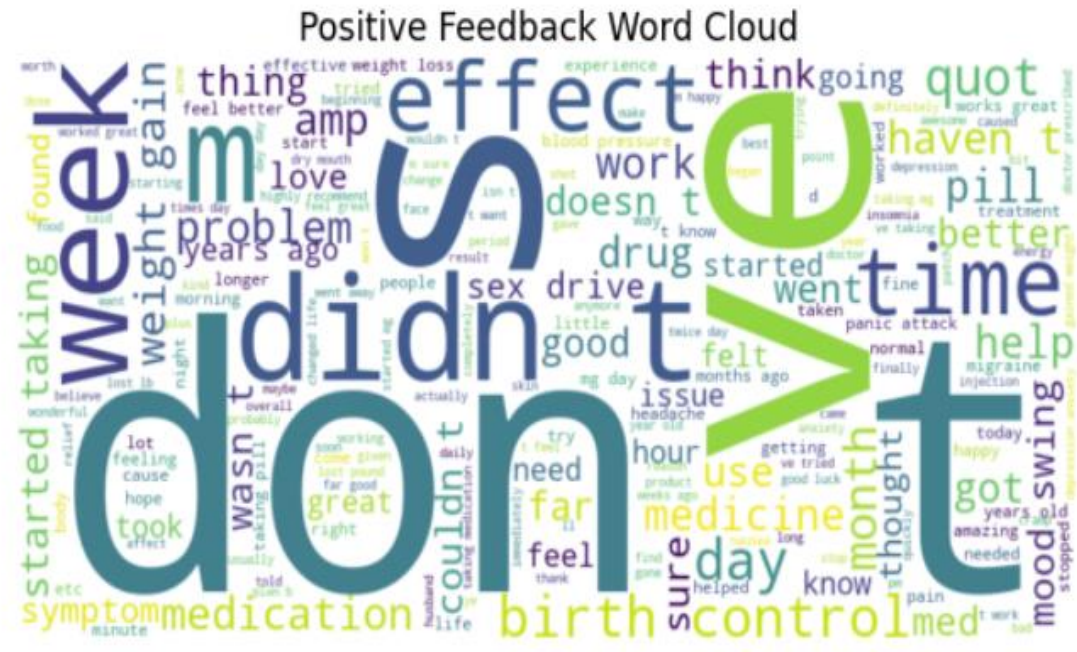
Appendix 6: Exploratory/ Data Visualization

```
# Plot the distribution of sentiment scores
plt.figure(figsize=(10, 6))
sns.histplot(df_review_data['sentiment_score'], bins=30, kde=True)
plt.title('Distribution of Sentiment Scores')
plt.xlabel('Sentiment Score')
plt.ylabel('Frequency')
plt.show()

Run Cell | Run Above | Debug Cell
#%%
# Plot the distribution of sentiment categories
plt.figure(figsize=(10, 6))
sns.countplot(data=df_review_data, x='sentiment')
plt.title('Distribution of Sentiment Categories')
plt.xlabel('Sentiment')
plt.ylabel('Count')
plt.show()
Run Cell | Run Above | Debug Cell
#%%
# Correlation between specific drugs and sentiment
# Assuming there's a 'drug_name' column in df_review_data
drug_sentiment = df_review_data.groupby('drug_name')['sentiment_score'].mean().reset_index()
drug_sentiment = drug_sentiment.sort_values(by='sentiment_score', ascending=False)

Run Cell | Run Above | Debug Cell
#%%
plt.figure(figsize=(15, 10))
sns.barplot(data=drug_sentiment, x='sentiment_score', y='drug_name')
plt.title('Average Sentiment Score by Drug')
plt.xlabel('Average Sentiment Score')
plt.ylabel('Drug')
plt.show()
```

Appendix 7: Continuation of Data Exploratory showing word cloud code

```
121    #%%
122    plt.figure(figsize=(15, 10))
123    sns.barplot(data=drug_sentiment, x='polarity', y='drugName')
124    plt.title('Average Sentiment Score by Drug')
125    plt.xlabel('Average Sentiment Score')
126    plt.ylabel('Drug')
127    plt.show()
128
       Run Cell | Run Above | Debug Cell | Go to [81]
129    #%%
130    # Generate word clouds for positive and negative feedback
131    positive_feedback = ' '.join(df_review_data[df_review_data['sentiment'] ==
132    negative_feedback = ' '.join(df_review_data[df_review_data['sentiment'] ==
133
134    positive_wordcloud = WordCloud(width=800, height=400, background_color='wh
135    negative_wordcloud = WordCloud(width=800, height=400, background_color='wh
136
       Run Cell | Run Above | Debug Cell | Go to [82]
137    #%%
138    # Plot the word clouds
139    plt.figure(figsize=(15, 7))
140    plt.subplot(1, 2, 1)
141    plt.imshow(positive_wordcloud, interpolation='bilinear')
142    plt.title('Positive Feedback Word Cloud')
143    plt.axis('off')
144
145    plt.subplot(1, 2, 2)
146    plt.imshow(negative_wordcloud, interpolation='bilinear')
147    plt.title('Negative Feedback Word Cloud')
148    plt.axis('off')
149
150    plt.show()
```

Appendix 8: Positive Feedback Word Cloud

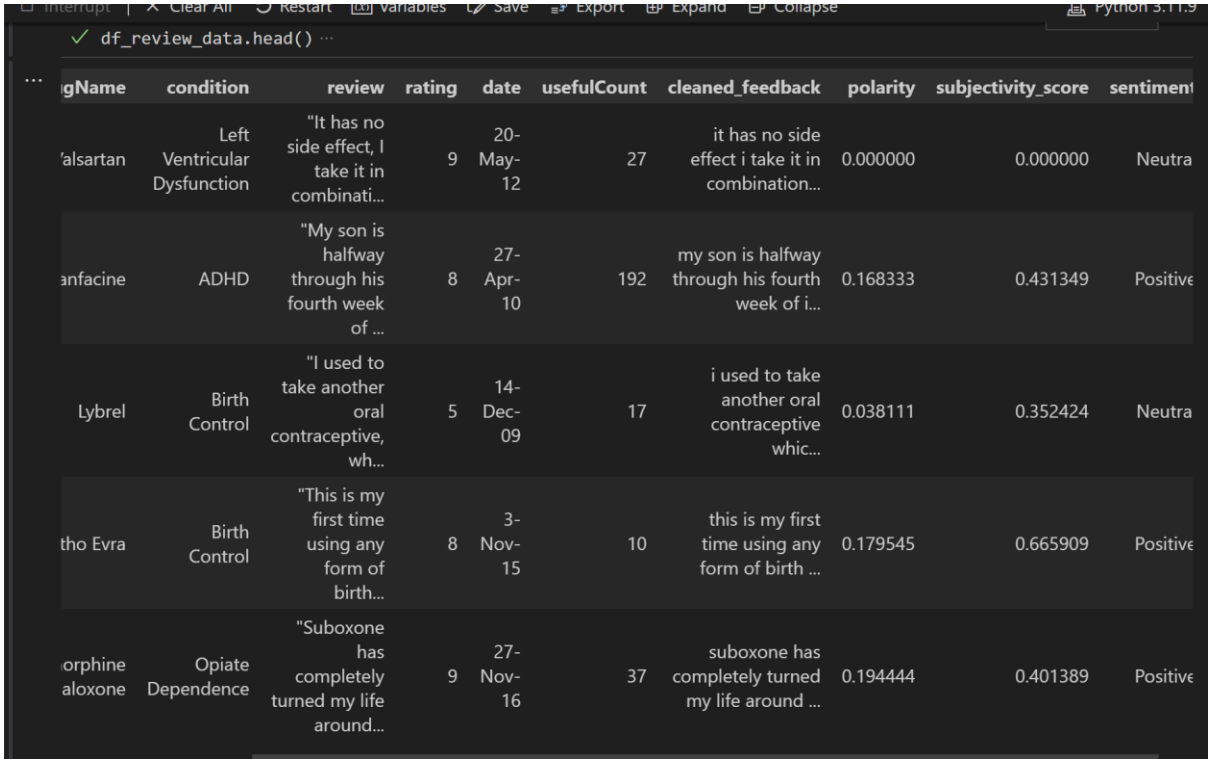


Appendix 9: Negative Feedback Word Cloud

Appendix 10: Screenshot of sentiment analysis application

```
#%%
# function to apply sentiment analysis with TextBlob
def analyze_sentiment(cleaned_feedback):
    blob = TextBlob(cleaned_feedback)
    return blob.sentiment.polarity, blob.sentiment.subjectivity

Run Cell | Run Above | Debug Cell
#%%
df_review_data[['polarity', 'subjectivity_score']] = df_review_data['cleaned_feedback'].

Run Cell | Run Above | Debug Cell
#%%
df_review_data.head()

Run Cell | Run Above | Debug Cell | Go to [13]
#%%
# Classify the sentiment
def classify_sentiment(polarity):
    if polarity > 0.1:
        return 'Positive'
    elif polarity < -0.1:
        return 'Negative'
    else:
        return 'Neutral'

Run Cell | Run Above | Debug Cell | Go to [14]
#%%
df_review_data['sentiment'] = df_review_data['polarity'].apply(classify_sentiment)

Run Cell | Run Above | Debug Cell | Go to [15]
#%%
df_review_data.head()
```

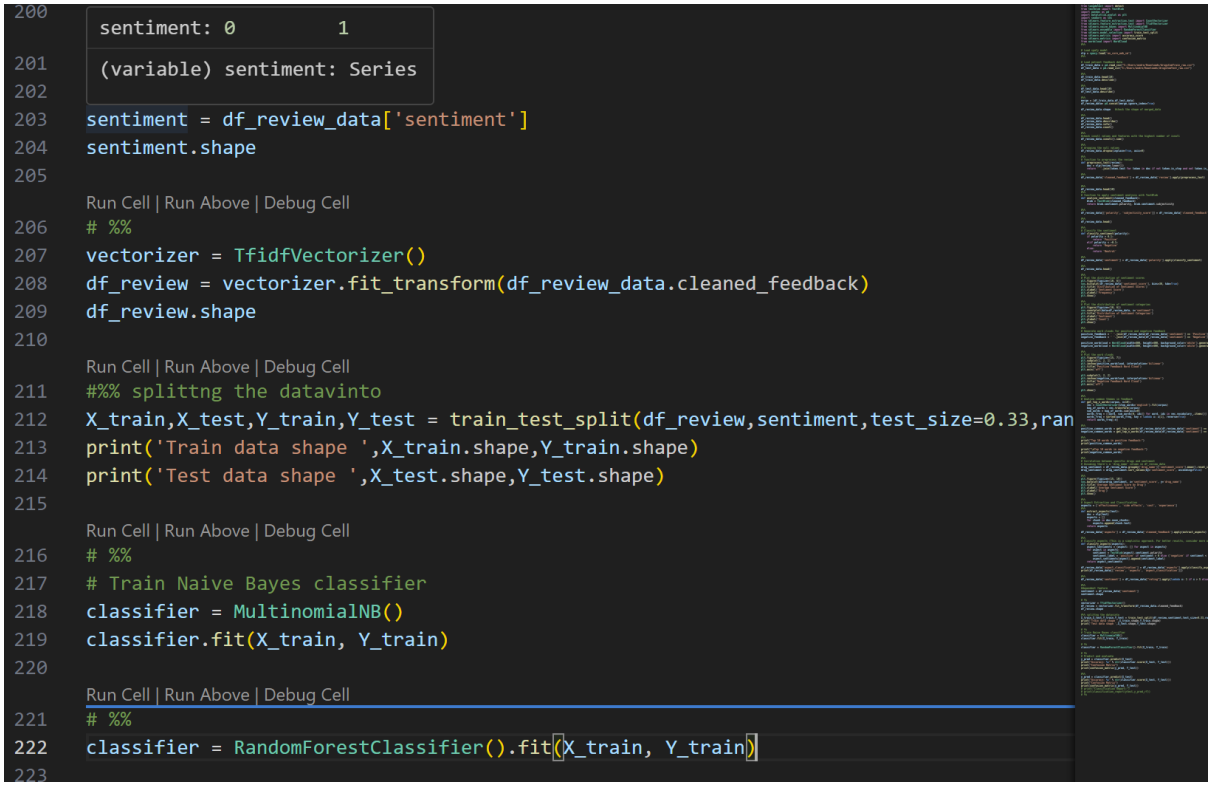
Appendix 11: sentiment analysis deliverable dataset including the added features.



df_review_data.head()

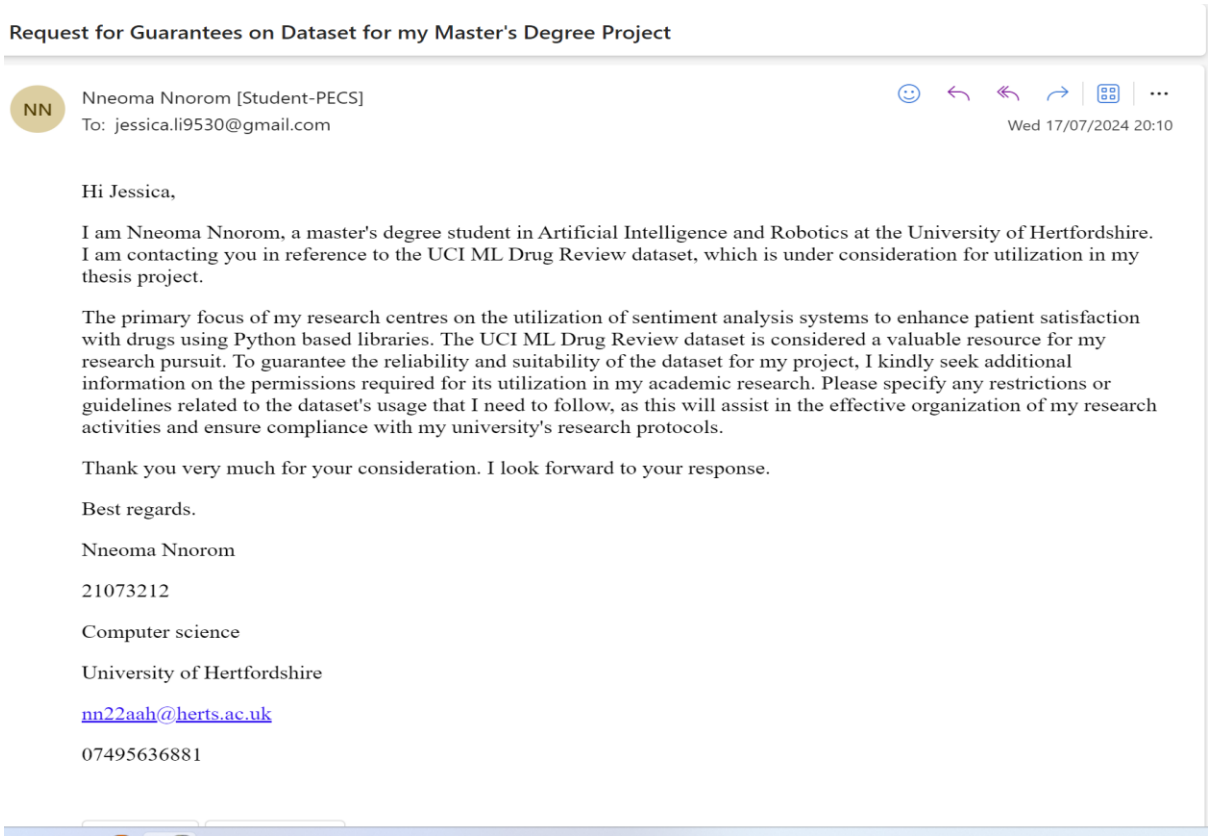
	gName	condition	review	rating	date	usefulCount	cleaned_feedback	polarity	subjectivity_score	sentiment
	'alsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27	it has no side effect i take it in combination...	0.000000	0.000000	Neutra
	anfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192	my son is halfway through his fourth week of i...	0.168333	0.431349	Positive
	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17	i used to take another oral contraceptive whic...	0.038111	0.352424	Neutra
	tho Evra	Birth Control	"This is my first time using any form of birth...	8	3-Nov-15	10	this is my first time using any form of birth ...	0.179545	0.665909	Positive
	orphine aloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27-Nov-16	37	suboxone has completely turned my life around ...	0.194444	0.401389	Positive

Appendix 12: Feature extraction, data splitting and model training



```
sentiment: 0         1
(variable) sentiment: Series

sentiment = df_review_data['sentiment']
sentiment.shape

Run Cell | Run Above | Debug Cell
# %%
vectorizer = TfidfVectorizer()
df_review = vectorizer.fit_transform(df_review_data.cleaned_feedback)
df_review.shape

Run Cell | Run Above | Debug Cell
#%% splittng the datavinto
X_train,X_test,Y_train,Y_test = train_test_split(df_review,sentiment,test_size=0.33,ran
print('Train data shape ',X_train.shape,Y_train.shape)
print('Test data shape ',X_test.shape,Y_test.shape)

Run Cell | Run Above | Debug Cell
# %%
# Train Naive Bayes classifier
classifier = MultinomialNB()
classifier.fit(X_train, Y_train)

Run Cell | Run Above | Debug Cell
# %%
classifier = RandomForestClassifier().fit(X_train, Y_train)
```

Appendix 13a: Evidence of request for guarantees on Dataset.



Request for Guarantees on Dataset for my Master's Degree Project

NN Nneoma Nnorom [Student-PECS]
To: jessica.li9530@gmail.com
Wed 17/07/2024 20:10

Hi Jessica,

I am Nneoma Nnorom, a master's degree student in Artificial Intelligence and Robotics at the University of Hertfordshire. I am contacting you in reference to the UCI ML Drug Review dataset, which is under consideration for utilization in my thesis project.

The primary focus of my research centres on the utilization of sentiment analysis systems to enhance patient satisfaction with drugs using Python based libraries. The UCI ML Drug Review dataset is considered a valuable resource for my research pursuit. To guarantee the reliability and suitability of the dataset for my project, I kindly seek additional information on the permissions required for its utilization in my academic research. Please specify any restrictions or guidelines related to the dataset's usage that I need to follow, as this will assist in the effective organization of my research activities and ensure compliance with my university's research protocols.

Thank you very much for your consideration. I look forward to your response.

Best regards.

Nneoma Nnorom

21073212

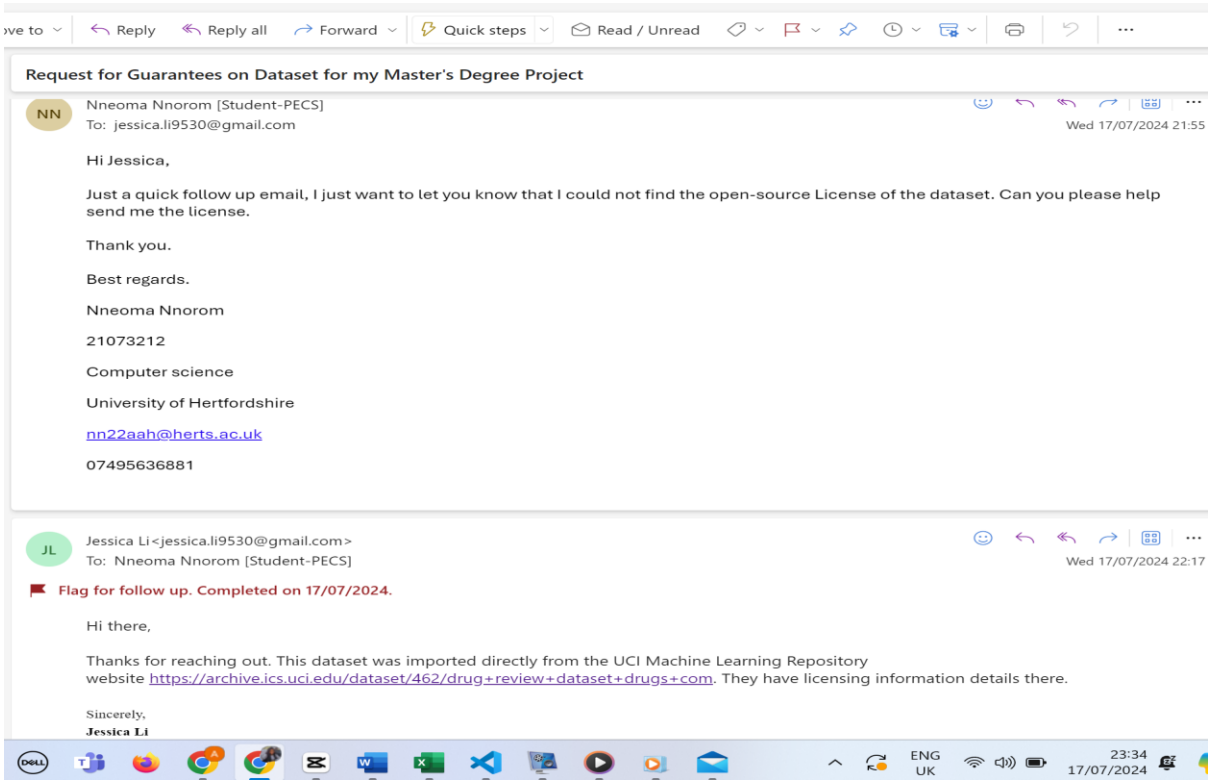
Computer science

University of Hertfordshire

nn22aah@herts.ac.uk

07495636881

Appendix 13b: Evidence of request for Dataset License.



ove to | Reply | Reply all | Forward | Quick steps | Read / Unread

Request for Guarantees on Dataset for my Master's Degree Project

NN Nneoma Nnorom [Student-PECS]
To: jessica.li9530@gmail.com
Wed 17/07/2024 21:55

Hi Jessica,

Just a quick follow up email, I just want to let you know that I could not find the open-source License of the dataset. Can you please help send me the license.

Thank you.

Best regards.

Nneoma Nnorom

21073212

Computer science

University of Hertfordshire

nn22aah@herts.ac.uk

07495636881

JL Jessica Li<jessica.li9530@gmail.com>
To: Nneoma Nnorom [Student-PECS]
Wed 17/07/2024 22:17

Flag for follow up. Completed on 17/07/2024.

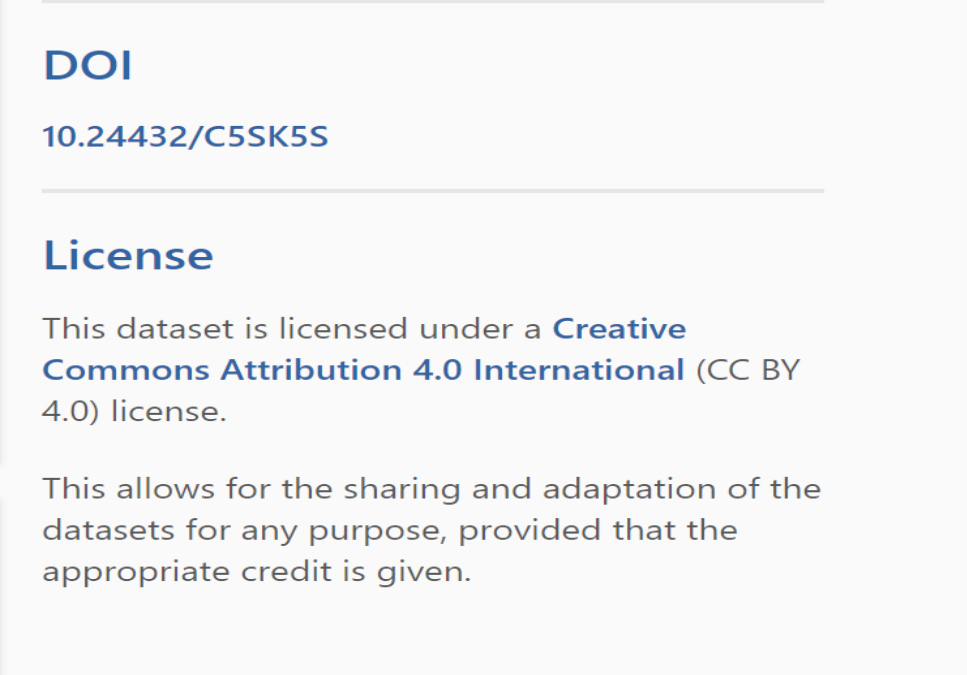
Hi there,

Thanks for reaching out. This dataset was imported directly from the UCI Machine Learning Repository website https://archive.ics.uci.edu/dataset/462/drug+review+dataset+drugs+com. They have licensing information details there.

Sincerely,
Jessica Li

ENG UK 23:34 17/07/2024

Appendix 14: Evidence of request for guarantees on Dataset.



DOI

10.24432/C5SK5S

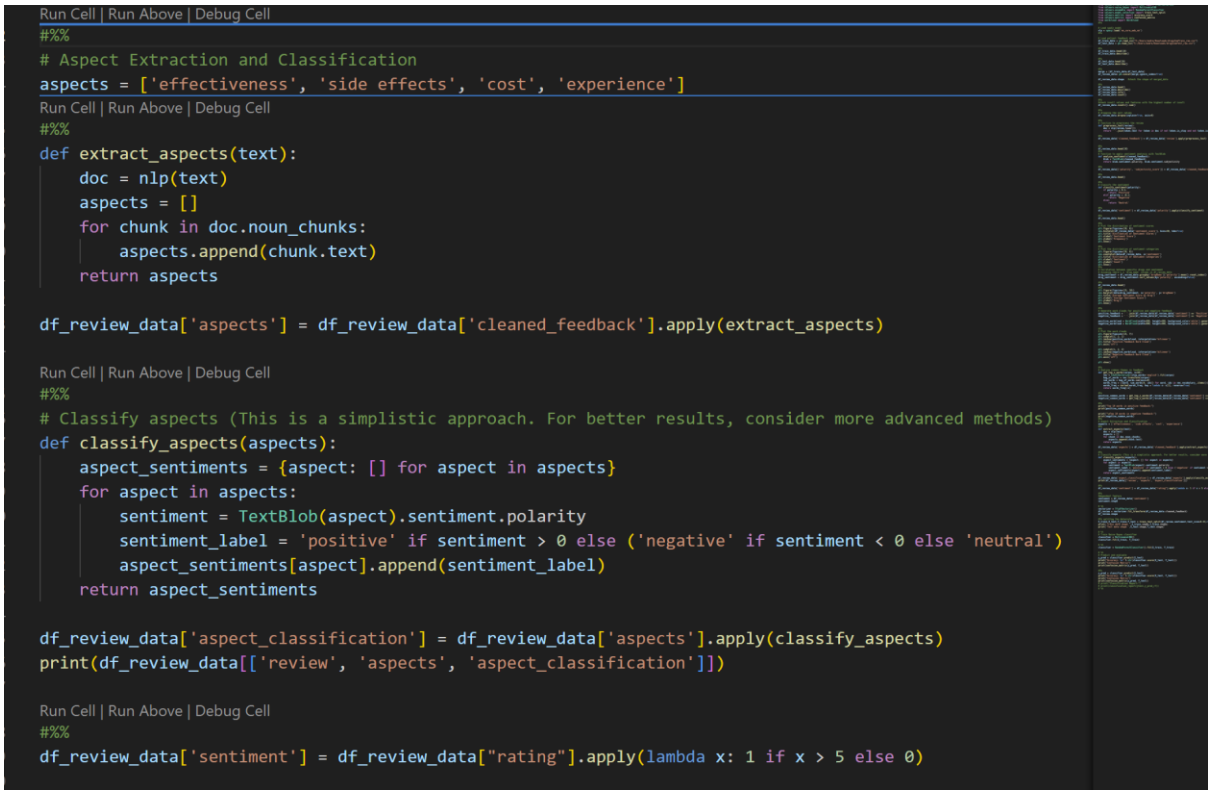
License

This dataset is licensed under a **Creative Commons Attribution 4.0 International** (CC BY 4.0) license.

This allows for the sharing and adaptation of the datasets for any purpose, provided that the appropriate credit is given.

Link (https://archive.ics.uci.edu/dataset/462/drug+review+dataset+drugs+com)

Appendix 15: Aspect-Based Extraction and Classification.



```
Run Cell | Run Above | Debug Cell
#%%
# Aspect Extraction and Classification
aspects = ['effectiveness', 'side effects', 'cost', 'experience']
Run Cell | Run Above | Debug Cell
#%%
def extract_aspects(text):
    doc = nlp(text)
    aspects = []
    for chunk in doc.noun_chunks:
        aspects.append(chunk.text)
    return aspects

df_review_data['aspects'] = df_review_data['cleaned_feedback'].apply(extract_aspects)

Run Cell | Run Above | Debug Cell
#%%
# Classify aspects (This is a simplistic approach. For better results, consider more advanced methods)
def classify_aspects(aspects):
    aspect_sentiments = {aspect: [] for aspect in aspects}
    for aspect in aspects:
        sentiment = TextBlob(aspect).sentiment.polarity
        sentiment_label = 'positive' if sentiment > 0 else ('negative' if sentiment < 0 else 'neutral')
        aspect_sentiments[aspect].append(sentiment_label)
    return aspect_sentiments

df_review_data['aspect_classification'] = df_review_data['aspects'].apply(classify_aspects)
print(df_review_data[['review', 'aspects', 'aspect_classification']])

Run Cell | Run Above | Debug Cell
#%%
df_review_data['sentiment'] = df_review_data["rating"].apply(lambda x: 1 if x > 5 else 0)
```

Appendix 16: Result of Aspect Extraction and Classification

```
df_review_data.head(10)
```

e	condition	review	rating	date	usefulCount	year	month	cleaned_feedback	aspects	aspect_classification
n	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	2012-05-20	27	2012	5	effect combination bystolic mg fish oil	[effect combination, mg fish oil]	{'effect combination': ['neutral'], 'mg fish o...
e	ADHD	"My son is halfway through his fourth week of ...	8	2010-04-27	192	2010	4	son halfway fourth week intuniv concerned begi...	[son, high dose day, hardly bed cranky sleep, ...	{'son': ['neutral'], 'high dose day': ['positi...
el	Birth Control	"I used to take another oral contraceptive, wh...	5	2009-12-14	17	2009	12	oral contraceptive pill cycle light period max...	[oral contraceptive pill cycle light period ma...	{'oral contraceptive pill cycle light period m...
ra	Birth Control	"This is my first time using any form of birth...	8	2015-11-03	10	2015	11	time form birth control glad go patch month de...	[time form, birth control, max cramp, intense ...	{'time form': ['neutral'], 'birth control': ['...
e e	Opiate Dependence	"Suboxone has completely turned my life around...	9	2016-11-27	37	2016	11	suboxone completely turn life feel healthy exc...	[suboxone, life, healthy excelling job money p...	{'suboxone': ['neutral', 'neutral'], 'life': [...

```
Press Shift+Enter to execute.
```

```
TERMINAL  PORTS  JUPYTER                                   powershell
```